

Dissertation

Statistical Approaches Supporting QbD Milestones via Bioprocess Digital Twins

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

Prof. Dr. Christoph Herwig

E166 Institut für Verfahrenstechnik, Umwelttechnik, Technische Biowissenschaften

und Mitbetreuer

Prof. Dr. Peter Filzmoser

E105 Computational Statistics Institute of Statistics and Mathematical Methods in Economics

eingereicht an der Technischen Universität Wien

Fakultät für Technische Chemie

von

Christopher Taylor MSc., B.S.

Matrikelnr.: 11736819

Wien, am

eigenhändige Unterschrift

Christopher M. Taylor, MSc

Statistical Approaches Supporting QbD Milestones via Bioprocess Digital Twins

Dissertation, December 12, 2022

Reviewers: Prof.Dr. Krist Gernaey and Prof. Dr-Ing. Ralf Takors

Supervisors: Prof. Dr. Christoph Herwig and Prof. Dr. Peter Filzmoser

Technische Universität Wien

Bioverfahrenstechnik

Institute of Chemical, Environmental and Bioscience Engineering

Faculty of Technical Chemistry

Karlsplatz 13

1040 and Vienna

Acknowledgements

Fulfilling the dream of defending this thesis was not achievable alone.

Firstly, to Professor Christoph Herwig, an enormous thank you for offering me the opportunity to pursue this goal. Your guidance combined with freedom to explore this field was invaluable to success in advancing this vision. Even more so, thank you for having the confidence in a mature student to persevere through the program. An additional heartfelt thank you to Professor Peter Filzmoser, who provided valuable support and statistical review in bringing this thesis into being.

My deeply felt thanks to my scientific colleagues and collaborators in this journey: Thomas Zahel, Lukas Marschall, Barbara Pretzner, and Daniel Borchert, who all provided a friendly, collegiate, and motivating atmosphere – and who brought the IPM to life. To my Ex-putec/Werum/Körper colleagues, thank you for your collaboration and team spirit: to Petra Lubitz, who offered support during the hard times; to my colleagues and team members throughout the years – Valentin Steinwandter, Damiano Totaro, Lukas Knosp, Ignasi Bofarull, and Norhan Mahfouz – your innovative spirits kept this work on track and optimistic.

To former colleagues & supervisors who had different, but essential impacts on my career path and to whom I am very grateful: Kiana Akbaroff & Darwin Richardson for giving me my start and mentorship; Hervé Morales & Katia Ceré-Monnat for believing in my transition to Europe; Jörg Gampfer for mentoring and supporting my dreams of advancing my education and experience in Vienna, where I've now spent 10 unexpected and lovely years.

On a personal level, I want to thank my family who have been with me through the thick and thin of this journey: my mother and father, Jamie and Mark, for their constant love & support throughout the years. To my mother Jamie also for her unwavering scientific support, idea sharing, and motivation. To my brothers, Zac, Cole, Wil, and Clint for listening to my rants and still asking for updates.

Above all, to Steph, Amelie, and Francis: you saw me through all of this – you are my rock and foundation. Steph, quite simply, I would have never made it this far without your love and support. This thesis is dedicated to you.

Abstract

Quality by Design is a life-cycle paradigm used by regulators to encourage pharmaceutical companies to consider product quality from the earliest development stages. The objective is to identify and document relationships between critical process parameters and product quality attributes via risk assessments and experimental results before validating for commercial manufacturing. This methodology has steadily gained traction over the last decade and applied statistical tools are increasingly leveraged to reach these goals quantitatively.

Nonetheless, significant gaps remain between the methodological intent and the current state-of-the-art practices. For example, during the identification of critical process elements, latent variables have largely been overlooked due to over-reliance on process knowledge and the absence of relevant extraction and multivariate methods. Subsequent risk assessments and data-driven models are siloed in the individual process steps (unit operations) and are not linked to the patient-relevant outcome: drug substance specifications. Lastly, there is an absence of data feedback loops between the above procedures and the manufacturing data in the commercial life-cycle.

This thesis addresses the above gaps via improvements and applications of an integrated process model; a framework centered on concatenating unit operation models and propagating error via Monte Carlo simulations. To realize this potential, novel procedures were first designed to uncover latent bioprocess variables via extraction and multivariate analysis. Once in place, an innovative Monte Carlo-based application was developed that establishes intermediate acceptance criteria for quality attributes via parameter sensitivity analysis. A further simulation procedure was created which, when combined with linearization techniques, enables the determination of parameter proven acceptable ranges and links these quantitatively to risk assessment severity rankings. Lastly, the integrated process model was substantially improved and inserted architecturally into manufacturing data feedback loops, enabling the model to react in real time to process conditions. The totality of these innovations depicts a major industry objective: a bioprocess digital twin.

Leveraging the developments in this thesis, the proposed integrated process model now quantitatively links process parameters and quality attributes to patient-relevant outcomes. Moreover, it does so with a technology that can iteratively adapt to new manufacturing data, ensuring that it accompanies the process throughout its life-cycle, and thereby establishes an engine for a digital twin. Thus, with holistic process quality as a central goal, the industry will be able to better fulfill both the intention and the potential of the Quality by Design paradigm.

Kurzfassung

"Quality by Design" ist das Paradigma, das von pharmazeutischen Aufsichtsbehörden verwendet wird, um Wissenschaftler zu verpflichten, die Produktqualität in den frühesten Design- und Entwicklungsphasen zu berücksichtigen. Diese Konzepte, die in mehreren kritischen Richtlinien zum Ausdruck kommen, wurden in den letzten zehn Jahren kontinuierlich in das Denken der Industrie integriert. Angewandtes statistisches Design und Analyse sind eine Hauptkomponente dieser Anforderungen, die verwendet werden, um Beziehungen zwischen Prozessparametern und den Produkteigenschaften zu quantifizieren.

Trotz dieser Verbesserungen sind die Lücken zwischen der Absicht der Quality by Design-Methodik und dem aktuellen Stand der Technik deutlich geworden. Die grundlegende Definition der kritischen Prozesselemente basiert immer noch weitgehend auf Prozesswissen, wobei latente Elemente in Ermangelung fortschrittlicher multivariater Verfahren übersehen werden. Qualitative Risikobewertungen und statistische Prozessschrittsmodelle sind nicht direkt mit dem Endergebnis verknüpft, nämlich den Spezifikationen der Arzneimittelsubstanz. Diese sind stellvertretend für die Auswirkungen auf den Patienten.

Diese Arbeit beabsichtigt daher, ein fortschrittliches integriertes Prozessmodell zu entwickeln, das als digitale Bioprozess-Zwillingsmaschine dienen kann, die die Erfüllung von Qualitätszielen durch Design ermöglicht. Dieses Ziel wurde erreicht und die innovativen Applikationen decken die latente Prozessvariablen auf, legen Zwischenakzeptanzkriterien fest, definieren Kontrollstrategien und verknüpfen sie sogar quantitativ mit Risikobewertungsrankings. Schließlich wurde das integrierte Prozessmodell selbst in Rückkopplungsschleifen der Herstelldaten integriert, wodurch das Modell in Echtzeit auf Prozessbedingungen reagieren konnte. Dieses Echtzeitmodell führt zu einem plausiblen digitalen Zwilling des Bioprozesses.

Die Implementierung eines digitalen Zwillings stellt einen neuen Meilenstein in der Kommerzialisierung von Bioprocessen dar. Mit der Qualität und der Wirkung auf den Patienten als zentrales Ziel und mit dem ganzheitlichen Prozess im Fokus kann die Industrie möglicherweise die Absicht und das Potenzial des Quality-by-Design-Paradigmas erfüllen.

Contents

I	Introduction	1
1	Introduction & Motivation	3
1.1	Background	3
1.2	Problem Statement	6
1.2.1	Methodological Gaps	6
1.2.2	Control Strategies	7
1.2.3	Process Knowledge and Risk Assessment	9
1.2.4	Real Time Applications	10
1.3	Why Now	11
1.4	Goal of this Thesis	12
	References	15
II	Results	21
2	Results, Findings & Achievements	23
2.1	Thesis Structure	23
2.2	Findings and Achievements	24
2.2.1	Multivariate Workflow for Latent Parameters	24
2.2.2	QbD Milestones Applications via IPM	25
2.2.3	Specification-Driven Acceptance Criteris	26
2.2.4	Integrated Process Model 2.0	27
3	Manuscripts	29
3.1	Multivariate MonitoringWorkflow for Formulation Fill and Finish Processes	29
3.2	Integrated Process Model Applications Linking Bioprocess Development to Quality by Design Milestones	46
3.3	Specification-Driven Acceptance Criteria for validation of biopharmaceutical processes	63
3.4	Architectural & Technological Improvements to Integrated Bioprocess Models towards Real-Time Applications	80

III	Conclusions	97
4	Conclusion	99
4.1	Summary	99
5	Impact	103
6	Outlook	105
6.1	Conclusion	106
IV	Appendix	109
A	Appendix	111
A.1	A1 Supporting Information: Integrated Process Model Applications Linking Bioprocess Development to Quality by Design Milestones	111
A.2	A2 Supporting Information: Specification-Driven Acceptance Criteria for validation of biopharmaceutical processes	123
A.3	A3 Supporting Information: Architectural & Technological Improvements to Integrated Bioprocess Models towards Real-Time Applications	127
	Declaration	133

Part I

Introduction

Introduction & Motivation

“ *Progress in modifying our concept of control has been and will be comparatively slow. In the first place, it requires the application of certain modern physical concepts; and in the second place it requires the application of statistical methods which up to the present time have been for the most part, left undisturbed in the journal in which they appeared*

— **Walter Shewhart**

“ *All chance systems of causes are not alike in the sense that they enable us to predict the future in terms of the past.*

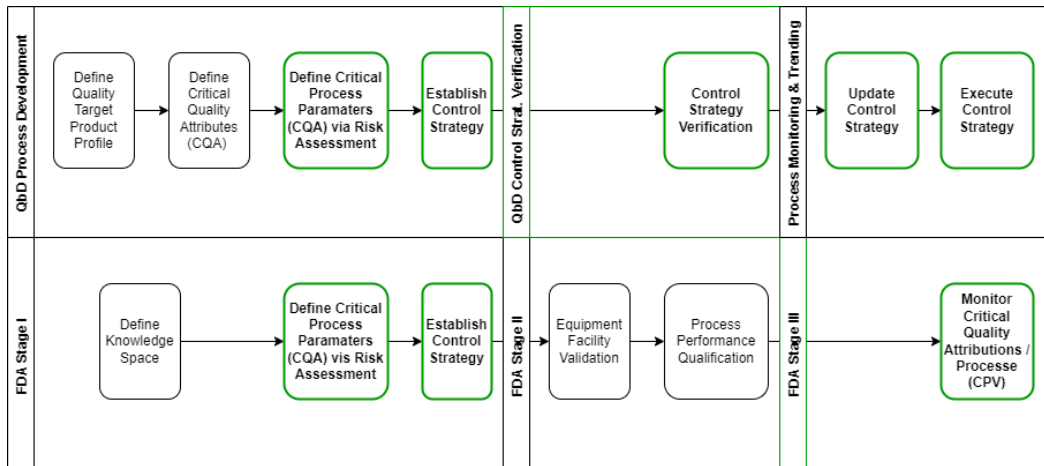
— **Walter Shewart**
-Postulate 1-

1.1 Background

When Joseph Juran formulated the concept of Quality by Design (QbD) for industry in the early 1990s, the novelty of his framework was its prodding of product developers to consider quality even as early as the design phase [17, 51]. Juran's principle criticism of the prevailing paradigm was that quality was considered primarily an 'a posteriori' concern. That is, the quality of the product could be inferred by successful quality testing at the finished product stage. Failures during quality testing lead to excess product discards, but otherwise do not call the process or product into question. This methodology is sometimes referred to as "testing into quality." In Juran's vision, quality should be a concern at least as important as product effectiveness. And in pursuit of this, development should employ resources to consider quality as early on as possible [19].

Naturally when the product has an impact on the health and well-being of its consumer, quality takes on a more urgent role than even Juran may have originally expressed [53]. Nowhere is this more evident than in the pharmaceutical industry, whose medical products ipso facto impact patient health. It is all the more so for biopharmaceutical products, most of which are administered parenterally and thus have a maximally direct effect on the patient's health. It is little wonder then that the US Food and Drug Administration (FDA),

Fig. 1.1.: Simplified swim lane diagrams of (bottom) the FDA Validations 2011 Process Validation: General Principles and Practices document [36] and the (top) QbD methodology per ICQ Q8 [1]. Key milestones, such as the control strategy, discussed in this dissertation are shown in bold green boxes. Despite minor differences, the two are nearly identical approaches. Broadly summarized: first define the critical attributes of a process, then define the parameters that affect them and finally, build a control strategy for these parameters



through the work of Azaj Hussain and others, as well as the European Medicines Agency (EMA), took up this concept and heavily promoted the QbD framework [43]. It has been an incremental and iterative ongoing process, boosted first by the 2002 Pharmaceutical Current Good Manufacturing Practice (cGMP) for the 21st Century Initiative [39] and in much more detail by the publishing of the International Conference on Harmonization (ICH) documents. The ICH guidelines outline a regulatory application system based expansively on the QbD concepts [53].

In recent years, QbD has become an established framework for biopharmaceutical development, with numerous articles highlighting systematic advances [40, 26, 46]. Regulatory authorities have translated the principles of QbD into industry language: concepts such as quality product profiles, risk assessment, risk management, and control strategies have become common parlance. The underlying philosophy of these translations appears in numerous research articles (in addition to Juran's writings), but was primarily codified in the 2011 FDA Validation Guidelines [36], supported by the ICH guidelines. The latter explicitly states that quality cannot be tested into products; quality should be built by design" [1].

If embedding the concepts in the industry language is critical for buy-in, establishing clear deliverables is critical for compliance [29, 9]. In particular, processes steps may often be depicted by quantitative mathematical models, and thus the QbD milestones should also be defined quantitatively and achieved through data from experiments or manufacturing. This serves the purpose of augmenting the existing paradigm of excessive reliance on human knowledge or expertise. While expertise is, of course, paramount to development work, such knowledge is easily lost or forgotten, not easily replaced, and occasionally prone to hubris

and error [38, 34, 50, 48]. Data is therefore the foundation of knowledge management and, by extension, the quality structure.

And in defining data-driven definitions of these concepts, the regulatory authorities ran quickly into several issues plaguing the biopharmaceutical sector. Foremost is the need for a stringent and structured approach to data driven experimental design and analysis in order to adequately prove quality in the development stages. Statistically underpinned design and analysis predate QbD by half a century. However, these tools were seldom applied in pharmaceutical development before the 1970s [43]. Even then, biopharmaceutical manufacturers have been slow to implement these principles, even as late as the new century [5]. This is in large part because biopharmaceuticals are extremely expensive to develop and thus the experimental resources are always very limited [31]. Nonetheless, even this hurdle has been incrementally overcome in the last decade due to increasingly clever designs, more accessible statistical tools, and greater buy-in from the regulators and management [4, 22].

Additional challenges arise from the regulatory introduction of the statistical process control concept found throughout the guidelines; paraphrased as "The process is the product [18, 51]." In other words, if the process is not under control across all process steps (or unit operations), no amount of quality testing performed at the final product step is sufficient to prove the overall quality of the product. Both the bioprocess and the resulting bioproducts are highly complex, requiring numerous interacting steps, and are often relatively poorly understood [10]. Therefore, at each critical step of the process, a procedure must be followed which assesses risk and then later proves control of the steps via monitoring the quality attributes [36]. This heightened control of the individual unit operation has been well discussed and translated into industry logic through the QbD methodology.

In summary, much progress has been made in the biopharmaceutical industry over the last 30 years. QbD concepts of pre-defined quality goals have been accepted. Statistical modeling has become a standard part of characterization. Methods of statistical process control methods have been investigated. With all these tools as a baseline, and the challenges still being faced, it becomes vital to revisit the original goals of QbD per ICH, FDA, and EMA guidelines and reassess how these may be derived as data-driven deliverables. Upon closer inspection, it becomes clear that our measures are not holistically answering the questions Juran sought to answer at the beginning of this journey. In the next chapter, we will describe these issues through three prisms of general methodological gaps: control strategy deficiencies, insufficient process knowledge management, and the absence of real-time solutions.

1.2 Problem Statement

1.2.1 Methodological Gaps

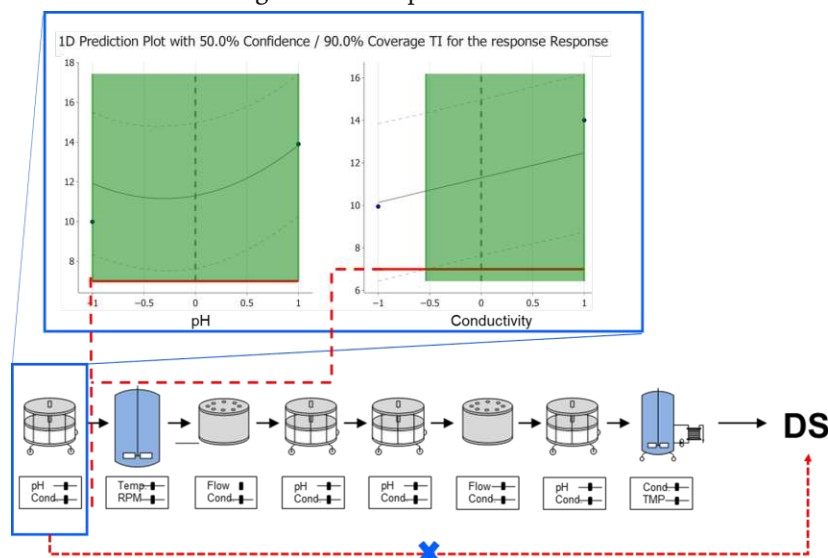
Despite the encouraging trends in the industry described above, persistent gaps appear between the intended deliverables and the actual output of state-of-the-art procedures, vis-à-vis the intent of the Quality by Design framework.

Firstly, the definition of controllable parameters is still under-explored. The established risk management approach requires the definition of critical process parameters (CPPs); those controllable factors which have an impact on critical quality attributes (CQAs) [1]. While many of these factors are straightforward, such as pH, temperature, or pressure, numerous other descriptors of the process remain latent within the increasingly vast sources of input data. Specifically, important process parameters may be a combination of multiple factors, or extractable only from time series data. Currently, such data sources are often prohibitively complicated, or require too many resources, to routinely extract and model these variables [41, 31, 32, 33]. This leads to missing key insights into the behavior or the process.

Secondly, bioprocesses are generally divided into numerous unit operations, unconnected in development to each other and the greater process chain. Indeed, the foundation of all quantitative applications of QbD methodology is the unit operation model. That is, current best practice models are generally mathematical models of these individual process steps in all three process segments: upstream processing (cell culture and harvest), downstream processing (purification), and formulation [12]. These models are generally data-driven and statistically underpinned, but of course, may also be mechanistic or hybrid for a given unit operation [47]. Critically, however, there are very few connections from these unit operations to the final drug product. Methods which attempt to make this connection include, amongst others, flowsheet models, most common in traditional pharmaceutical processes, which connect outputs of one unit operation to inputs of another, but without concentrating on maintaining CQA quality across all unit operations [27]. Bayesian approaches have also been proposed, though not yet widely implemented, which would allow combining models as informed prior distributions in order to present a posterior distribution, from which the subsequent unit operation's prior distribution could be sampled. However, these proposals do not describe a way to handle extrapolations [30].

An integrated process model (IPM) framework has been described in 2017, which proposes concatenating regression models to predict quality holistically across the process chain. This is achieved by describing the relationship between CQA pool values (outputs) at a given unit operation and the subsequent unit operations's CQA load values (input) as a means of linking unit operations. The IPM has an elegant and simple inferential backbone, allowing easy interpretation for regulatory submission. Nonetheless, it also has similar limitations to those described above [52]. While being the preferred framework for holistic process

Fig. 1.2.: An example of a univariate control strategy (PAR) plot. The model of the two factors (pH and conductivity) is shown against a response, including a tolerance interval (dotted line) and local intermediate acceptance criteria (red). Wherever the model plus tolerance interval intersects this threshold marks the end of the proven acceptable ranges. This definition is problematic in that it does not connect to downstream UOs and chooses acceptance criteria only relevant to the current UO and not the drug substance specifications



predictions, the IPM, as originally proposed, has a complex and algebraically sub-optimal data model and has no described link to a real-time environment.

The following sections explore the specific QbD objectives impacted by the above described gaps in technology and methodology. Control strategies procedures are lacking linkage to final drug substance, to manufacturing scale, and to statistically important data(1.2.2). Process knowledge and risk assessments are missing not only this link to final drug substance, but also lack quantitative feedback loops to each other(1.2.3). Finally, lack of real time adaptation of the above two topics leads to risk of quality loss and opportunity(1.2.4).

1.2.2 Control Strategies

The process control strategy's primary goal is to "describe and justify how in-process controls..., intermediates... contribute to the final product quality [1]." That is, monitoring of the intermediate steps of the process should prove consistent quality at the patient-impacting drug substance step. ICHQ11 further reinforces this issue by clearly saying that a control strategy must ensure that the drug substance CQAs are "within the appropriate range limit or distribution," in the described unit operation. The implication is that, while we can most easily control the last unit operation before drug substance, we should be able to describe the relationship of these variables to the final CQAs at any upstream unit operation [2].

As described above, in practice, the control strategies are set for each of the unit operations individually. This answers questions about the statistical control of that given unit operation, but it does not answer the fundamental question posed by the guideline: does the process impact the CQAs (i.e. patient impact), which are assessed at drug substance? Current control strategies may be able to describe the local impact, but are disconnected from the Drug Substance specifications.

The primary reason for this disconnect is the statistical complexity and resources required to model the impact of CPPs on CQAs across multiple unit operations, each with its own unique set of potential model terms. In a given model, the number of terms (and thus degrees of freedom) needed for the full model equation expands linearly with the number of main effects and more exponentially with increasing higher order effects¹. Adding multiple unit operations will cause the interaction effects to explode. Linkage studies have therefore often been seen as a proxy for true inter-unit operation multivariate statistics, wherein CPPs are adjusted in a given unit operation and the responses are measured only in later unit operations. However, these ignore the interactions entirely and would thus also quickly fall prey to the lack of available resources to effectively gain sufficient degrees of freedom to model inter-unit operation interactions [23].

Secondly, these control strategies do not address *Scale* of the process in the holistic goals of the FDA process validation guidelines. The FDA seeks to have validation be linear from development through to commercial manufacturing. In practice, there is a significant disconnect between the scale of the data sources in the different steps. For example, during process characterization (Stage I), most of the experimental data is generated at lab-scale. It is with this small mock-up of the industrial process that design spaces are explored and control strategies set [12]. This data typically does not include any large, commercial scale data at all, even if there likely exists at least minimal data from this scale. This leaves out critical information and may ignore scale offsets.

Conversely, in the commercial continued process verification stage (Stage III), the control strategy is executed solely at manufacturing scale, and importantly, is only updated with further manufacturing scale data. This data, being mostly at set point, is not as informative as the original characterization data. But, in our experience, this original design space exploration data is never included in the control strategy updates. Indeed to our knowledge, holistic use of large and small-scale data at all points in the validation life-cycle is not common practice and is therefore missing out on key insights, in spite of some efforts to pull these sources together [34].

Reasons for this range from operational to compliance related. Much has to do with the simple availability of development and manufacturing data in a combined central repository [11]. But above all the mixing of GMP data from manufacturing with potentially non-GMP data at small-scale, is often seen as an excessive compliance risk. Thus it is perceived that non-GMP and manufacturing data not mix [6]

¹more specifically, in the classic 'response surface model' wherein interactions and quadratic effects of the main effects are considered, the increase in terms is $2n + \frac{n(n-1)}{2}$ when considering maximally 2-factor interactions and $2n + \frac{n^3-n}{6}$ if considering 3-factor interactions.

1.2.3 Process Knowledge and Risk Assessment

The lack of a translation of quantitative relationships into codified process knowledge is another major gap in realizing the full QbD methodological. Quantified process knowledge can be defined here as the model-described relationship of the interactions by and between process parameters on patient quality. That is, these relationships are fitted and then converted to knowledge that can be stored and further applied [14, 49]. By this reasoning, such a repository of process knowledge could be as simple as the collection of regression analyses between independent variables (CPPs and their interactions) and independent variables (CQAs) per unit operation.

However, in practice, the primary repository for process knowledge is populated by process risk assessments [20, 3]. Risk assessments are generally performed as a variation of a Failure Mode and Effect Analysis (FMEA) or equivalent semi-quantitative evaluation of quality risk, generally measured by a discrete ranking scale. In brief, risk assessments evaluate and assign a rank to the impact of a process parameter's deviation from set point on the measured CQAs. Process parameters that have risks rated higher than a certain predefined threshold are assigned the designation CPP and require a control strategy. This rating is evaluated by expertise (and literature if possible) on the categories of severity, frequency, and detectability of the deviation [16].

It is important to note that in current standard practice, data does not quantitatively affect the rankings, but rather influences the expert discussion around the ranking [25]. This lack of a direct or quantitative relationship between data and risk rankings is a source of current research in the industry. Several articles have been successful in putting forth procedures to establish this relationship, but none of these address the linkage between the current unit operation and subsequent unit operations. Instead, risk rankings are dependent on assessing the quantitative impact against the intermediate acceptance criteria in the current unit operation. This is in turn reliant on the justification of the intermediate acceptance criteria as proxy for drug substance [42, 25, 24, 13]. Therefore, even cutting-edge methods of linking data to risk assessments are insufficiently connecting these to patient impact.

Inadequate process knowledge, even if quantitative, can have a direct effect on process control. If the risk assessment is the principal form of documenting process knowledge, the control strategy may be seen as the application of this process knowledge towards upholding quality. And if the control strategy is too wide (e.g. insufficient data to tighten ranges), it could be subject to regulatory push-back as being insufficiently stringent in keeping statistical control of the process. Worse, if the process knowledge leads to too-tight a control strategy (e.g. few data points are randomly too close together), the Out-of-Specification (OOS) and discard rates will increase. This represents both a lack and a loss of process knowledge leveraged in the commercial manufacturing process.

1.2.4 Real Time Applications

The QbD methodology is meant to describe living documents, which can therefore be considered life-cycle companions. Hence these process knowledge repositories must be updated with new data [15]. This is most clearly seen in the establishment of the Continued Process Verification Plan (CPV) in Stage III (commercial manufacturing) of the FDA guidelines. Stage III governs the confirmation that the process consistently remains in a validated state of quality. CPV is performed through a series of iterative formal quality evaluations (i.e. yearly or other regular intervals) as well as more frequent and flexible statistical trending and monitoring plans. Ideally, this should be done in real-time as the data is produced. However, in practice, this is virtually never performed as such, since many of the data collecting systems require hours, days, or even weeks to generate the pertinent results [8].

This lack of timely updating of the process information can be categorized in the following ways. First, deviations to the process, in particular those which may have a quality impact, must be investigated in a timely manner as required by the regulators [36]. Exactly how 'timely' is left to the manufacturers to defend, but quality investigations must be thorough and therefore lag behind the actual decision-making vis-à-vis whether to proceed or abort the process [45]. This has the obvious effect that the comprehensive results of any investigation risk not being used to mitigate the current batch, either because the process was provisionally accepted and the process was continued, or because the process was subsequently aborted. In either case, both the quality and business case are clearly at an advantage if the results of deviation investigations can be immediately applied to the QbD methods. Such a rapid risk mitigation tool would lead to both quality and business improvements.

Secondly, on the opposite end of investigative outcomes, CPV is also recommended to look for data indicative of optimizations [36]. However, such process optimization often occurs long after discovery. Seemingly to understand this probability, the FDA makes a subtle, but important, implication within the guidelines. CPV-discovered improvements should be those which can be implemented without a major change to the process, otherwise, they would require a post-approval change notification (a lengthy and expensive regulatory process). This implication is further supported by an even subtler one. CPV plans monitor processes running at set point, given variation in the Normal Operation Range (NOR, a generally submitted parameter value [37]). That is, optimizations within this design space should already be permissible to change with minimal notification to the authorities (of course, provided characterization has proven this range acceptable [37]). In summary, CPV should seek optimizations and improvements that may be rapidly implemented with minimal regulatory oversight.

Nonetheless, this is rarely done in practice, as the CPV plan is typically only reviewed at intervals conducive to analytics and statistical analysis. That is, results are only interpreted after sufficient results have been obtained to provide adequate power for an augmented analysis [21]. While periodic updating of the QbD deliverables is in any case essential for regulatory reporting, this only occurs long after the batch in question could be improved.

Optimizations of individual batches, within the submitted design space and linked to drug substance quality, should ideally be available for immediate implementation.

More generally, process knowledge (for use in both risk mitigation and process optimization) is not reliably placed in a central repository at rapid intervals. ICHQ8 encourages the inclusion of all data-driven process knowledge to be included in the iterations of the CPV protocol. However, the frequency of these changes is not sufficiently fast – certainly not in real time – and additionally the data behind these proposed changes are not centralized in a format that can be easily assessed. The underlying reason stems from misaligned data structures and nomenclature between development and manufacturing, hybrid paper-electronic solutions which increase effort, and a general suspicion of combining data from sources with different levels of data quality [44] as described in 1.2.3. All this leads to a lack of timely updating of process knowledge and equally infrequent optimization and risk mitigation.

In summary, much progress has been made in establishing QbD as the governing principle for bioprocess knowledge and life cycle management. Nonetheless, under-explored process parameters, silo unit operation models, and challenges in creating holistic process models inhibit the realization of the original Quality-by-Design goals. This is most clearly manifested in the QbD concepts around control strategies, process knowledge, and real-time feedback loops. There is significant room for innovations that would better attain QbD goals. Improvement in these categories would see a system fully realized to attain the principles laid out in Juran's (and the regulators') vision.

1.3 Why Now

Enabling a critical review of the current QbD practices first required substantial implementation of basic concepts. And indeed, many advances have been made in recent years with regard to model-based approaches to development deliverables. Much of the improvements described in section 1.1 have been supported by the increasing presence of applied statisticians and data scientists in the marketplace and even more so by the availability of standard software tools [41]. However, as described above, advancing the QbD aims requires statistical and data management methods currently not well established in the industry. For example, most of the process knowledge is being held in silos per unit operation. And the current inability of standard software, to manage complex model chains has ultimately created a natural ceiling for the realization of concatenated unit operation models.

In recent years, however, this ceiling looks more likely to be breached. With the increasing openness to statistically powerful programming languages such as Python and R, customizable computational applied statistics have become more accessible to process engineers and scientists. These tools can significantly reduce the cost of development for latent variable detection algorithms and advanced regression analyses. Thus the barrier of entry for the exploration of unit operation model linkages across full processes has been lowered [41].

Moreover, the data itself has also only in recent years become available for such analysis. Data lakes, warehouses, and historians have become an increasing part of the IT architecture of biopharma companies. For the first time, these systems offer both access to comprehensive datasets as well as the ability to contextualize them in meaningful process order; all of which was nearly unobtainable a decade ago [44].

Finally, increasing computational power has rendered even standard laptop machines able to fit more complex models or to search more complex design spaces. Such computational power is required as more process parameters are being investigated and the degrees of freedom are increasing accordingly. Furthermore, once models are in place, extensive simulation procedures are now available, allowing the integrated propagation of error through random variation (and, implicitly, the acceptance of statistical uncertainty). Without the constraint of reserving super-computer time to perform analyses and simulations, the barrier of entry for design space exploration has been reduced[44].

With flexible programming frameworks available, computational power increased, and the acceptance of multivariate statistics heightened, the most daunting hurdles towards innovation become surmountable. It is a clear moment for a step forward that can answer the questions that most truly correspond to Juran's vision of quality.

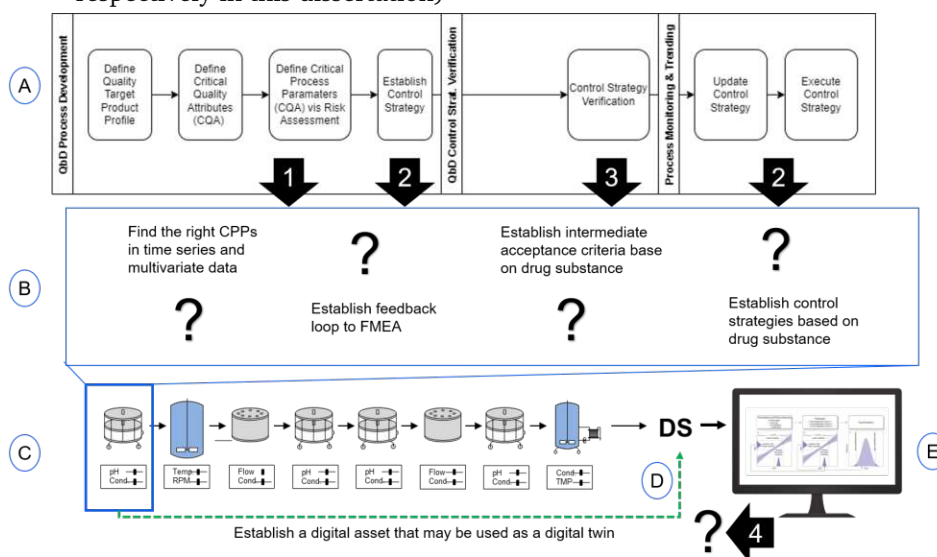
1.4 Goal of this Thesis

The goal of this thesis is to demonstrate that leveraging combined architectural and statistical innovations of an advanced IPM technology will lead not only to more technically rigorous and precise simulations, but will also support the establishment of QbD deliverables that are based on holistic process data, represent patient quality throughout the process, and run in real-time as an asset to a digital twin.

This demonstration seeks to include the following innovations:

- a novel latent feature extraction for more relevant CPP modeling (3.1).
- an improved, statistically more rigorous data model for the IPM (3.4).
- innovative IPM applications (i.e. Monte Carlo-based parameter sensitivity analyses) to directly produce QbD deliverables (3.2 and 3.3).
- a framework to deploy the IPM as a digital asset described, which would depict the first holistic bioprocess digital twin (3.4).

Fig. 1.3.: Levels of objectives within this dissertation:(A)Interpret Qbd methodology maximally to patient interest, (B)establish innovations at key data-driven points in the Qbd methodology (C) per unit operation, (D) based on results at drug substance, which are key to patient safety, and (E) produce a deployable digital twin to maintain all the above. The key innovations correspond to manuscripts 1-4, respectively in this dissertation)



References

- [1] ICH Q8 (R2). *Pharmaceutical Development Q8 (R2)*. 2009.
- [2] ICH Q11 (Step 4). *Development and Manufacture of Drug Substances (Chemical Entities and Biotechnological/Biological Entities) - Q11 - Step 4*. 2012. URL: http://sh.st/st/787f28ed3e745c14417e4aec27303038/www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q11/Q11_Step_4.pdf (visited on 10/18/2016).
- [3] “A-Mab: A Case Study in Bioprocess Development”. In: (), p. 278.
- [4] Vinzenz Abt et al. “Model-Based Tools for Optimal Experiments in Bioprocess Engineering”. In: *Current Opinion in Chemical Engineering* 22 (2018). Biotechnology and bioprocess engineering, pp. 244–252. ISSN: 2211-3398. DOI: 10.1016/j.coche.2018.11.007. URL: <https://www.sciencedirect.com/science/article/pii/S221133981830056X>.
- [5] Joseph S. Alford. “Bioprocess Control: Advances and Challenges”. In: *Computers & Chemical Engineering* 30.10-12 (Sept. 2006), pp. 1464–1475. ISSN: 00981354. DOI: 10.1016/j.compchemeng.2006.05.039. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0098135406001475> (visited on 09/12/2022).
- [6] Deb Autor et al. “PDA Points to Consider: Best Practices for Document/Data Management and Control and Preparing for Data Integrity Inspections”. In: *PDA Journal of Pharmaceutical Science and Technology* 72.3 (2018), pp. 332–337. ISSN: 1079-7440, 1948-2124. DOI: 10.5731/pdajpst.2018.008573. URL: <http://journal.pda.org/lookup/doi/10.5731/pdajpst.2018.008573> (visited on 09/27/2022).
- [7] Daniel Borchert et al. “Quantitative CPP Evaluation from Risk Assessment Using Integrated Process Modeling”. In: *Bioengineering* 6.4 (2019), p. 16. DOI: 10.3390/bioengineering6040114.
- [8] M. Boyer et al. “A Roadmap for the Implementation of Continued Process Verification”. In: *PDA Journal of Pharmaceutical Science and Technology* 70.3 (May 1, 2016), pp. 282–292. ISSN: 1079-7440. DOI: 10.5731/pdajpst.2015.006395. URL: <http://journal.pda.org/cgi/doi/10.5731/pdajpst.2015.006395> (visited on 09/15/2022).
- [9] Williams Calvert. “PHARMACEUTICAL RESEARCH AND MANUFACTURERS OF AMERICAD’OOD AND DRUG ADMINISTRATION COOPERATION IN BUILDING CAPABILITIES FOR ELECTRONIC SUBMISSION OF POSTMARKETING SAFETY DATA”. In: (), p. 7.
- [10] Jessica Carmen et al. “Developing Assays to Address Identity, Potency, Purity and Safety: Cell Characterization in Cell Therapy Process Development”. In: *Regenerative Medicine* 7.1 (Jan. 2012), pp. 85–100. ISSN: 1746-0751, 1746-076X. DOI: 10.2217/rme.11.105. URL: <https://www.futuremedicine.com/doi/10.2217/rme.11.105> (visited on 09/14/2022).

- [11] Salim Charaniya, Wei-Shou Hu, and George Karypis. "Mining Bioprocess Data: Opportunities and Challenges". In: *Trends in Biotechnology* 26.12 (Dec. 2008), pp. 690–699. ISSN: 01677799. DOI: 10.1016/j.tibtech.2008.09.003. URL: <https://linkinghub.elsevier.com/retrieve/pii/S016777990800245X> (visited on 09/27/2022).
- [12] Pauline M. Doran. *Bioprocess Engineering Principles*. 2nd ed. Previous ed.: 1995. Amsterdam ; Boston: Elsevier/Academic Press, 2013. 919 pp. ISBN: 978-0-12-220851-5.
- [13] Fiorenzo Franceschini and Maurizio Galetto. "A New Approach for Evaluation of Risk Priorities of Failure Modes in FMEA". In: *International Journal of Production Research* 39.13 (Jan. 2001), pp. 2991–3002. ISSN: 0020-7543, 1366-588X. DOI: 10.1080/00207540110056162. URL: <http://www.tandfonline.com/doi/abs/10.1080/00207540110056162> (visited on 05/17/2019).
- [14] Helena Bigares Grangeia et al. "Quality by Design in Pharmaceutical Manufacturing: A Systematic Review of Current Status, Challenges and Future Perspectives". In: *European Journal of Pharmaceutics and Biopharmaceutics* 147 (Feb. 2020), pp. 19–37. ISSN: 09396411. DOI: 10.1016/j.ejpb.2019.12.007. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0939641119313189> (visited on 11/10/2022).
- [15] "ICH Guideline Q10 on Pharmaceutical Quality System - Step 5". In: (), p. 20.
- [16] "ICH Guideline Q9 on Quality Risk Management". In: (), p. 20.
- [17] Feroz Jameel et al., eds. *Quality by Design for Biopharmaceutical Drug Product Development*. Vol. 18. AAPS Advances in the Pharmaceutical Sciences Series. New York, NY: Springer New York, 2015. ISBN: 978-1-4939-2315-1 978-1-4939-2316-8. DOI: 10.1007/978-1-4939-2316-8. URL: <http://link.springer.com/10.1007/978-1-4939-2316-8> (visited on 09/12/2022).
- [18] Angela Faustino Jozala et al. "Biopharmaceuticals from Microorganisms: From Production to Purification". In: *Brazilian Journal of Microbiology* 47 (Dec. 2016), pp. 51–63. ISSN: 15178382. DOI: 10.1016/j.bjm.2016.10.007. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1517838216310413> (visited on 11/10/2022).
- [19] J.M. Juran. *Juran on Quality by Design: The New Steps for Planning Quality into Goods and Services*. Free Press, 1992. URL: <https://books.google.de/books?id=b48uzwECAAJ>.
- [20] Brian Kelley, Mary Cromwell, and Joe Jerkins. "Integration of QbD Risk Assessment Tools and Overall Risk Management". In: *Biologicals* 44.5 (Sept. 2016), pp. 341–351. ISSN: 10451056. DOI: 10.1016/j.biologicals.2016.06.001. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1045105616300331> (visited on 11/10/2022).
- [21] Eun Ji Kim et al. "Process Analytical Technology Tools for Monitoring Pharmaceutical Unit Operations: A Control Strategy for Continuous Process Verification". In: *Pharmaceutics* 13.6 (June 21, 2021), p. 919. ISSN: 1999-4923. DOI: 10.3390/pharmaceutics13060919. URL: <https://www.mdpi.com/1999-4923/13/6/919> (visited on 11/11/2022).
- [22] Lisa M. LaVange. "Statistics at FDA: Reflections on the Past Six Years". In: *Statistics in Biopharmaceutical Research* 11.1 (Jan. 2, 2019), pp. 1–12. ISSN: 1946-6315. DOI: 10.1080/19466315.2019.1571322. URL: <https://www.tandfonline.com/doi/full/10.1080/19466315.2019.1571322> (visited on 11/10/2022).

- [23] Fredric J. Lim, Jagannathan Sundaram, and Alavattam Sreedhara. “Application of Quality by Design Principles to the Drug Product Technology Transfer Process”. In: *Quality by Design for Biopharmaceutical Drug Product Development*. Ed. by Feroz Jameel et al. Vol. 18. AAPS Advances in the Pharmaceutical Sciences Series. New York, NY: Springer New York, 2015, pp. 661–692. ISBN: 978-1-4939-2315-1 978-1-4939-2316-8. DOI: 10.1007/978-1-4939-2316-8_27. URL: http://link.springer.com/10.1007/978-1-4939-2316-8_27 (visited on 09/14/2022).
- [24] Hu-Chen Liu et al. “A Novel Approach for Failure Mode and Effects Analysis Using Combination Weighting and Fuzzy VIKOR Method”. In: *Applied Soft Computing* 28 (Mar. 2015), pp. 579–588. ISSN: 15684946. DOI: 10.1016/j.asoc.2014.11.036. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1568494614005985> (visited on 05/11/2019).
- [25] Hu-Chen Liu et al. “Improving Risk Evaluation in FMEA With Cloud Model and Hierarchical TOPSIS Method”. In: *IEEE Transactions on Fuzzy Systems* 27.1 (Jan. 2019), pp. 84–95. ISSN: 1063-6706, 1941-0034. DOI: 10.1109/TFUZZ.2018.2861719. URL: <https://ieeexplore.ieee.org/document/8423656/> (visited on 05/11/2019).
- [26] Francesca Luciani et al. “Implementing Quality by Design for Biotech Products: Are Regulators on Track?” In: *mAbs* 7.3 (May 4, 2015). meiner ansicht nach zu viel bullshit bingo, pp. 451–455. ISSN: 1942-0862, 1942-0870. DOI: 10.1080/19420862.2015.1023058. URL: <http://www.tandfonline.com/doi/full/10.1080/19420862.2015.1023058> (visited on 10/07/2015).
- [27] Nirupaplava Metta et al. “Dynamic Flowsheet Model Development and Sensitivity Analysis of a Continuous Pharmaceutical Tablet Manufacturing Process Using the Wet Granulation Route”. In: *Processes* 7.4 (Apr. 24, 2019), p. 234. ISSN: 2227-9717. DOI: 10.3390/pr7040234. URL: <https://www.mdpi.com/2227-9717/7/4/234> (visited on 07/09/2022).
- [28] Johannes Möller et al. “Model-Assisted Design of Experiments as a Concept for Knowledge-Based Bioprocess Development”. In: *Bioprocess and Biosystems Engineering* (Feb. 26, 2019). ISSN: 1615-7591, 1615-7605. DOI: 10.1007/s00449-019-02089-7. URL: <http://link.springer.com/10.1007/s00449-019-02089-7> (visited on 03/25/2019).
- [29] Justina A. Molzon. “The Value and Benefits of the International Conference on Harmonisation (ICH) to Drug Regulatory Authorities: Advancing Harmonization for Better Public Health”. In: *Global Approach in Safety Testing*. Ed. by Jan Willem van der Laan and Joseph J. DeGeorge. Vol. 5. AAPS Advances in the Pharmaceutical Sciences Series. New York, NY: Springer New York, 2013, pp. 23–27. ISBN: 978-1-4614-5949-1 978-1-4614-5950-7. DOI: 10.1007/978-1-4614-5950-7_3. URL: http://link.springer.com/10.1007/978-1-4614-5950-7_3 (visited on 11/06/2022).
- [30] Liliana Montano Herrera et al. “Holistic Process Models: A Bayesian Predictive Ensemble Method for Single and Coupled Unit Operation Models”. In: *Processes* 10.4 (Mar. 29, 2022), p. 662. ISSN: 2227-9717. DOI: 10.3390/pr10040662. URL: <https://www.mdpi.com/2227-9717/10/4/662> (visited on 05/18/2022).
- [31] Douglas C Montgomery. *Design and Analysis of Experiments*. 9th ed.
- [32] Douglas C. Montgomery. *Statistical Quality Control*. 7th ed. Wiley, 1991.

- [33] Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci. *Introduction to Time Series Analysis and Forecasting*. 2nd ed. Wiley, 1976. ISBN: 978-1-118-74511-3.
- [34] Harini Narayanan et al. “Bioprocessing in the Digital Age: The Role of Process Models”. In: *Biotechnology Journal* 15.1 (Jan. 2020), p. 1900172. ISSN: 1860-6768, 1860-7314. DOI: 10.1002/biot.201900172. URL: <https://onlinelibrary.wiley.com/doi/10.1002/biot.201900172> (visited on 09/12/2022).
- [35] Thomas Oberleitner et al. “Holistic Design of Experiments Using an Integrated Process Model”. In: *Bioengineering* 9.11 (Nov. 3, 2022), p. 643. ISSN: 2306-5354. DOI: 10.3390/bioengineering9110643. URL: <https://www.mdpi.com/2306-5354/9/11/643> (visited on 11/13/2022).
- [36] “Process Validation: General Principles and Practices”. In: (), p. 22.
- [37] “Questions and Answers: Improving the Understanding of NORs, PARs, DS_p and Normal Variability of Process Parameters”. In: *Questions and answers* (), p. 4.
- [38] Anurag S. Rathore, Anshuman Bansal, and Jaspinder Hans. “Knowledge Management and Process Monitoring of Pharmaceutical Processes in the Quality by Design Paradigm”. In: *Measurement, Monitoring, Modelling and Control of Bioprocesses*. Ed. by Carl-Fredrik Mandenius and Nigel J Titchener-Hooker. Vol. 132. Advances in Biochemical Engineering/Biotechnology. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 217–247. ISBN: 978-3-642-36837-0 978-3-642-36838-7. DOI: 10.1007/10_2012_172. URL: http://link.springer.com/10.1007/10_2012_172 (visited on 11/06/2022).
- [39] Anurag S. Rathore and Helen Winkle. “Quality by Design for Biopharmaceuticals”. In: *Nature biotechnology* 27.1 (2009), pp. 26–34.
- [40] Bryan S. Riley and Xuhong Li. “Quality by Design and Process Analytical Technology for Sterile Products—Where Are We Now?” In: *AAPS PharmSciTech* 12.1 (Mar. 2011), pp. 114–118. ISSN: 1530-9932. DOI: 10.1208/s12249-010-9566-x. URL: <http://link.springer.com/10.1208/s12249-010-9566-x> (visited on 09/12/2022).
- [41] Michael I. Sadowski, Chris Grant, and Tim S. Fell. “Harnessing QbD, Programming Languages, and Automation for Reproducible Biology”. In: *Trends in Biotechnology* 34.3 (Mar. 2016), pp. 214–227. ISSN: 01677799. DOI: 10.1016/j.tibtech.2015.11.006. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0167779915002462> (visited on 09/15/2022).
- [42] Arash Shahin. “Integration of FMEA and the Kano Model: An Exploratory Examination”. In: *International Journal of Quality & Reliability Management* 21.7 (Sept. 2004), pp. 731–746. ISSN: 0265-671X. DOI: 10.1108/02656710410549082. URL: <https://www.emeraldinsight.com/doi/10.1108/02656710410549082> (visited on 05/11/2019).
- [43] Ronald D. Snee. “Quality by Design: Building Quality into Products and Processes”. In: *Nonclinical Statistics for Pharmaceutical and Biotechnology Industries*. Ed. by Lanju Zhang. Statistics for Biology and Health. Cham: Springer International Publishing, 2016, pp. 461–499. ISBN: 978-3-319-23557-8 978-3-319-23558-5. DOI: 10.1007/978-3-319-23558-5_18. URL: http://link.springer.com/10.1007/978-3-319-23558-5_18 (visited on 09/12/2022).

- [44]Valentin Steinwandter, Daniel Borchert, and Christoph Herwig. “Data Science Tools and Applications on the Way to Pharma 4.0”. In: *Drug Discovery Today* (June 14, 2019). ISSN: 1359-6446. DOI: 10.1016/j.drudis.2019.06.005. URL: <http://www.sciencedirect.com/science/article/pii/S1359644618305324>.
- [45]Susan J. Schniepp. “Investigation Timeliness vs. Thoroughness_ Finding the Right Balance.Pdf”. In: *BioPharm International* 31.12 (Dec. 12, 2018), pp. 49–50.
- [46]Judith P. ter Horst et al. “Implementation of Quality by Design (QbD) Principles in Regulatory Dossiers of Medicinal Products in the European Union (EU) Between 2014 and 2019”. In: *Therapeutic Innovation & Regulatory Science* 55.3 (May 2021), pp. 583–590. ISSN: 2168-4790, 2168-4804. DOI: 10.1007/s43441-020-00254-9. URL: <http://link.springer.com/10.1007/s43441-020-00254-9> (visited on 09/12/2022).
- [47]Apostolos Tsopanoglou and Ioscani Jiménez del Val. “Moving towards an Era of Hybrid Modelling: Advantages and Challenges of Coupling Mechanistic and Data-Driven Models for Upstream Pharmaceutical Bioprocesses”. In: *Current Opinion in Chemical Engineering* 32 (June 2021), p. 100691. ISSN: 22113398. DOI: 10.1016/j.coche.2021.100691. URL: <https://linkinghub.elsevier.com/retrieve/pii/S221133982100023X> (visited on 09/14/2022).
- [48]James Vesper. “What Are Risk Appetite & Risk Tolerance In Pharma & Medical Devices?” In: (Feb. 23, 2022), p. 5. URL: <https://www.pharmaceuticalonline.com/doc/what-are-risk-appetite-risk-tolerance-in-pharma-medical-devices-0001>.
- [49]Tim Voigt, Martin Kohlhasse, and Oliver Nelles. “Incremental DoE and Modeling Methodology with Gaussian Process Regression: An Industrially Applicable Approach to Incorporate Expert Knowledge”. In: *Mathematics* 9.19 (Oct. 4, 2021), p. 2479. ISSN: 2227-7390. DOI: 10.3390/math9192479. URL: <https://www.mdpi.com/2227-7390/9/19/2479> (visited on 11/10/2022).
- [50]Mark F Witcher. “Failure Mode and Effect Analysis (FMEA) as a Quality by Design (QbD) Tool for Managing Biopharmaceutical Product Development and Manufacturing Risks”. In: (), p. 9.
- [51]Lawrence X. Yu et al. “Understanding Pharmaceutical Quality by Design”. In: *The AAPS Journal* 16.4 (July 2014), pp. 771–783. ISSN: 1550-7416. DOI: 10.1208/s12248-014-9598-3. URL: <http://link.springer.com/10.1208/s12248-014-9598-3> (visited on 08/04/2022).
- [52]Thomas Zahel et al. “Integrated Process Modeling—A Process Validation Life Cycle Companion”. In: *Bioengineering* 4.4 (Oct. 17, 2017), p. 86. DOI: 10.3390/bioengineering4040086. URL: <http://www.mdpi.com/2306-5354/4/4/86> (visited on 10/24/2017).
- [53]Jesús Zurdo et al. “Early Implementation of QbD in Biopharmaceutical Development: A Practical Example”. In: *BioMed Research International* 2015 (2015), pp. 1–19. ISSN: 2314-6133, 2314-6141. DOI: 10.1155/2015/605427. URL: <http://www.hindawi.com/journals/bmri/2015/605427/> (visited on 08/01/2015).

Part II

Results

Results, Findings & Achievements

“*What matters, however, are not so much the individual bits, but the successive patterns into which you arrange them, then break them up and rearrange them.*”

— Arthur Koestler

“*Constant systems of chance do exist in nature*

— Walter Shewhart
-Postulate 2-

2.1 Thesis Structure

This thesis is comprised of four peer-reviewed and published manuscripts. The following chapter will shortly describe the findings and achievements of each of the manuscripts, and then subsequently in section chapter 3 the published manuscripts are included in full.

Manuscript Title	Authors	Status	Date	Reference page
Multivariate Monitoring Workflow for Formulation, Fill and Finish Processes	Pretzner Taylor Dorozinski Dekner Liebminger Herwig	Published	3. June 2020	32
Integrated Process Model Applications Linking Bioprocess Development to Quality by Design Milestones	Taylor Marschall Kunzelmann Richter Rudolph Vajda Presser Zahel Studts Herwig	Published	24. October 2021	48
Specification-Driven Acceptance Criteria for validation of biopharmaceutical processes	Marschall Taylor Zahel Kunzelmann Wiedemann Presser Studts Herwig	Published	23. September 2022	66
Architectural and Technological Improvements to Integrated Bioprocess Models towards Real-Time Applications	Taylor Pretzner Zahel Herwig	Published	9. October 2022	83

2.2 Findings and Achievements

2.2.1 Multivariate Workflow for Latent Parameters

State of the Art & Problem Statement

In the course of a project to establish a CPV workflow in the Fill & Finish (FF) process segment of a biopharmaceutical product on the market, several challenges were noted pertaining to establishing QbD relationships between CPPs and CQAs. Firstly, a lyophilization and filtration step series produced data mostly found in multiple time series and multidimensional data. This led manufacturing scientists to, at best, only leverage univariate assessments of unfolded data, which may overlook domain-relevant CPP features. This is partly due to the lack of procedures in the literature to explore these latent variables within this segment of bioprocessing. Secondly, this exploration was generally given less priority, due to the focus on drug substance specifications, which are evaluated one step prior to FF. That is, operators had invested less time in these unit operations, as long as the primary CQAs met specification at drug substance and then, in reduced form, at packaging several steps downstream. The likelihood may be high that conform specifications at drug substance will remain so through

finalization of drug product, but nonetheless does not conform to the intent of the QbD methodologies.

Findings

Within the project, we developed an innovative domain-relevant procedure to automate the exploration of FF multi-dimensional and time series data starting from the data alignment phase. While contextualizing the different data sources, a single batch object was created and, using process expertise, a series of dynamic phase procedures were described that allow more relevant extraction of potentially critical parameters for further exploration. Finally, a robust PCA was established to check for further sources within multivariate variation. Once implemented as a single procedure, the outcome is a list of newly defined potentially critical parameters that can be immediately trended, monitored, and further investigated at a multivariate level.

Christopher Taylor's Contribution

As project lead, I supervised the development of the algorithms and procedures, managed the case study, and contributed to the writing and review of the publication.

Publication

Pretzner, Barbara, Christopher Taylor, Filip Dorozinski, Michael Dekner, Andreas Liebming, and Christoph Herwig. "Multivariate Monitoring Workflow for Formulation, Fill and Finish Processes." *Bioengineering* 7, no. 2 (June 3, 2020): 50. <https://doi.org/10.3390/bioengineering7020050>.

2.2.2 QbD Milestones Applications via IPM

State of the Art & Problem Statement

Despite the increasingly common application of process models towards development goals, most models are maximally applied to a singular unit operation. Furthermore, their conclusions are not concatenated out to the drug substance specifications. Two QbD goals are of particular interest here: the severity ranking of a risk assessment and the establishment of a control strategy. The severity ranking of a given CPP on a CQA is usually evaluated solely by process expertise. This assessment should be evaluated on the parameter deviation's final patient impact (i.e. at specifications) but in practice is performed based on their influence in the given unit operation. Control strategies define the range within which a CPP does not have a critical impact on a CQA. This range is submitted as a proven acceptable range (PAR) to the authorities as the acceptable manufacturing range. As with the risk assessment, the PAR is nearly always established within the context of a single UO.

Findings

Leveraging the original IPM technology, two innovative procedures were established to propagate conclusions at the unit operation out to drug substance. For the control strategy, a

parameter sensitivity analysis (PSA) was successfully proposed that uses Monte Carlo simulations to survey the whole of a given CPPs investigated range, resulting in the relationship of the Out-of-Specification (OOS) rate across the full CPPs range. This is used to set the PAR at the point at which an unacceptable percentage of lots may not conform to specifications. The risk assessment builds on the parameter sensitivity analysis with an additional linearization technique to compare slopes of OOS percentages to a customizable rubric of risk assessment severity scores. After adopting the rubric to the company risk assessment procedure, the severity rankings can be directly calculated from the IPM PSA.

Thus, both the PAR and the risk assessment severity rankings will be based on propagated impact at drug substance. This procedure can be very simply automated via the IPM.

Christopher Taylor's Contribution

I designed and implemented the risk assessment PSA linearization technique and conducted the case study. I wrote the manuscript with inputs from the co-authors. As project lead, I supervised the development and implementation of all described innovations.

Publication

Taylor, Christopher, Lukas Marschall, Marco Kunzelmann, Michael Richter, Frederik Rudolph, Judith Vajda, Beate Presser, Thomas Zahel, Joey Studts, and Christoph Herwig. "Integrated Process Model Applications Linking Bioprocess Development to Quality by Design Milestones." *Bioengineering* 8, no. 11 (October 24, 2021): 156. <https://doi.org/10.3390/bioengineering8110156>.

2.2.3 Specification-Driven Acceptance Criteris

State of the Art & Problem Statement

Control strategies at the individual unit operation cannot use drug substance specifications directly as intermediate acceptance criteria. For one, this may not be an appropriate range for a given CQA at that unit operation. Secondly, this effectively leads back to the testing-into quality issue that QbD tends to solve. Nonetheless, CQA results are generated at the individual unit operation and must be assessed against some threshold to ensure that they are keeping with the expected quality. Classically, these acceptance criteria are generated by 3 multiples of the existing data set's standard deviation without any further connection to downstream steps. Standard methods to create this connection are based on spiking studies, which are difficult to perform and run into sample matrix issues. Monte Carlo simulations have also been proposed to find worst-case scenarios given the known functional relationship. However, none of these approaches are able to directly link intermediate acceptance criteria to the drug substance specifications.

Findings

After establishing a definition of the term 'intermediate acceptance criteria' based on regulatory guidelines, we explored and discussed a QbD-based redefinition of the wording

'intended use.' Based on this definition, we are able to establish a procedure that leverages the IPM and MC simulations to assay the limits of a given load material on subsequent unit operations. A range of feasible load material values are defined per unit operation and the full range is assessed by simulations propagated out to drug substance, generating an OOS percentage. Intermediate acceptance criteria are drawn from the point at which an unacceptable percentage of OOS is simulated. This procedure can be easily automated and acceptance criteria can be quickly established across all unit operations; all directly linked to the process capability of meeting specifications.

Christopher Taylor's Contribution

I designed the acceptance criteria functionality, including the definition of acceptance criteria leading to the defined procedure. I performed all statistical modeling, and drafted the manuscript. Lukas Marschall and I contributed equally to this research.

Publication

Marschall, Lukas, Christopher Taylor, Thomas Zahel, Marco Kunzelmann, Alexander Wiedemann, Beate Presser, Joey Studts, and Christoph Herwig. "Specification-Driven Acceptance Criteria for Validation of Biopharmaceutical Processes." *Frontiers in Bioengineering and Biotechnology* 10 (September 23, 2022): 1010583. <https://doi.org/10.3389/fbioe.2022.1010583>.

2.2.4 Integrated Process Model 2.0

State of the Art & Problem Statement

Numerous advances have now been made in linking unit operation models and applying them to various outcomes of process development. Nonetheless, most of these are still under-exploring the definitions proposed by QbD and the regulatory guidelines. Reasons for this include statistical issues of combining data sets, overlooking variance in error propagation, and insufficient ability to link results throughout the process.

The IPM itself in its originally proposed form has shown over time to require innovation. The combination of two models from two regressions is mathematically suboptimal and overlooks scale effects. Other scale-dependent effects were not described in the original publication. And in the case of simulations that run outside of the explored data range, no extrapolation mechanism has been implemented (even a very conservative one).

Furthermore, none of these approaches has yet proposed a full real-time digital asset for potential implementation as a digital twin. The final step of producing holistic models that correspond truly to the regulatory requirements is to insert these models into the manufacturing architecture as a companion to life cycle management in real time.

Findings

An IPM technology was described that was based on the original IPM concept of linking UOs via load and pool CQA values. However, the data model was simplified, describing rather a

single matrix and regression model rather than two separate models. This model is more robust and addresses scale differences directly. Procedures for scale-dependent variables were established. A conservative extrapolation procedure was developed, which, despite the challenges of extrapolation in a data-driven environment, works within the risk management systems described by the authorities and allows baseline assumptions to be made regarding the control strategies.

Finally, a real-time framework is described for the IPM that allows model establishment, feedback loop, and instantaneous simulation of new conditions within a database containing relevant manufacturing data. This should provide a plausible digital asset for a bioprocess digital twin.

Christopher Taylor's Contribution

As project lead, I lead the conceptualization of the innovations, developed the scale-dependent variable procedure, described a real-time use case, performed the case study, and drafted the manuscript.

Publication

Taylor, Christopher, Barbara Pretzner, Thomas Zahel, and Christoph Herwig. 2022. "Architectural and Technological Improvements to Integrated Bioprocess Models towards Real-Time Applications" *Bioengineering* 9, no. 10 (October 9, 2022): 534. <https://doi.org/10.3390/bioengineering9100534>


Manuscripts

3

3.1 Multivariate Monitoring Workflow for Formulation Fill and Finish Processes

Article

Multivariate Monitoring Workflow for Formulation, Fill and Finish Processes

Barbara Pretzner ^{1,2,*}, Christopher Taylor ^{1,2}, Filip Dorozinski ³, Michael Dekner ³,
Andreas Liebming ⁴ and Christoph Herwig ^{1,3} 

¹ Exputec GmbH, Mariahilfer Straße 88A/1/9, 1070 Vienna, Austria; christopher.taylor@exputec.com (C.T.); christoph.herwig@tuwien.ac.at (C.H.)

² Research Area Biochemical Engineering, Vienna University of Technology, Gumpendorferstrasse 1a, 1060 Vienna, Austria

³ Department of Manufacturing Sciences, Takeda, 1070 Vienna, Austria; filip.dorozinski@takeda.com (F.D.); michael.dekner@takeda.com (M.D.)

⁴ Plasma Derived Therapies R&D, Takeda, 1070 Vienna, Austria; andreas.liebming@takeda.com

* Correspondence: barbara.pretzner@exputec.com

Received: 29 April 2020; Accepted: 1 June 2020; Published: 3 June 2020



Abstract: Process monitoring is a critical task in ensuring the consistent quality of the final drug product in biopharmaceutical formulation, fill, and finish (FFF) processes. Data generated during FFF monitoring includes multiple time series and high-dimensional data, which is typically investigated in a limited way and rarely examined with multivariate data analysis (MVDA) tools to optimally distinguish between normal and abnormal observations. Data alignment, data cleaning and correct feature extraction of time series of various FFF sources are resource-intensive tasks, but nonetheless they are crucial for further data analysis. Furthermore, most commercial statistical software programs offer only nonrobust MVDA, rendering the identification of multivariate outliers error-prone. To solve this issue, we aimed to develop a novel, automated, multivariate process monitoring workflow for FFF processes, which is able to robustly identify root causes in process-relevant FFF features. We demonstrate the successful implementation of algorithms capable of data alignment and cleaning of time-series data from various FFF data sources, followed by the interconnection of the time-series data with process-relevant phase settings, thus enabling the seamless extraction of process-relevant features. This workflow allows the introduction of efficient, high-dimensional monitoring in FFF for a daily work-routine as well as for continued process verification (CPV).

Keywords: multivariate monitoring; CPV; formulation; fill finish process; data science; time-series analysis; feature extraction

1. Introduction

In 2011, the Food and Drug Administration (FDA) published a guideline that emphasizes the importance of undertaking continued process verification (CPV) in biopharmaceutical manufacturing as an integral and final part of the process validation lifecycle [1] within the Quality by Design (QbD) approach. CPV ensures that the product quality and process performance stays in control throughout the commercial part of the product life cycle. The core element of a CPV plan is the control and monitoring strategy of certain critical process parameters (CPPs) and critical quality attributes (CQAs), as well as the method for analyzing the collected data. The FDA stresses that the collected data should be evaluated with appropriate statistical process control technology, but leaves the selection of a concrete monitoring strategy and statistical tools to the individual developer. In the biopharmaceutical industry, most manufacturers use simple out-of-specification or univariate trending charts to show

their control over their process [2,3]. While classic biopharma process segments, namely upstream and downstream processing resulting in the drug substance, typically have established CPV plans leading into validation, when it comes to advanced control technologies, formulation, fill, and finish (FFF) is often deprioritized [4]. Therefore, current and well-described CPV or monitoring plans for FFF are very difficult to find in the literature. This last unit operation is needed to turn the purified drug substance into a final dosage form, applicable for the market [5]. Freezing of the purified protein bulk, thawing of the bulk, formulation, sterile filtration, filling and on occasion lyophilization are the common steps within FFF to obtain a safe, stable, final product, ready to be transported. Although the FFF process is chemically and biologically more straightforward than a fermentation process, any variation in FFF can influence the stability, safety or final dosage form of the product [5,6].

Current process monitoring strategies in biopharmaceutical FFF steps are generally limited to a univariate assessment of two distinct and separate data sources:

- single-point data (called “feature data”) from intermediates or from Quality Database (QDB) testing (see Figure 1, State of the Art, QDB Data—Univariate Assessment). Examples: lyophilization duration, sterile filtration hold time, amount of various formulation buffer ingredients, etc.
- time-series data during the individual unit operations (see Figure 1, State of the Art, Lyophilization/Filtration Data—Univariate Assessment). Examples: online measurement of product temperature over process time, online measurement of pressure during lyophilization over process time, etc.

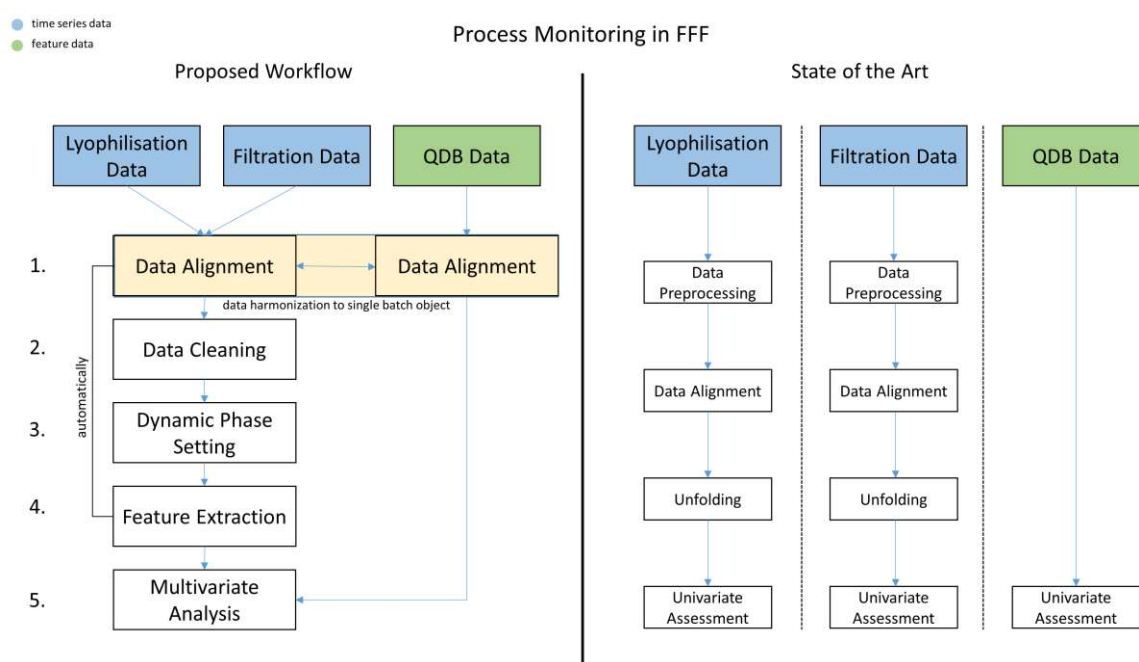


Figure 1. Comparison of workflows for process monitoring in formulation, fill, and finish (FFF) processes with respect to filtration, lyophilization and quality database (QDB) data. Compared to the current workflow, the proposed workflow examines the different FFF data sources in more detail, harmonizes the data sources to one single batch object, allowing the analysis of the data in a broader context, and uses their key feature data to perform multivariate analysis. All intermediate steps in the proposed workflow are automated and applicable on further FFF QDB data.

This univariate control strategy is often insufficient to uncover root cause variation related to interactions between these data types in a multivariate space, as D.C. Montgomery already stated in 1991 [7]. Moreover, time-series data can give an additional deep insight on how the process actually performs in certain states or phases [8].

Most FFF control strategies only use very rudimentarily extracted time-series features such as the length, maximum, or minimum value of time-series data, overlooking the information hidden in time-series patterns [9–11]. One method to analyze a time-series in more detail, is to unfold the signal along its time-points and then use the data as input for multivariate data analysis (MVDA) [12]. In fermentation or in chemical engineering, the in-depth analysis of time series is established by analyzing the process with Fourier transform near-infrared spectroscopy and evaluating the data in conjunction with principal component analysis (PCA) and partial least squares regression [13,14]. Another method is to identify certain patterns by dividing the signal into smaller phases to enable advanced feature extraction. The advantage here is that only process-relevant information will be further analyzed with MVDA [15].

Although the identification of patterns is easily performed by the human eye, the analysis of several time-series patterns can be extremely time intensive and failure-prone, if not done automatically [9]. Today's program languages, e.g., Python, offer a great variety of machine learning algorithms or neural networks (NN), which may be used for pattern recognition. Most of these data science tools demand a huge training dataset. In areas such as biology, genomics or biopharmaceutical manufacturing, it often occurs that the observed data holds a higher number of features p than number of observations N , also known as the $p \gg N$ problem. This high-dimensionality of the data often leads to problems when applying NN or machine learning algorithms [16,17].

Deeper insight to the process is not limited to the extraction of advanced features from the time-series data, but can be extended to the subsequent calculation of supplementary key performance indicators, which may also be subjected to MVDA [15]. Suarez-Zuluaga et al. showed in an upstream process case study how a basic, dynamic phase-setting algorithm, followed by key performance indicator extraction and MVDA accelerated the development of their process [18].

In order to establish MVDA for FFF in a CPV plan, the input dataset must be in a certain shape and must also be easily accessible, which is rarely the case in commercial manufacturing [19]. Usually, the data collected from the various monitoring equipment for individual unit operations of the FFF process is rarely stored in the same place and is often not aligned with each other, which results in a highly time-consuming task of establishing an analytical MVDA dataset.

This paper presents a novel, holistic, multivariate process monitoring strategy, combining the individual FFF data sources via time-series feature extraction using a dynamic phase-setting approach. The assessment of the power of this new method takes place on real biopharmaceutical manufacturing data and is compared to historical data evaluations that occurred based on traditional process monitoring strategies. The goal is herein to establish a multivariate, automated FFF process monitoring workflow, which uses all existing FFF data (time-series and QDB data), makes use of process-relevant time-series patterns, and is followed by robust principal component analysis (ROBPCA) to detect lots which perform atypically related to reference lots. This approach should ease and accelerate the identification of abnormal behavior within the FFF process and point to root causes for this via parameter loadings. The roadmap to realize this goal can be described in the following steps:

1. Assign the available data to the corresponding lots (see Figure 1, Proposed Workflow, Data Alignment).
2. Enhance the quality of information by reducing interference signals within the time series (see Figure 1, Proposed Workflow, Data Cleaning).
3. Identify process-relevant characteristics of the time-series pattern and leverage for further feature extraction (see Figure 1, Proposed Workflow, Dynamic Phase Setting).
4. Create an analytical data set based on the extracted features and combine with already available features (see Figure 1, Proposed Workflow, Feature Extraction).
5. Perform robust principal component analysis to assess the data set in step 6 (see Figure 1, Proposed Workflow, Multivariate Analysis).

2. Materials and Methods

2.1. Data

The analytical data set for our innovative CPV approach was derived from an industrial FFF process of a parenteral biopharmaceutical therapy, which is stored in a liquid phase. This data consisted of 58 lots and was compiled from three different FFF data sources, as shown in Table 1.

Table 1. Overview of the quality data, sterile filtration (SF) and lyophilization (LP) data sources, including their data content, abbreviation and unit.

Quality Data		Sterile Filtration			Lyophilization			
Data Type	Feature	Time-series			Time-series			
	Description	Abbr.	Description	Abbr.	Unit	Description	Abbr.	Unit
Monitored Output	CQA data of bulk drug substance (BDS)	CQA data	Temperature of product	SF1	(°C)	Inlet temperature	LP1	(°C)
	Time stamps of start and end of sterile filtration and lyophilization	Time stamps	Applied pressure	SF2	(bar)	Outlet temperature	LP2	(°C)
			Weight of unfiltered product	SF3	(kg)	Chamber vacuum 1	LP3	(bar)
						Chamber vacuum 2	LP4	(bar)
						Temperature of liquid nitrogen	LP5	(°C)
						Condenser pressure	LP6	(bar)
						Condenser vacuum	LP7	(bar)

2.2. Software

The commercially available software inCygnt® Web version 2019.08 (Exputec GmbH, Vienna, Austria) and Python 3.5 (Python Software Foundation, <https://www.python.org/>) was used for data preprocessing, algorithm development, and multivariate data analysis. The statistical software JMP® (SAS Institute, USA) was used for MVDA result comparison.

2.3. Statistical Methods

Robust principal component analysis (ROBPCA) [20] was performed for MVDA. In contrast to the conventional PCA [21], the ROBPCA is less influenced by outlying observations and can recover principal components of a data matrix even though its entries might be sparse to a certain extent. The ROBPCA analysis allows the evaluation of whether observations are more or less similar to each other in the multivariate space, by plotting the orthogonal distance against the score distance. A high value in the score distance means that the observation does obey the multivariate model, but certain variables have a higher or lower value compared to the average of the other observations. A high value in the orthogonal distance indicates that the observation does not follow the multivariate model and shows a different correlation. The contribution of the score and orthogonal distance of each observation allows the identification of which variables are responsible for the observed abnormality.

3. Results

3.1. Step 1: Data Alignment

The current state of data management within FFF describes the following procedure: measure various process parameters from all FFF process steps with different monitoring equipment and store the collected data on the monitoring equipment’s databases [19]. Interfaces, which enable us to harmonize and align the data between the different sources, without requiring major manual input, are rarely available. Some data harmonization usually takes place offline by manually adding certain CQA data to a quality database (QDB), which usually contains intermediate data from each FFF process step. However, the CQA data does not cover all the information, which is available within the time series.

This decentralized data management makes further multivariate data analysis effectively inaccessible. Furthermore, some systems—such as in this case the sterile filtration and lyophilization equipment—do not contextualize the collected data to any specific corresponding lot (i.e., what differentiates one lot to another is not defined within the data collection system), which makes the raw data impossible to be used for any further multivariate analysis.

In contrast to the state of the art, where every process signal is preprocessed and aligned separately (see Figure 1), we present an automated workflow, where every FFF data source is aligned and contextualized to a unique batch object within the inCyght database. By merging all available FFF data independently of their source or format into batch objects, a comprehensive insight of the data for each lot is given, which is a necessity to facilitate an automated multivariate CPV workflow.

To realize the harmonization of the various data origins, the data sources must be linked to each other. The most straightforward procedure is to use the lot name as linkage. However, as in the case of the sterile filtration data, no lot name was available, since the data was continuously recorded resulting in one continuous time series over months. In this case we used timestamps, stored in the QDB which provided information when each lot was filtrated, to contextualize the filtration data to the batch objects. We developed a robust interface in Python, which automatically uploads all FFF data to inCyght, as shown in Figure 2.

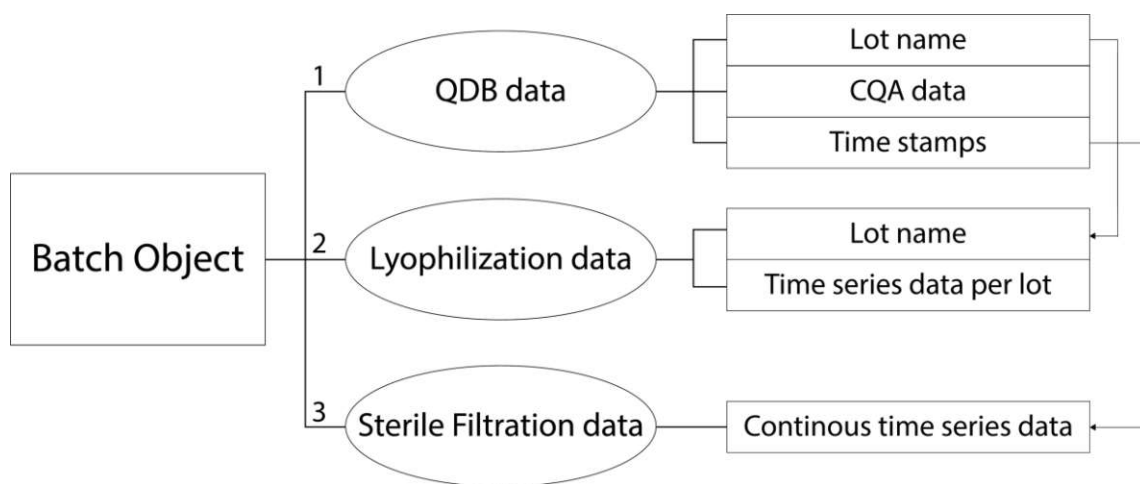


Figure 2. Scheme of the data alignment, showing the workflow and the individual contents of the data sources. All available data are stored in a batch object within the inCyght data base (IDB). The data alignment from various data sources to one batch object can.

3.2. Step 2: Data Cleaning

The outcome of any analysis, machine learning, or the phase-setting algorithms in the following chapters, is strongly influenced by the quality of the data [22]. Real-world data is never ideal from the analytical perspective and consists of the real signal and accompanying noise. The two main

sources for noise are random noise and interference signals. Random noise is usually introduced by measurement tools such as sensors, whereas interference signals are commonly caused by equipment failure or operator error. To enhance the quality of the data, the signal-to-noise ratio (SNR) must be maximized. A common tool to enhance the SNR is the application of filters such as median or Savitzky–Golay [23]. However, the application of the wrong filter might lead to loss of data quality. Furthermore, strong interference signals might not be reduced nor removed by the usage of filters, instead leading to false-positive alerts, if not removed before the MVDA. In order to preprocess the data correctly, it is important to understand which errors may exist and to what extent they might affect the data, using domain knowledge. Different data-cleaning algorithms for each data source were developed to enhance the quality of the data accordingly, as described in the following paragraphs.

The QDB data included only CQA data and time stamps (Table 1) and was not further preprocessed. The SNR of the lyophilization data was already sufficient, where further filter application would possibly lead to information loss.

Data from the sterile filtration included anomalies that did not affect the product quality, but might affect the results of the MVDA and therefore need to be removed. At the end of the filtration, the pressure increases, leading to a high peak at the end of SF2 (SF—sterile filtration), as shown in Figure 3A. Those high peaks are typical within the process, but not relevant for data analysis in this approach. Therefore, the data-cleaning algorithm was adjusted to remove the last slope from SF2, as shown in Figure 3B.

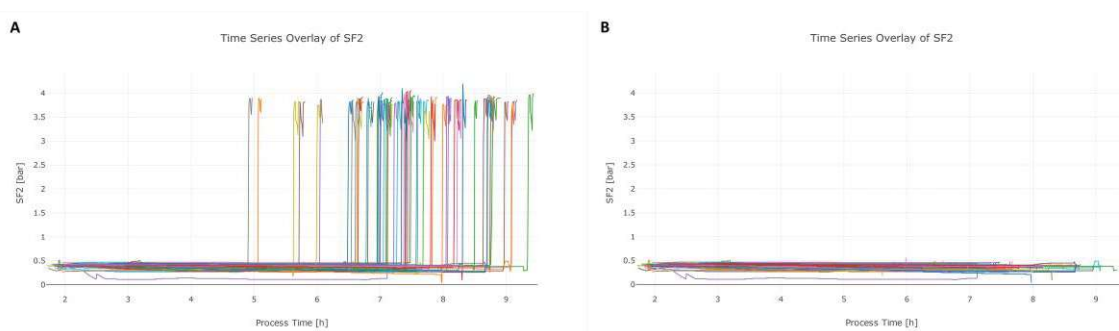


Figure 3. Data cleaning algorithm applied on signal SF2. The raw SF2 signals of various lots (different colors) are depicted in (A), whereas the high peak at the end of the process time is not relevant for monitoring and needs to be removed to enhance the performance for multivariate data analysis (MVDA). In (B), the corrected SF2 signal is depicted after the cleaning algorithm was applied.

SF3 is used to track the progress of the filtration and therefore should slowly decrease over time. However, SF3 showed pronounced peaks during the process, which were identified to be caused by manufacturing personnel stepping on the scale, as shown in Figure 4A. As these events are not linked to the manufacturing process, these occurrences have no impact to the quality of the product. However, these interferences have a negative influence on the data quality. Since the overall slope in SF3 can be expected to be very low, peaks with high slopes could be easily removed by the data-cleaning algorithm, as seen in Figure 4B.

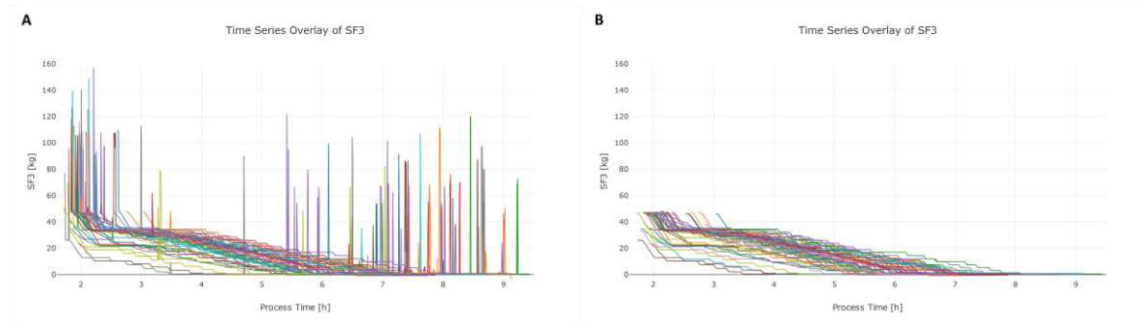


Figure 4. Data-cleaning algorithm applied on signal SF3. The raw SF3 signals of various lots (different colors) are depicted in (A), whereas the high peaks within the signal are caused due to stepping on the scale. These peaks have no further process-relevant information content and therefore need to be removed to enhance the performance for MVDA. In (B), the corrected SF3 signal is depicted after the cleaning algorithm was applied.

3.3. Step 3: Dynamic Phase Setting

3.3.1. General

Time-series data are frequently used for control charts or other univariate monitoring tools, whereas only the maximum or minimum value or the length of the signal are commonly used for further analysis in the biopharmaceutical industry. Although the pattern of a time series explains the behavior of a process in greater detail, it is usually overlooked when it comes to FFF monitoring plans.

Most statistical analysis and monitoring tools require a two-dimensional data set. Since time-series data is a three-dimensional data format (batches (N) \times variables (K) \times time (J)), the data has to be dimensionally rearranged into a two-way matrix structure; this process is also called “unfolding”. One possibility is to unfold the data based on every single time point, resulting in $N \cdot J \times K$ matrix. This kind of unfolding leads to a dataset with many features which are not necessarily process-relevant and might lead to a higher demand of data storage [24,25]. Another possibility is to separate the time series into phases, based on process expert knowledge, e.g., cooling or heating phases within the temperature signal. Certain features (e.g., median, mean, min, and max value) of this phase and signal can be subsequently extracted, which are more valuable for monitoring than time-dependent features. Therefore, we developed novel phase-setting algorithms, specific for FFF data, which enable automated phase setting in the data.

3.3.2. Step Signal Phase Settings Algorithms

FFF data often occur as rectangular or step signals. This is caused by the stepwise adjustment of the pH, pressure, or temperature within the FFF process, which has an immediate effect on the system. This results in time-series data, most of which consist only of plateaus and sharp slopes, which may be divided algorithmically into phases.

The starting point of a slope can be determined by searching for a value within the signal, where the difference of neighboring data is above 0 in the y-direction. The end point of the slope is reached, when the change of the neighboring points in the y-direction is 0.

$$slope_{start} = abs(y_t - y_{t+i}) > 0 \quad (1)$$

$$slope_{end} = abs(y_t - y_{t+i}) = 0 \quad (2)$$

Like any other signal, the stepwise signal might contain some noise, as stated in subchapter Step 2: Data Cleaning. This could result in misidentification of slopes, since the noise could randomly indicate a gradient or plateau. In this workflow, we used several methods to tackle this problem. First,

if the signal-to-noise ratio is estimable, a certain threshold for equation 1 and 2 can be chosen, instead of assuming a change of 0.

$$slope_{start} = abs(y_t - y_{t+i}) > threshold\ value \tag{3}$$

$$slope_{end} = abs(y_t - y_{t+i}) \leq threshold\ value \tag{4}$$

The threshold was set based on the standard deviation of the noise within the individual time-series data and was verified for practical relevance by the process experts. However, the noise might not always be estimable or unaffected by outliers, which again would result in the wrong phase setting. Such false identifications can be reduced by taking the duration of the slope and plateaus into account. If the minimum or maximum duration of the plateaus or slopes is assessable (derivable from FFF standard operation procedures or by consultation with the process expert), the current slope or plateau duration can be checked to determine if it meets the criterion of the expected length of the duration. If the criterion is not fulfilled, the current proposal for the end of the slope or plateau is discarded and the algorithm searches for the next best guess, as illustrated in Figure 5.

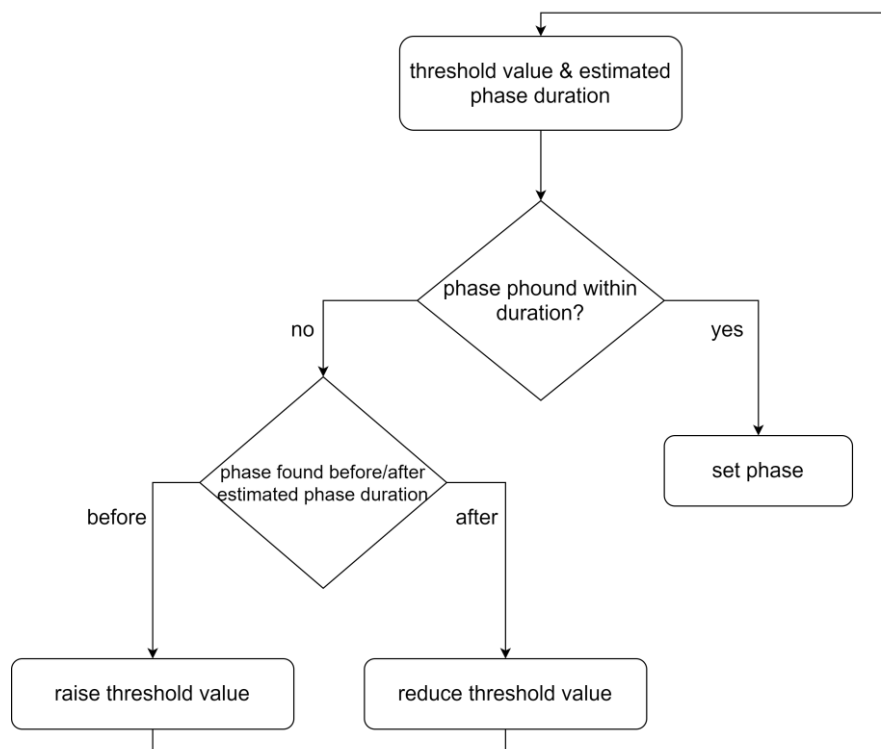


Figure 5. Flowchart of the step signal phase settings algorithm. The algorithm starts with an initial threshold value and an estimated phase duration range. If the algorithm detects a phase (slope or plateau) the algorithm stops and sets the phase accordingly. However, if the algorithm detects a phase before or after the estimated phase duration, the threshold value is raised or reduced, respectively. After that the algorithm starts again with the changed threshold value.

Based on this approach, 97% of the phases were correctly identified, which was verified by an expert from the FFF facility.

3.3.3. Intertwined Phase Settings Algorithms

Not all time-series data can be divided into slopes or plateaus only, or lead to process-relevant features. Depending on the observed signal, other states might be of interest, but cannot be

algorithmically divided into phases with only one input time series, since the pattern of interest results from the combination of several signals.

Thus, a principle advantage of having centralized and aligned data is that the information from the individual time-series signals can be combined to enable process-relevant phase setting. As an example, the process engineers wanted to monitor the LP4 signal from lyophilization in more detail, by dividing the signal into four phases, to identify any deviations within the lots, as seen in Figure 6.

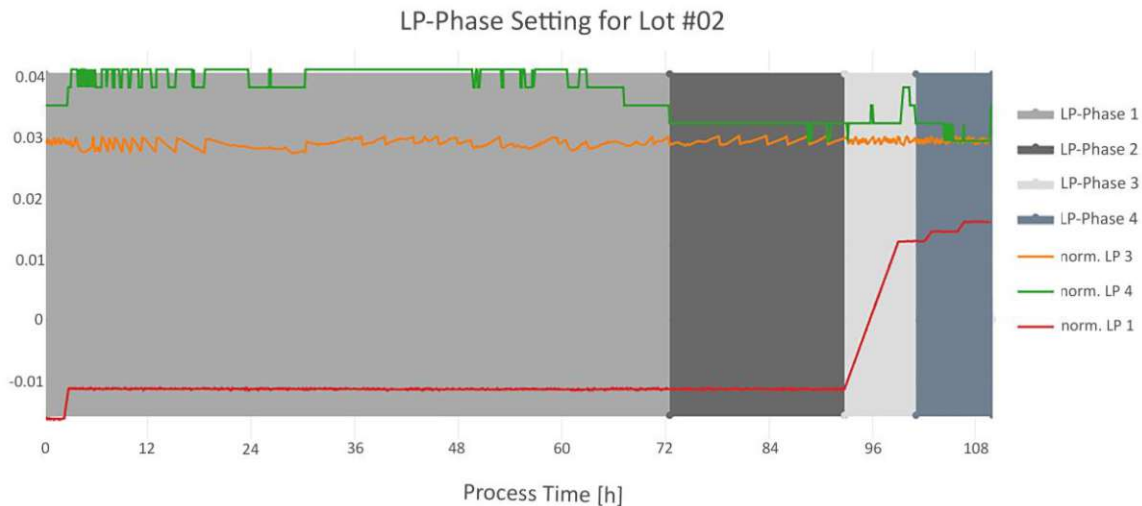


Figure 6. Example results for lot #2 of the intertwined phase setting method. The LP-Phases 1–4 are shown as filled rectangles in different grey colors. The normalized signal of LP3 is shown in orange, the normalized signal of LP4 is shown in green and the normalized signal of LP1 is shown in red.

For this approach the LP1, LP3 and LP4 signal (LP—lyophilization) were used. The detailed phase-setting conditions for this algorithm are described in Table 2.

Table 2. Condition description on how the four phases LP-Phase 1-4 were set. The phases were set accordingly to the “run order”.

Phase Name	Condition	Run Order
LP-Phase 1	Starts with the beginning of LP4 signal.	2
LP-Phase 2	Ends with the start of LP-Phase 2. First timestamp, where the difference between LP3 and LP4 is below 20%.	1
LP-Phase 3	Ends with the increasing slope of LP1. Starts with the end of LP-Phase 2. Ends when LP4 has the same value as at the beginning of LP-Phase 2, within certain time range.	3
LP-Phase 4	Starts with the end of LP-Phase 3. Ends with end of LP4.	4

3.4. Step 4: Feature Extraction

After the phases were set, the mean, maximum, and minimum value were extracted. Furthermore, the residuals of the slope and the standard deviation of the plateau phases were calculated. Moreover, the duration of the LP Phases described in the Section 3.3.3. Intertwined Phase Settings Algorithms were extracted.

All extracted features, as well as the features from the QDB data, were joined in a feature matrix, as schematically shown in Figure 7. The feature extraction resulted in 130 new variables, making a total in 252 variables (including 122 QDB features). Features without variance within the lots were removed from the feature matrix, resulting in 208 features per lot in total for further robust PCA analysis (see Supplementary Materials).

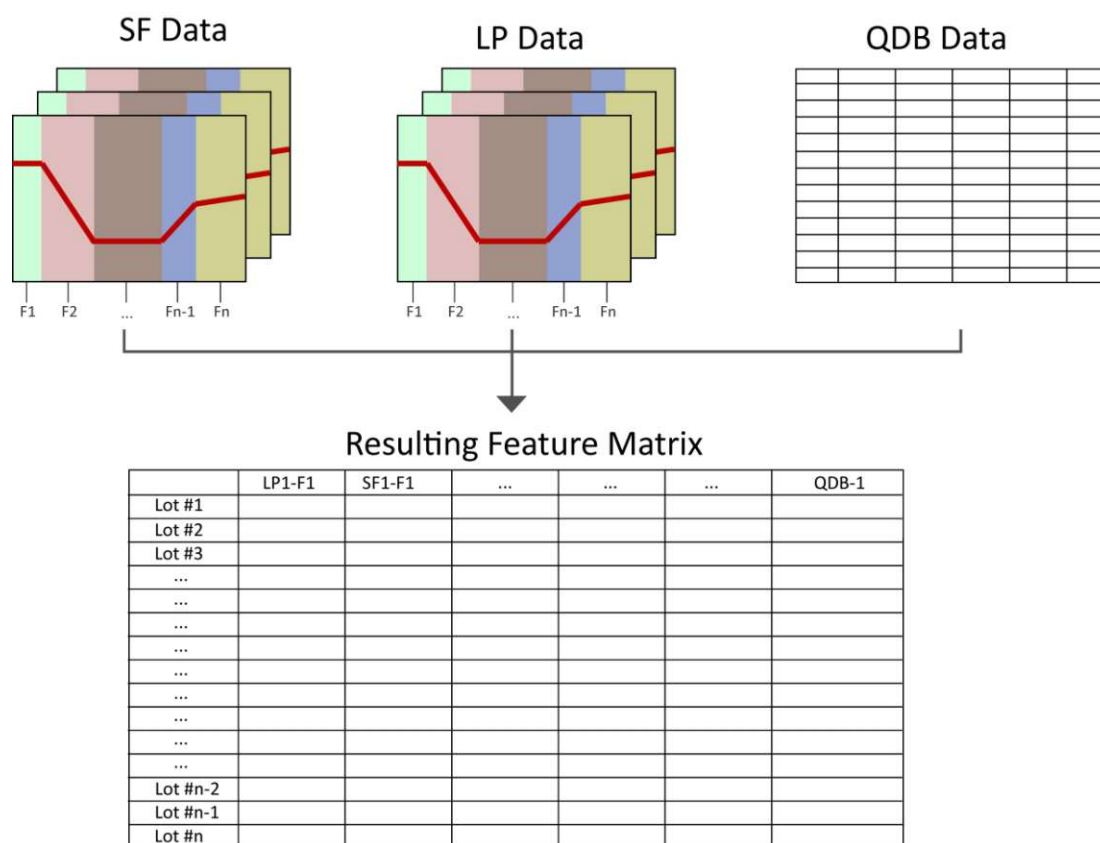


Figure 7. Schematic workflow of feature extraction. After the time-series data of SF and LP are separated into phases, various features are extracted (shown as F1, F2, . . . , Fn-1, Fn) from the data per lot and summarized in an overall feature matrix. Furthermore, the features of the QDB data are added to the feature matrix. The resulting feature matrix is ready for to be further processed by MVDA.

3.5. Step 5: Multivariate Analysis

Since there was no CPV plan in place as of the start of this project, which would have specified any normal operating ranges for the multivariate space, we looked to identify batches that differ from the majority by certain features. This allows us to detect differences within the batches that might remain undiscovered in the univariate space. The robust PCA weighs all 252 features and 58 lots equally, making all features and lots equally important. As described and developed by M. Hubert et al., it is important to distinguish between regular and abnormal observations. Therefore a robust score distance (SDi) and an orthogonal distance (ODi) for each observation is calculated respective to the Mahalanobis distance [26,27]. This calculation is followed by plotting the distances on the y-axis and x-axis for each observation (blue points), as shown in Figure 8 [20]. The plot is divided by a cutoff line (black dashed line) in two yellow, one red, and one green quadrant to differentiate between outliers and normal observations. The cutoff values for the vertical and horizontal line are both 97.5% (97.5% quantile of a chi-squared distribution), again following the calculation developed by M. Hubert et al. If the lots observe the multivariate model, they are located in the green quadrant. However, if the observation is identified as an orthogonal or score outlier, it is found in the left-top yellow quadrant

or right-bottom yellow quadrant, respectively. When a lot is identified as both a score and orthogonal outlier, it is located in the red/orange quadrant [20].

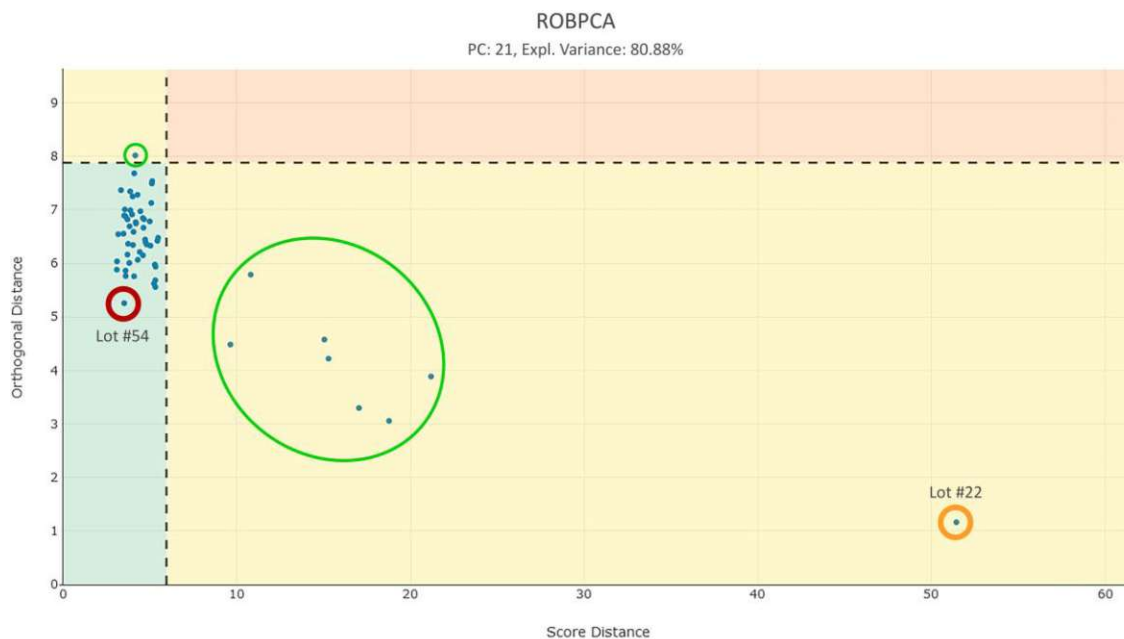


Figure 8. Outlier plot showing all 58 lots (blue points) with their associated scores and orthogonal distance. Additionally, the 97.5% tolerance intervals are indicated as dashed lines, which separate the plot in four differently colored quadrants. Observations in the green quadrant obey the multivariate model, whereas observations in the yellow quadrants indicate orthogonal (upper-left) or score distance (lower-right) outliers. Observations in the yellow quadrants, but close to the 97.5% TI, can be considered as moderate outliers (green circles). The red quadrant includes lots which are suspicious in both directions. Lot #22, shown as encircled in orange, is identified as most prominent outlier within the score distance. Lot #54, shown encircled in red, is closest to the model mean (0,0). No score and orthogonal outlier were observed. The robust principal component analysis (ROBPCA) was built with 21 principal components and explains 80.88% of the variance.

The majority of lots are found in the green quadrant, indicating good multivariate model for the majority of the observations. Furthermore, there are no lots, which are identified as both score and orthogonal outliers, which may be indicative of a stable process. Nonetheless, eight lots are noteworthy in their score distance, of which seven lots are moderately outlying, whereas lot #22 is the most prominent representative of this group and can be assumed to be an outlier, as seen in Figure 8.

In order to discover the underlying root cause for lot #22 abnormality in the score distance, the contribution plot may be applied. This plot displays the contribution of each variable to the score distance of the observation to the model plane. Variables which have a significant contribution to a lot's score distance from the model mean have large values in the bar plot. As seen in Figure 9 a strong pattern for #22 is identified, which is not observed for lot #54 (the lot closest to the center of the proposed model). Of note, the extracted plateau features from variable LP1 and LP2 show a very striking behavior in the scores direction (see "LP1" and "LP2", respectively, Figure 9). Additionally, a CQA feature from the QDB data exhibits the highest contribution in the score distance for lot #22, shown in Figure 9. as "QDB-1". These features explain the distance of lot #22 to the model mean in the outlier plot and therefore may be prioritized for further investigation.

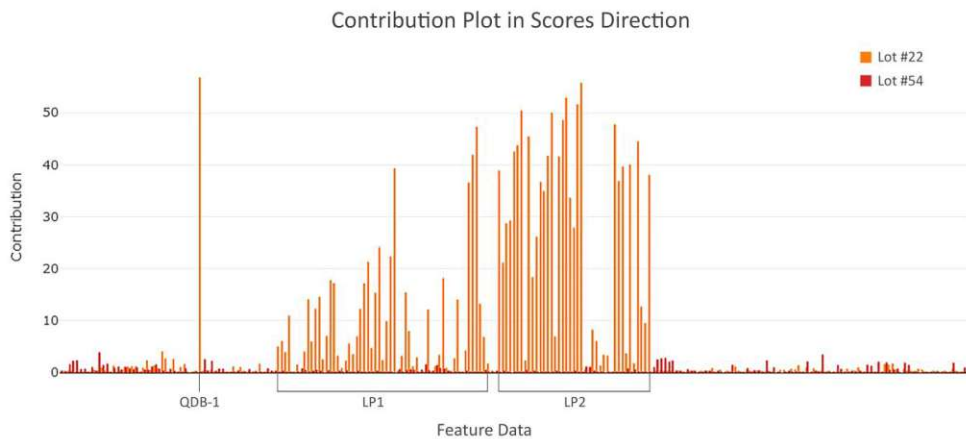


Figure 9. Contribution plot in the scores direction, showing which feature data (x-axis) has most impact to the contribution (y-axis) for the outlier lot and the lot closest to the center of the model. All 252 variables of lot #22 (orange bars) and lot #54 (red bars) are depicted. A strong pattern for lot #22 is identified, which is nonexistent in lot #54. Of particular note is ‘QDB-1’ and the extracted plateau features from variable LP1 and LP2, which exhibit a high contribution (≤ 52) and lead to the outlying score of lot #22.

When comparing the LP2 signal of lot #22 with lot #54, it also becomes clear that the LP2 signal of lot #22 has a ~30% higher noise in the slopes and plateaus as the reference lot, as shown in Figure 10. Using state-of-the-art methods for monitoring in FFF, this noise had been undetected so far.

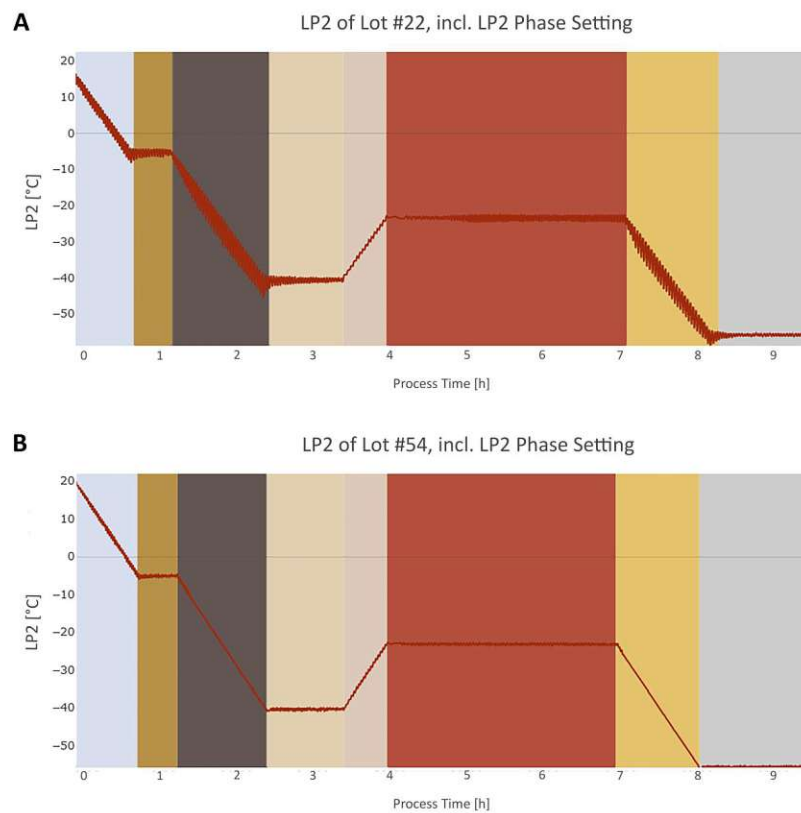


Figure 10. Segment of the LP2 signal (red line) over the process time of lot #22 (A) and lot #54 (B). The colored rectangles represent the individual phases for LP2, discussed in chapter 3.3.2. Step signal phase settings algorithms divide the signal into plateaus and slopes. The LP2 signal in (A) has a ~30% higher noise level than the LP2 signal in (B), which led to the high score distance in the robust PCA, as shown in Figure 8.

4. Discussion

Our presented workflow clearly demonstrates that our approach is capable of achieving the intended purpose of monitoring process-relevant abnormalities in high dimensional FFF data and pinpointing potential root causes, whereas conventional monitoring procedures would have overlooked these abnormalities. The approach represents the first workflow wherein FFF data, regardless of data type, is comprehensively and algorithmically aligned, extracted, and analyzed in a centralized monitoring system.

Data alignment is the first essential and indispensable step when carrying out multivariate analysis on data from various sources. Yet data alignment is still a big obstacle in biopharma and its solution strongly depends on the individual manufacturer [19]. After data alignment and cleaning, we reduce the data dimension by separating the time-series data into phases, where various features are extracted. It is important to note that process expert knowledge is required to decide which phases should be set and which features should be extracted to ensure that all significant production information is being monitored [28]. This is the differentiating factor to the often-used unfolding process, which does not take this process knowledge into account and thus leaves analysis purely in the hands of the data analytical tool.

Lastly, the resulting overall feature matrix requires statistical know-how to robustly identify any multivariate outlier in the data. The following subchapters discuss the applied phase setting and outlier detection in greater detail and compares them to the state of the art.

4.1. Phase Setting

Time-series data includes valuable information, but are often not looked at closely, since analysis is not straight forward as with tabular data records. To apply any statistical tool to the time series for analysis, the data dimension has to be reduced. Commonly, the signals are often unfolded variable- or batch-wise on every single time point [24]. This results in a dataset with many features, which are not necessarily important for process monitoring. Therefore, we present a time-series dimension reduction, where the signals are divided in phases, followed by feature extraction based on process expert knowledge.

The method for phase setting strongly depends on the shape of the signal, as well as on the sample size. In FFF time series, data from lyophilization and filtration processes from lot to lot are very similar and mainly consists of slopes and plateaus. Therefore, we were able to use fixed thresholds for phase setting, whereas the value of the thresholds can be easily evaluated with the help of process knowledge. However, one drawback of this method is, that these thresholds are very rigid and unexpected changes within the process might lead to failed phase settings. Machine learning algorithms, such as random forest could improve the robustness of our proposed phase-settings algorithms [29]. Unfortunately, as with most machine learning algorithms, success depends strongly on N to p ratio, that is, the quality and size of the training data set, which is rarely the case in real biopharma manufacturing processes [30].

4.2. Outlier Detection

The identification of abnormal observations within the monitored data is the most critical aspect of CPV. As in engineering or genetics, the data of biotechnological processes are high-dimensional. However, extreme values within the data can be easily identified by scatter or boxplots when analyzing the data in one dimension. The univariate analysis of a multivariate dataset ignores completely the orthogonal relation between the observations. Therefore, we highly recommend using MDVA for high-dimensional data to identify possible multivariate outliers. One of the most popular statistical tools is PCA, which helps to understand high-dimensional data better and to identify which variables have most effect on the variations within the data. This method tries to explain the covariance by the means of principal components (PC), which are linear combinations of the variables. The PCA can be used not only to determine latent variables in the data but also to identify outliers. Despite its

popularity, one drawback of the classical PCA (CPCA) is that it is highly sensitive to outliers, resulting in a disturbed multivariate model leading to false interpretations, as shown by M. Hubert et al. and V. Todorov et al. [20,31]. Hence, we strongly recommend robust methods, as ROBPCA, when trying to identify outliers. However, the ROBPCA method uses a hard 97.5% cutoff value to distinguish between normal and abnormal observations. P. Filzmoser developed an adaptive method to estimate the cutoff value based on the data structure and sample size, which could further improve the outlier detection results of ROBPCA [32].

Other statistical software such as JMP® offer the possibility to analyze the data with CPCA and use the Hotelling T^2 test and the distance to the model in the X-data (DModX) plot to identify possible outliers, as depicted in Figure 11. After analyzing the final feature matrix with the Hotelling T^2 test and normalized DModX plot four outliers in the score distance and nine outliers in the orthogonal distance were detected, respectively. Lot #22 was also identified as outlier in the Hotelling T^2 test, but its outlyingness was not as pronounced as in ROBPCA compared to the other lots. The identification of the other outliers, identified by CPCA (see Table 3) was not reasonable, since we were not able to identify any reason for their distance-to-model-center when looking at the feature matrix. The false positive outliers might be derived from the fragile multivariate model influenced by the outlier, since CPCA is a nonrobust method.

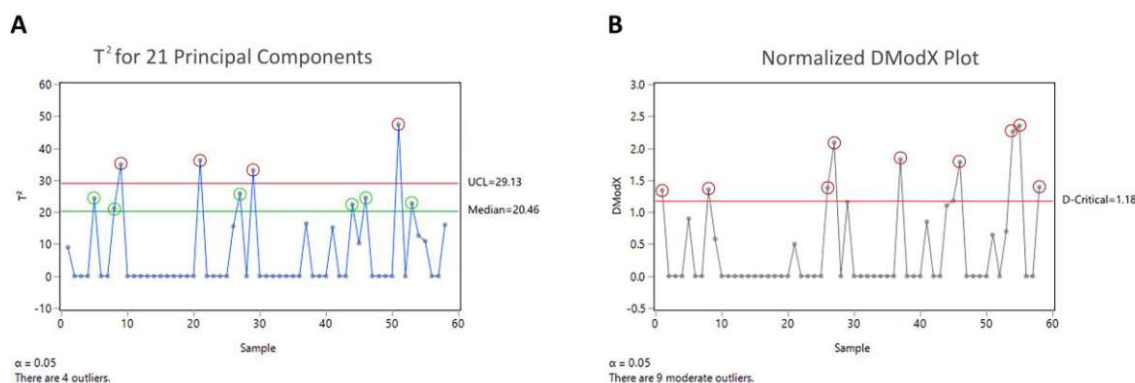


Figure 11. A classical principal components’ analysis (CPCA) was performed on the final feature matrix in JMP®. As with ROBPCA, 21 principal components were used for analysis. The Hotelling T^2 test (A) and normalized DModX plot (B) were implemented in this software for multivariate outlier detection. The Hotelling T^2 plot is used to identify outlying batches in the score direction, if a batch exceeds the red line, representing the 95% confidence interval (CI) of the model population, the batch can be assumed to be an outlier, whereas batches above the green line (median of the model population) can be considered as moderate outlier. The DModX plot is used to identify outlying batches in the orthogonal direction, if a batch exceeds the red line, representing the 95% CI of the model population, the batch can be assumed to be a moderate outlier. Outliers in both tests are marked as red circles.

Table 3. Comparison of CPCA and ROBPCA in terms of outlier identification.

	Score Distance		Orthogonal Distance	
	Moderate Outlier	Outlier	Moderate Outlier	Outlier
CPCA	6	4	-	9
ROBPCA	7	1	1	0

5. Conclusions

We have shown that the presented multivariate FFF monitoring workflow presents a uniquely holistic centralization of all FFF data, while robustly detecting data abnormalities, which have been undiscovered with the current state-of-art methods. These alignment methods have consistently been underused, since data alignment and cleaning are resource-intensive tasks, if not done in an automated

fashion. Furthermore, the classical PCA, established in most statistical software, is not ideal for outlier detection, leading to false-positive results.

We have successfully shown that the workflow is highly automatable through Python scripts and can be therefore used in CPV plans or in daily process-monitoring routines. By using automated phase-setting methods, followed by the extraction of process-relevant features and subsequent robust PCA analysis, multivariate data abnormalities can be easily identified at a glance. Once implemented, this should be easily executable by process engineers and related experts.

Once the requirements are fulfilled, such as sufficient data management and the agreement on certain phase definitions and threshold settings, process experts are able to robustly identify multivariate outliers in their FFF monitoring data with our developed algorithms in inCyght®. The developed algorithms could be further improved by using supervised machine learning methods for faster and more accurate threshold setting.

Supplementary Materials: The feature matrix is available online at <http://www.mdpi.com/2306-5354/7/2/50/s1>.

Author Contributions: B.P. designed the research, conducted the analysis and wrote the manuscript with inputs from C.T. F.D. supported the work by collecting the data and incorporated process knowledge. C.T. and C.H. assisted in planning and writing the manuscript. M.D. helped with the organization of this project and reviewed the manuscript. A.L. supported the work with statistical knowledge. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Baxter AG now part of Takeda in the course of a project with Exputec GmbH.

Acknowledgments: Open Access Funding by TU Wien Bibliothek.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. U.S. Department of Health and Human Services. *Process Validation: General Principles and Practices*; U.S. Department of Health and Human Services: Washington, DC, USA, 2011.
2. Nelson, L.S. The shewhart control chart—Tests for special causes. *J. Qual. Technol.* **1984**, *16*, 237–239. [[CrossRef](#)]
3. Boyer, M.; Gampfer, J.; Zamamiri, A.; Payne, R. A roadmap for the implementation of continued process verification. *PDA J. Pharm. Sci. Technol.* **2016**, *70*, 282–292. [[CrossRef](#)]
4. BPOG. Continued Process Verification: An Industry Position Paper with Example Plan; Biophorum Operations Group. Available online: <https://docplayer.net/21494332-Continued-process-verification-an-industry-position-paper-with-example-plan.html> (accessed on 3 June 2020).
5. Patro, S.Y.; Freund, E.; Chang, B.S. Protein formulation and fill-finish operations. In *Biotechnology Annual Review*; Elsevier: Amsterdam, The Netherlands, 2002; Volume 8, pp. 55–84. ISBN 978-0-444-51025-9.
6. Rathore, N.; Rajan, R.S. Current perspectives on stability of protein drug products during formulation, fill and finish operations. *Biotechnol. Prog.* **2008**, *24*, 504–514. [[CrossRef](#)]
7. Montgomery, D.C. *Statistical Quality Control*, 7th ed.; Wiley: Hoboken, NJ, USA, 1991.
8. Montgomery, D.C.; Jennings, C.L.; Kulahci, M. *Introduction to Time Series Analysis and Forecasting*, 2nd ed.; Wiley: Hoboken, NJ, USA, 1976; ISBN 978-1-118-74511-3.
9. Geurts, P. Pattern extraction for time series classification. In *Principles of Data Mining and Knowledge Discovery*; De Raedt, L., Siebes, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2001; Volume 2168, pp. 115–127. ISBN 978-3-540-42534-2.
10. Stephanopoulos, G.; Locher, G.; Duff, M.J.; Kamimura, R.; Stephanopoulos, G. Fermentation database mining by pattern recognition. *Biotechnol. Bioeng.* **1997**, *53*, 443–452. [[CrossRef](#)]
11. Golabgir, A.; Gutierrez, J.M.; Hefzi, H.; Li, S.; Palsson, B.O.; Herwig, C.; Lewis, N.E. Quantitative feature extraction from the Chinese hamster ovary bioprocess bibliome using a novel meta-analysis workflow. *Biotechnol. Adv.* **2016**, *34*, 621–633. [[CrossRef](#)]
12. Chiang, L.H.; Leardi, R.; Pell, R.J.; Seasholtz, M.B. Industrial experiences with multivariate statistical analysis of batch process data. *Chemom. Intell. Lab. Syst.* **2006**, *81*, 109–119. [[CrossRef](#)]

13. Vo, A.Q.; He, H.; Zhang, J.; Martin, S.; Chen, R.; Repka, M.A. Application of FT-NIR analysis for in-line and real-time monitoring of pharmaceutical hot melt extrusion: A technical note. *AAPS PharmSciTech* **2018**, *19*, 3425–3429. [CrossRef]
14. Chen, J.; Liu, K.-C. On-line batch process monitoring using dynamic PCA and dynamic PLS models. *Chem. Eng. Sci.* **2002**, *57*, 63–75. [CrossRef]
15. Borchert, D.; Suarez-Zuluaga, D.A.; Sagmeister, P.; Thomassen, Y.E.; Herwig, C. Comparison of data science workflows for root cause analysis of bioprocesses. *Bioprocess Biosyst. Eng.* **2019**, *42*, 245–256. [CrossRef]
16. Donoho, D.L. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Chall. Lect.* **2000**, *1*, 1–32. Available online: https://www.researchgate.net/publication/220049061_High-Dimensional_Data_Analysis_The_Curses_and_Blessings_of_Dimensionality (accessed on 2 June 2020).
17. Friedman, J.; Hastie, T.; Tibshirani, R. High-dimensional problems: $P \gg N$. In *The Elements of Statistical Learning*; Springer: Berlin, Germany, 2017; Volume 2, pp. 649–699.
18. Suarez-Zuluaga, D.A.; Borchert, D.; Driessen, N.N.; Bakker, W.A.M.; Thomassen, Y.E. Accelerating bioprocess development by analysis of all available data: A USP case study. *Vaccine* **2019**, *37*, 7081–7089. [CrossRef]
19. Steinwandter, V.; Borchert, D.; Herwig, C. Data science tools and applications on the way to Pharma 4.0. *Drug Discov. Today* **2019**. [CrossRef]
20. Hubert, M.; Rousseeuw, P.J.; Vandenberghe, K. ROBPCA: A new approach to robust principal component analysis. *Technometrics* **2005**, *47*, 64–79. [CrossRef]
21. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417–441. [CrossRef]
22. Zhu, X.; Wu, X. Class noise vs. attribute noise: A quantitative study. *Artif. Intell. Rev.* **2004**, *22*, 177–210. [CrossRef]
23. Brownrigg, D.R.K. The weighted median filter. *Commun. ACM* **1984**, *27*, 807–818. [CrossRef]
24. Lee, J.-M.; Yoo, C.K.; Lee, I.-B. Enhanced process monitoring of fed-batch penicillin cultivation using time-varying and multivariate statistical analysis. *J. Biotechnol.* **2004**, *110*, 119–136. [CrossRef]
25. Agrawal, R.; Nyamful, C. Challenges of big data storage and management. *Glob. J. Inf. Technol.* **2016**, *6*. [CrossRef]
26. De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D.L. The mahalanobis distance. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 1–18. [CrossRef]
27. Brereton, R.G. The Mahalanobis distance and its relationship to principal component scores: The Mahalanobis distance and PCA. *J. Chemom.* **2015**, *29*, 143–145. [CrossRef]
28. Charaniya, S.; Hu, W.-S.; Karypis, G. Mining bioprocess data: Opportunities and challenges. *Trends Biotechnol.* **2008**, *26*, 690–699. [CrossRef] [PubMed]
29. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; pp. 278–282.
30. Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 20. [CrossRef]
31. Todorov, V.; Templ, M.; Filzmoser, P. Detection of multivariate outliers in business survey data with incomplete information. *Adv. Data Anal. Classif.* **2011**, *5*, 37–56. [CrossRef]
32. Filzmoser, P. A Multivariate Outlier Detection Method. 2004. Available online: <http://file.statistik.tuwien.ac.at/filz/papers/minsk04.pdf> (accessed on 2 June 2020).





© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

3.2 Integrated Process Model Applications Linking Bioprocess Development to Quality by Design Milestones

Article

Integrated Process Model Applications Linking Bioprocess Development to Quality by Design Milestones

Christopher Taylor ^{1,2}, Lukas Marschall ^{1,2}, Marco Kunzelmann ³, Michael Richter ³, Frederik Rudolph ³, Judith Vajda ³, Beate Presser ³, Thomas Zahel ¹, Joey Studts ³ and Christoph Herwig ^{1,2,*}

- ¹ Koerber Pharma Software PAS-X Savvy, Mariahilfer Straße 88A/1/9, 1070 Vienna, Austria; christopher.taylor@werum.com (C.T.); lukas.marschall@werum.com (L.M.); thomas.zahel@werum.com (T.Z.)
- ² Research Area Biochemical Engineering, Vienna University of Technology, Gumpendorferstrasse 1a, 1060 Vienna, Austria
- ³ Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Str. 65, 88397 Biberach an der Riss, Germany; marco.kunzelmann@boehringer-ingelheim.com (M.K.); michael.richter@boehringer-ingelheim.com (M.R.); frederik.rudolph@boehringer-ingelheim.com (F.R.); judith.vajda@boehringer-ingelheim.com (J.V.); beate.presser@boehringer-ingelheim.com (B.P.); joey.studts@boehringer-ingelheim.com (J.S.)
- * Correspondence: christoph.herwig@tuwien.ac.at

Abstract: Maximizing the value of each available data point in bioprocess development is essential in order to reduce the time-to-market, lower the number of expensive wet-lab experiments, and maximize process understanding. Advanced in silico methods are increasingly being investigated to accomplish these goals. Within this contribution, we propose a novel integrated process model procedure to maximize the use of development data to optimize the Stage 1 process validation work flow. We generate an integrated process model based on available data and apply two innovative Monte Carlo simulation-based parameter sensitivity analysis linearization techniques to automate two quality by design activities: determining risk assessment severity rankings and establishing preliminary control strategies for critical process parameters. These procedures are assessed in a case study for proof of concept on a candidate monoclonal antibody bioprocess after process development, but prior to process characterization. The evaluation was successful in returning results that were used to support Stage I process validation milestones and demonstrated the potential to reduce the investigated parameters by up to 24% in process characterization, while simultaneously setting up a strategy for iterative updates of risk assessments and process controls throughout the process life-cycle to ensure a robust and efficient drug supply.

Keywords: digital twin; QbD; integrated process model; statistical modelling; bioprocess; control strategy; FMEA; severity rankings; development; risk assessment; DoE



Citation: Taylor, C.; Marschall, L.; Kunzelmann, M.; Richter, M.; Rudolph, F.; Vajda, J.; Presser, B.; Zahel, T.; Studts, J.; Herwig, C. Integrated Process Model Applications Linking Bioprocess Development to Quality by Design Milestones. *Bioengineering* **2021**, *8*, 156. <https://doi.org/10.3390/bioengineering8110156>

Academic Editor: Joaquim M. S. Cabral

Received: 9 September 2021
Accepted: 19 October 2021
Published: 24 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

State-of-the-art bioprocess development seeks a balance between reducing time to market, while satisfying the regulatory requirements as described by the American Food and Drug Administration's proposed validation cycle [1]. These guidelines require extensive process understanding to provide sufficient control in order to ensure that later process changes, whether for scale-up, optimization, or trouble-shooting, do not lead to substantially different product attributes [2]. Generating sufficient data to ensure this process control in situations of potentially unknown variability requires significant experimental resources and time.

An emphasis on the implementation of models as early as possible in process development could potentially save constrained resources and maximize process understanding. In silico process models can also create a baseline upon which late-phase characterization is augmented, ensuring that all data is used at all stages of development, and that all data generated supports process knowledge and control at the most critical steps. Additionally,

connecting these potentially separate data sources and models would enable the iterative improvement of holistic process understanding as described by quality by design (QbD, Figure 1) [3]. For ease of terminology, development data refers here to all data generated prior to process characterization studies (PCSs).

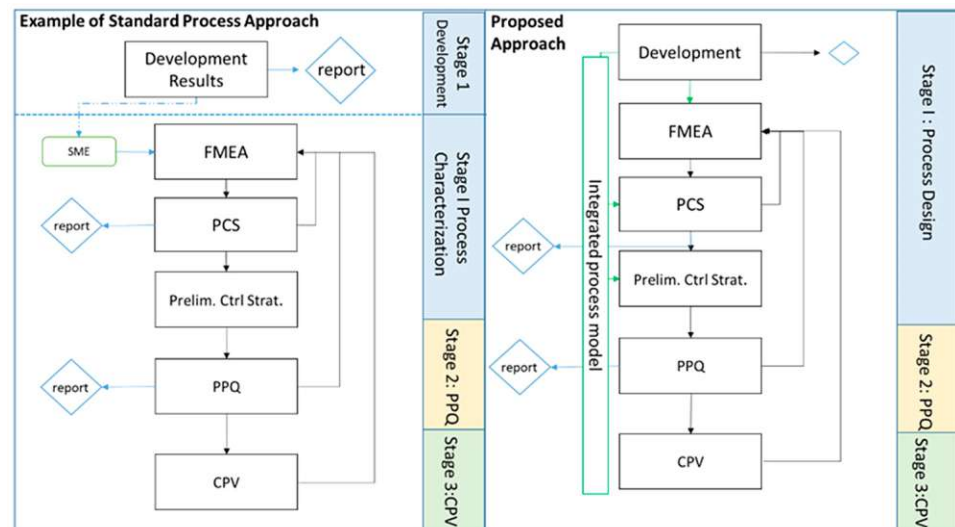


Figure 1. Example of a standard characterization workflow (left) vs. the proposed IPM-based workflow (right). Three aspects are of note. First, an algorithmic link between development and PCS can iteratively add data to an integrated process model. Second, the baseline FMEA severity ranking and preliminary control strategy are statistically underpinned. Third, since the FMEA and preliminary control strategy results are automatically generated in parallel, it removes linear effort in task completion. These activities then supplement the process performance qualification (PPQ) and continued process verification (CPV) stages of validation.

In particular, two deliverables from the FDA QbD framework (i.e., Stage I Process Design) could be significantly supported by quantitatively leveraging models from development:

- Risk assessment, often in the form of a failure mode and effects analysis (FMEA), which are conducted by process experts in order to assess relationships between potential critical process parameters (pCPPs) and critical quality attributes (CQAs), which then are iteratively reinforced by subject matter expertise and experimental results;
- Preliminary control strategy establishment for the CPPs by determining the CPPs' proven acceptable range (PAR) for all CQAs.

One promising technology in this field is the integrated process model—an *in silico* depiction of the entire process chain, where relatively sparse unit operation data can nonetheless be statistically fitted for use in an overarching model. The unit operations are modelled individually. These models are then connected by using the output (dependent variable) of one model as input (independent variable) for the subsequent model, forming the integrated process model [4]. Critically for bioprocesses, the relationships between the unit operations are to be depicted in the overall model so that these applications' predictions can be propagated out to drug substance, rather than being limited to the individual unit operation. That is, the impact of even the first unit operation's process parameters can be modelled *in silico* onto the CQAs at drug substance.

Within this collaboration, we present novel IPM applications, leveraging only existing pre-process characterization data to (1) algorithmically establish risk assessment severity rankings, thereby determining CPPs, while also (2) generating PARs for those CPPs. The presented approach provides a concrete link between existing development data and pro-

cess characterization, and potentially removes subjectivity where data-driven conclusions are otherwise challenging.

1.1. Risk Assessment Severity

Risk assessments such as the FMEA are perhaps the QbD deliverable most sensitive to the quality and quantity of available data, as they must overcome certain numerical [5] and psychological challenges [6]. Specifically, critical impact or severity on a CQA by a process parameter (which leads to the CPP designation) is determined through discussion and process expertise within the perspective of a single unit operation [7]. Rankings are revised iteratively throughout the product life cycle.

Inclusion of all available data in the risk assessment via models creates the benefit of a more quantitative approach to process understanding. A number of complex procedures to quantify these rankings have been proposed and discussed in recent years [8–10]. While all of these methods contribute to the improved modelling of risk, these methods generally do not include the link between each of the unit operations in order to connect conclusions to the final impact on drug substance.

Thus, we seek to establish a quantitative and automated determination of the FMEA severity rankings, based on a model of the available data, that assesses all pCPPs' impact on drug substance, irrespective of that parameter's unit operation, by propagating impact through the entire process flow. We then look to algorithmically assign the severity ranking via a linearization of modelled pCPP impact on the drug substance out-of-acceptance (OOA) likelihood against a predefined critical limit.

1.2. Preliminary Control Strategy

The control strategy is the final manufacturing goal after designating a parameter as a CPP and is typically preliminarily established before process validation Stage II (Process Performance Qualification). A control strategy is any system designed to ensure that CPPs remain in a constant state of control that is required to ensure quality during manufacturing [11]. An essential metric for control strategies is the proven acceptable range (PAR): the range for a CPP within which all CQAs consistently remain within acceptance criteria, while all other parameters stay at set point or within normal operating variation [12].

Djuris provides an overview of available tools and techniques used to achieve this milestone [13]. Most of the approaches used for classifying pCPPs into critical and non-critical designations compare them by their potential impact on quality attributes and the likelihood to cause a result exceeding predefined limits [14,15]. The determination of control ranges is often made based on statistical models, where the limits are defined for the input parameters in such a way that the response variables meet predefined limits [16]. In assessments of process repeatability, parameters to be monitored are also compared against predefined acceptable ranges [17].

All these approaches have in common that they are highly dependent on which quality attribute results are acceptable. For the last unit operation before drug substance, specification limits (if available) can be directly applied; however, for intermediate process steps the assignment of acceptable ranges becomes a challenge, as it is not known which quality attribute levels are acceptable. An improvement would therefore be to use linked data across all unit operations to determine all intermediate ranges.

Eon Duval aims to solve this by linking the statistical models of individual unit operations, feeding the output of the previous unit operation as load into the subsequent unit operation [18]. As input concentrations, they use the worst-case prediction within the normal operating range. This approach has some drawbacks:

- For each CPP, there exists a worst-case condition for each response. The worst case for CQA 1 might not be the worst case for CQA 2.

- The worst case is not the most likely condition. Processes are usually performed at set point conditions and the normal operating ranges represent the uncontrollable variation, meaning that the most probable condition is the set point.
- For the uncertainty of the model predictions, the lower or upper 95% confidence interval is used as a worst case. This does not take into account that the most likely prediction is the model mean at set point.
- The models are only based on small-scale data and manufacturing data is not considered.

Peterson pointed out that for setting up control ranges (i.e., design space), uncertainty in the model prediction needs to be taken into account [19]. Additionally, the uncertainty around the process parameters (i.e., normal operating range) needs to be taken into account as well [20].

To our knowledge, the IPM covers all of the aspects discussed above [21]. We aim to leverage this methodology to set up a control strategy that considers the linkage between unit operations and the uncertainty around process parameter set points [21].

In this collaboration, we will present the following applied techniques with the above goals in mind:

- Create an IPM by concatenating development generated statistical models, thereby establishing an in silico version of the process [21].
- Assess risk assessment severity rankings by application of an IPM parameter sensitivity analysis and rank linearization algorithm that quantifies the behavior of each parameter with regard to its OOA probability, and compares this against a predefined critical OOA rate, assigning an FMEA severity ranking based on the impact at drug substance.
- Propose PAR limits per CPP and CQA by detecting increases in simulated drug substance OOA results across the CPP screening range and assigning a cut-off representative of a predefined acceptable OOA probability.

Finally, we present the above as a proof-of-concept case study using a candidate monoclonal antibody process provided by Boehringer Ingelheim to assess the results of the above procedures.

2. Materials and Methods

2.1. Candidate Process for Case Study

For the applications in this case study, in a collaboration with Boehringer Ingelheim in Biberach, Germany, data sets stemming from development and pre-PCS studies were made available for a candidate monoclonal antibody that depicts a potential platform bioprocess. For this process, the following aspects were relevant to the IPM:

- Eight downstream unit operations, consisting of a capture step (CAP), an acid treatment step (AT), an anion exchange chromatography step (AEX), and a cation exchange chromatography step (CEX), for which process development activities were carried-out. Additional data exists at set point for the following unit operations: depth filtration (DF), ultra-diafiltration (UFDF), viral filtration (VF), and the resulting drug substance (DS).
- DoE-based ordinary least square models, which were fitted, discussed, and selected with subject matter experts before inclusion in the IPM. The experiments were carried out in small-scale systems representative of the manufacturing scale. Subject matter experts evaluated the suitability of these systems prior to experimentation. Model variable selection was based on a standard procedure of selecting the model with lowest Akaike information criterion, and the following diagnostics were then assessed for model significance: R^2_{adj} , Q^2 , RMSE, and partial p -values, as well as the model residuals. All selected models were then discussed with the process experts for process plausibility before acceptance. Acceptable regression models were found for the unit operations CAP, AT, AEX, and CEX.

- Manufacturing data for specific clearance models and yield/clearance calculations as required for the unit operation linkage described elsewhere for all unit operations from two industrial scales [21]:
 - Two manufacturing-scale runs;
 - Three pilot-scale runs.
- Four CQAs typical of monoclonal antibody products, depicting three impurities (CQA1_{imp}, CQA2_{imp}, CQA3_{imp}) and one desired product attribute (CQA1_{prod}):
 - Each of the above CQAs has an acceptance limit at drug substance in place.

2.2. Data for the Integrated Process Model

The IPM technology used here is described in detail elsewhere [21]. Regression models that depict the performance of a unit operation as a function of its process parameters were fitted to the four described responses' specific clearance (SC). Briefly, specific clearance is used here as a general term for the specific, non-volumetric increase or decrease of a CQA, although a desirable product trait may also be accurately described as yield. This term enables a cross-unit operation transfer of the output units as seen in Equation (1) below:

$$SC = PP \cdot \beta_{PP} + \beta_0 + \varepsilon \quad (1)$$

where SC is a vector of the measured specific clearances, PP is a ($n * p$) matrix of the process parameter settings of each DoE run, β_{PP} is the regression coefficient, and β_0 is the intercept. These models will be referred to as DoE models in the following.

Additionally, unifactorial regression models describing the specific clearance performance of a unit operation as a function of the specific load concentration (SLC in Equation (2)) onto pool specific concentration were calculated similarly, providing the link between unit operations. These models will be referred to as load models in the following.

$$SC = SLC \cdot \beta_{SLC} + \beta_0 + \varepsilon \quad (2)$$

For cases where a DoE model and a load model were both available for a unit operation, the results of the prediction were combined according to Equation (3). The predicted clearance at the sampled process parameter settings was corrected by the change in clearance due to a change in SLC.

$$\hat{SC}_i = \hat{SC}(PP_i) \cdot \frac{\hat{SC}(SLC_i)}{\hat{SC}(\overline{SLC}_{DoE})} \quad (3)$$

where $\hat{SC}(PP_i)$ is the expected clearance at the process parameter settings for the current simulation cycle, $\hat{SC}(\overline{SLC}_{DoE})$ is the expected clearance at the mean SLC from the DoE runs, and $\hat{SC}(SLC_i)$ is the expected clearance at the SLC of the current simulation cycle.

The starting concentration for each CQA was assumed to be normally distributed. The mean and standard deviation were estimated from manufacturing-scale runs.

For the case study described, all identified and applied models are summarized as a heat map in Table 1 and again in greater detail in the supplemental information Tables S1 and S2. The underlying assumption is that during the subsequent simulations, sampled prediction variation will be depicted in proportion to the RMSE and should generally correspond to the standard deviation of any reproducible (i.e., set point) runs.

Table 1. Summary of models used for Monte Carlo simulation. DoE models are multiple linear regression models from statistically underpinned designs. SC models are single-factor linear regression models of load to specific clearance. If both types of model were available, ‘Both’ is marked in the table.

Unit Operation	CQA1 _{imp}	CQA2 _{imp}	CQA3 _{imp}	CQA1 _{product}
CAP	DoE	DoE		DoE
AT	Both	DoE	DoE	DoE
DF			SC Model	
AEX	DoE	Both	Both	DoE
CEX	Both	DoE	DoE	DoE
VF				
UFDF	SC Model	SC Model		

2.3. Parameter Sensitivity Analysis

The parameter sensitivity analysis (PSA) is a specialized application of the IPM Monte Carlo simulation [21]. The PSA assesses how the change of each parameter across the full investigated screening range influences OOA events at drug substance. For each process parameter per unit operation for which there was a statistical model, the PSA was conducted per the following procedure:

The individual process parameter’s experimental screening range was divided into 10 equidistant points, referred to as grids, with 10 being the grid size for this study.

At start of the simulation, the parameter was fixed at the grid point at the lowest end of the screening range.

All other process parameters were allowed to vary around their set point within the described normal operating ranges. The process parameters were assumed to be normally distributed with the set point being the mean and the normal operating range being ± 3 standard deviations.

The full Monte Carlo simulation was performed 1000 times at the above conditions for each CQA. An average OOA (%) result was recorded for each CQA, based on a pre-determined acceptance limit.

The individual CPP was then fixed to the next grid and the cycle was re-performed.

Once all grids in the grid size were simulated, the %OOA result was plotted across the screening range.

OOA probability per CQA was calculated according to Equations (4) and (5) below. While the calculation of OOA percent likelihood could further be optimized to include non-parametric procedures for the selected statistical models, the normality assumption at the drug substance level largely applies.

$$OOA = P(X \leq \text{Lower Spec Limit} \cup X \geq \text{Upper Spec Limit}) \quad (4)$$

where X is a normal random variable, $N(\bar{x}, s^{*2})$, where s^* is the upper one-sided confidence limit of the standard deviation.

$$s^* = s \cdot \sqrt{\frac{n-1}{\chi^2(\gamma, n-1)}} \quad (5)$$

The upper confidence interval of the standard deviation was used as an estimate of precision in order to allow for a fair comparison between the OOA rate between observed large-scale data and in silico runs. The sample size of the available large-scale data was very small compared to the runs generated in silico.

PSA results can be plotted in various ways. The individual CQA results can be plotted against an overlay of all CPPs with scaled and centered parameter screening ranges, with

the goal of rapidly determining the most impacting parameters. Conversely, the results can be plotted per CPP against all CQAs. The CPP-based plot is useful for determining a control strategy (described in the next chapter), while the CQA plot allows for quick interpretation of focal points for each CQA during the process. The suggested procedure would be to first identify CQAs strongly impacted by certain CPPs (i.e., plot by CQA) and then to drill down into the CPP-based plots.

Here, we additionally describe an innovative plotting overlay for all CPPs per CQA, referred to as ‘relative screening range’: a plot of a combination of standard data coding, but fixed to the manufacturing set point. The rescaling of the data is generally standardized on a $-1,1$ scale, but in this case it was additionally allowed to shift such that the set point always represented 0. This implies that in the most extreme case where the upper/lower limit of the screening range also represented the set point, the screening range was coded -2 to 0 (or 0,2). Equation (6) is as follows for each point in the screening range:

$$x_i' = \frac{x_i - x_{sp}}{\frac{x_{max} - x_{min}}{2}} \quad (6)$$

where x_i' is the rescaled value of the i th value of parameter x , x_i is the original i th value of parameter x , x_{sp} is the manufacturing set point of the parameter corresponding to x , and x_{min} and x_{max} are the minimum and maximum points of the screening range for parameter x .

This not only makes it possible to quickly see the magnitude with which parameters impact any number of CQAs, but also to see where the set point lies with regard to the explored range, giving an indication of where there may be room for the restriction or expansion of parameter ranges.

3. Results

The development process data provided by Boehringer Ingelheim were successfully used to create an IPM consisting of specific clearance and multi-linear regression models for all four CQAs, to be simulated across seven unit operations.

To implement the proposed novel procedure, the following activities were executed, as described below: (1) IPM plausibility check and confirmation as an adequate model collection, (2) automated generation of risk assessment severity rankings leading to CPP designation, and (3) generation of the preliminary control strategy PAR settings for the CPPs.

3.1. IPM Plausibility Check

The quality of each OLS model contained and concatenated within the IPM was previously individually assessed based on R^2_{adj} , Q^2 , RMSE, p -values, and residual analysis. The linking of the models and the results of the Monte Carlo simulation were additionally checked visually by 1000 cycles executed under the manufacturing conditions (i.e., set point for all parameters with sampling from a normal distribution around the set point). A check of the actual data to the simulated data was performed, and the predicted *in silico* OOA results were compared to the observed OOA results.

Given the sparsity of manufacturing data at this stage of development, the simulated data corresponded adequately well to the existing manufacturing runs, and this repository of data can be seen as a baseline data set that may be used to proceed with further IPM applications (see example in Figure 2 below; remaining plots in the supplementary material, Figures S1–S3).

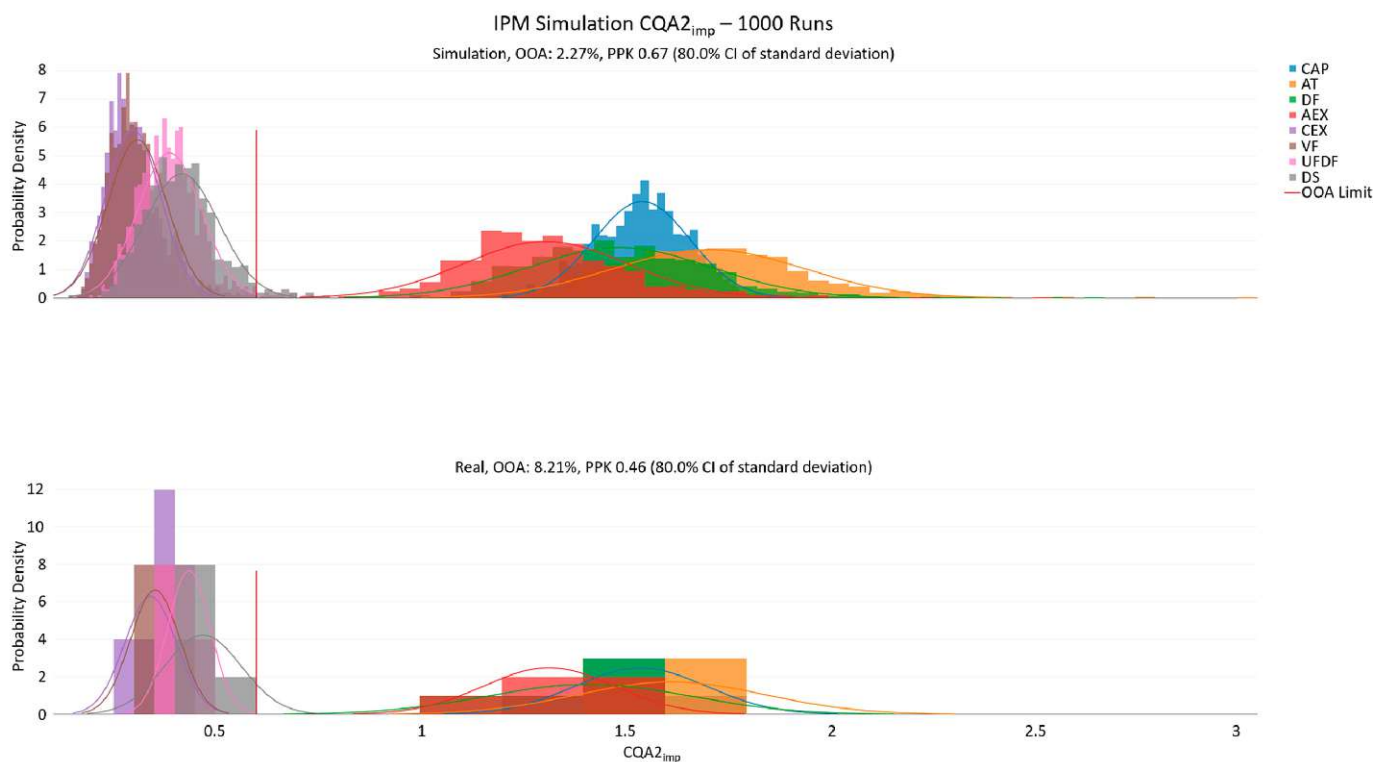


Figure 2. IPM trending plot for 1000 simulations (upper subplot) compared to real data performed (lower subplot) for a CQA impurity. The unit operation order is in descending order in the legend. One can clearly see the descent towards the final drug substance result. The simulation distributions align well with the sparse manufacturing data and the predicted OOA is slightly less than the predicted OOA incidence based on a normal distribution around the real manufacturing data.

3.2. Automated Generation of FMEA Severity Rankings

Upon completion of the PSA simulation, the results were used for the FMEA severity ranking linearization. This procedure describes the ratio of the slope of simulated CQA OOA results to a predefined critical frequency of OOA results. A critical effect slope is defined here as the maximum allowable effect of the pCPP on the CQA between the manufacturing set point and the edge of the screening range (in units %OOA).

3.2.1. Critical Effect Slope Determination

First, the critical effect is depicted as the slope between half the screening range (i.e., set point to the edge of screening range) and the intersection with a predefined critical frequency of OOA results. Here a limit of 5% allowable OOA results was chosen, corresponding to a population outside 2 standard deviations of the normally distributed results for a given CQA. This 5% limit was determined with the subject matter experts and the underlying risk management system.

Specifically, starting from the mean simulated value at the manufacturing set point, out to the intersection of the screening range (x -axis) and the critical effect (5% OOA, y -axis), a line was fit and subsequent slope was calculated. This was performed twice, on both sides of the manufacturing set point. Once established, these lines represent the maximum allowable severity of the pCPP impact on a CQA. These slopes are hereafter referred to as the 'critical slopes' (see Figure 3 as well as Equation (7)).

$$m_{critical\ effect} = \frac{OOA_{SR,5\%} - OOA_{SP}}{CPP_{SR} - CPP_{SP}} \quad (7)$$

where $m_{critical\ effect}$ is the critical slope, OOA is the out-of-acceptance result as a percentage, SR is the screening range limit (max or min), $SR,5\%$ is the intersection of the screening range and 5% OOA , and SP is the set point.

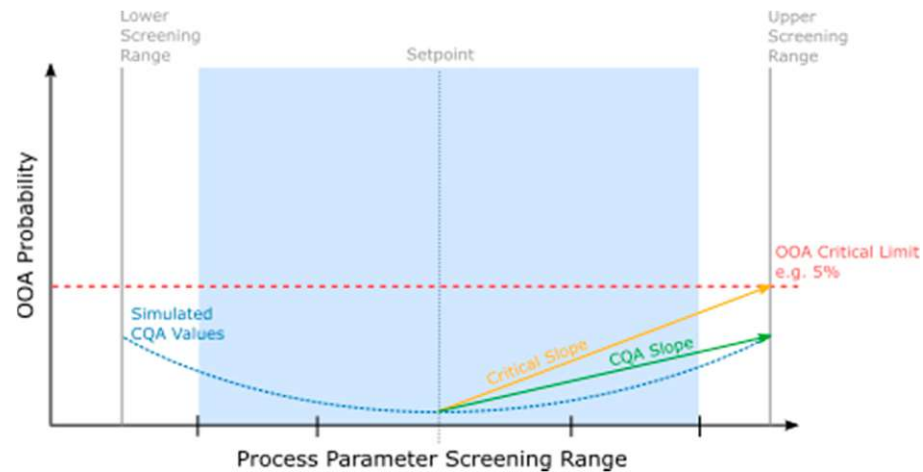


Figure 3. Example of FMEA linearization procedure. The shaded area represents the currently proposed manufacturing range.

3.2.2. CQA Slope Determination

The CQA slope was fit in a similar manner, but based only on the simulated data for the CQA in question (see Figure 3). A line was fit between the results of the simulation at the set point value and the simulated mean value at the screening range limit. This slope was also calculated and repeated for both sides of the set point. Once established, these slopes represent the simulated relationship between the pCPP and the CQA in the OOA results. This slope's relationship to the critical slope is essential to quantify how close it is to an unacceptable impact. These slopes are referred to as CQA slopes in Equation (8) below:

$$m_{cqa} = \frac{OOA_{SR} - OOA_{SP}}{CPP_{SR} - CPP_{SP}} \tag{8}$$

where m_{cqa} is the CQA slope, OOA is the out-of-acceptance result as a percentage, and SP is the set point.

3.2.3. FMEA Ranking Algorithm

Once both the critical slope and the CQA slope were determined, the two slopes were compared as a ratio, with the critical slope being the reference slope and the CQA slope being calculated as a % reference value in Equation (9) below:

$$\%ref = \frac{m_{cqa}}{m_{critical\ effect}} * 100 \tag{9}$$

The calculation was performed twice, once for each side of the screening range. The 'worst-case' slope of the two was chosen for the ranking. This %ref value was then compared to a classification rubric to determine the corresponding FMEA severity value (Table 2). As a baseline, any CQA slope equal to or larger than the critical slope must by design correspond to the highest severity ranking.

The FMEA ranking was determined using the agreed-upon rubric (Table 2), based on an internal company 10-point FMEA severity ranking scale. Other ranking scales may be used as well.

Table 2. FMEA ranking algorithm.

% Reference	FMEA Ranking
≥0.8	10
0.5–0.8	7
0.3–0.5	3
<0.3	1

As a case study of the linearization methodology, the risk rankings of two selected process parameters that were already assessed in a current best-practices FMEA process (selected by process experts for proof of concept) were algorithmically assessed in the IPM and the results were compared.

The risk rankings from the FMEA expert team and the risk rankings from the IPM application are compared in Figure 4. The results and rankings of the process-expertise-based FMEA evaluation were not made available before the completion of IPM in order to avoid any form of bias. Both the FMEA team and the IPM severity ranking agree that the parameter AT_pH should be considered critical. The IPM results also generally agree with the expert assessment of the CAP_Residence Time as a non-critical parameter.

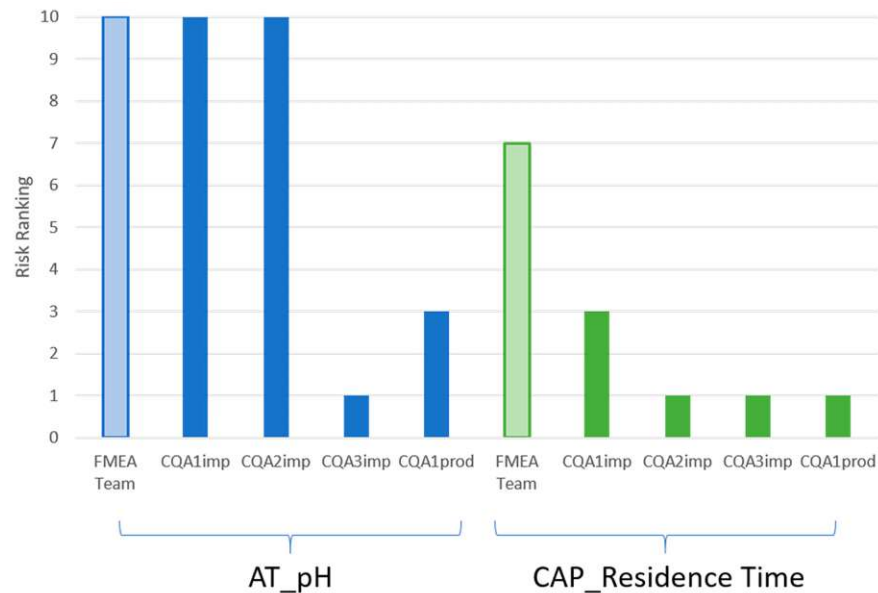


Figure 4. Results of the algorithmic setting of the FMEA severity ranking per CQA. The FMEA team bar represents the current process-expertise-based ranking with the granularity of final ranking for the worst-case CQAs. The IPM ranking assesses each of the CQAs against the 2 parameters in question. As CPPs are classified as ranking 10, AT_pH will be considered a CPP in both assessments, whereas CAP_Residence Time will be considered a PP. While both methods agree with the final ranking, there is a difference in the assessment of CAP_Residence Time, with the FMEA team ranking this as a more critical parameter than the PSA.

3.2.4. Preliminary Control Strategy Setting Procedure

In establishing the control strategy, the ICH Q8(R2) guideline provides a description of the proven acceptable range that could be translated to a quantitative description that may then be used to establish a control strategy using the PSA. According to the ICH, a proven acceptable range is ‘A characterized range of a process parameter for which operation within this range, while keeping other parameters constant, will result in producing a material meeting relevant quality criteria.’

We defined the PAR for each process parameter as the range where CQA results were within acceptance limits at drug substance at a certain probability level. Therefore, we first defined the critical level (i.e., the out-of-acceptance probability) that the manufacturer

is willing to accept. Again, we defined this as 5% out-of-acceptance results (roughly equivalent to accepting 2 standard deviations of the population distribution).

CQA OOA probability results are plotted across the screening range. Once the CQAs were identified as having high areas of risk, that is, the OOA >5% limit was crossed (see Figure 5), the PAR was set to the point at which 5% was intersected. This was iterated through all CQAs for all pCPPs and the most restrictive of the PARs was chosen as the ultimate PAR limit.

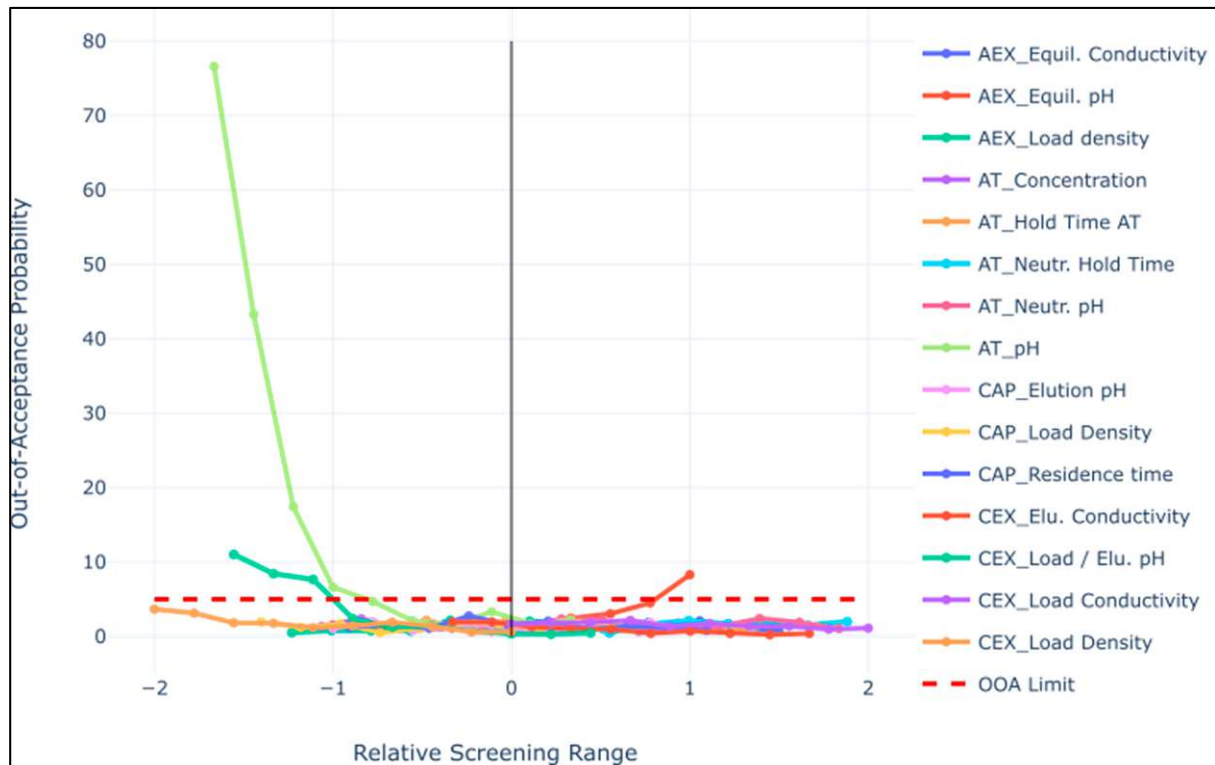


Figure 5. Relative screening range of CQA1_{imp} for all tested potential CPPs. The set point was coded to zero and the design space varied between −2 and 2 based on where the set point lies.

The CPP-based PSA plots can be used for easy interpretation of the above procedure. In Figure 6, the behavior of all CQAs is shown across the full screening range for a given CPP. A line is drawn as each of the CQAs crosses the 5% threshold, and the final PAR is the most restrictive range. Lastly, areas of the screening range within PAR are in grey. Fifteen pCPPs were assessed in this manner. The control strategy plot for AT_pH is depicted below. Further plots can be found in the supplementary information per CPPs in Table 3 (Figures S4–S7).

Table 3. Summary overview of automated results for PAR setting by PSA.

Unit Operation	Process Parameter	Proposed Control Range (If Any)	Justification of PAR
AT	pH AT	Restriction Low	CQA4 _{prod} , CQA1 _{imp} , CQA3 _{imp}
CEX	Conductivity	Restriction High	CQA1 _{imp} , CQA2 _{imp} , CQA3 _{imp}
	Load Density	Restriction High	CQA1 _{imp}

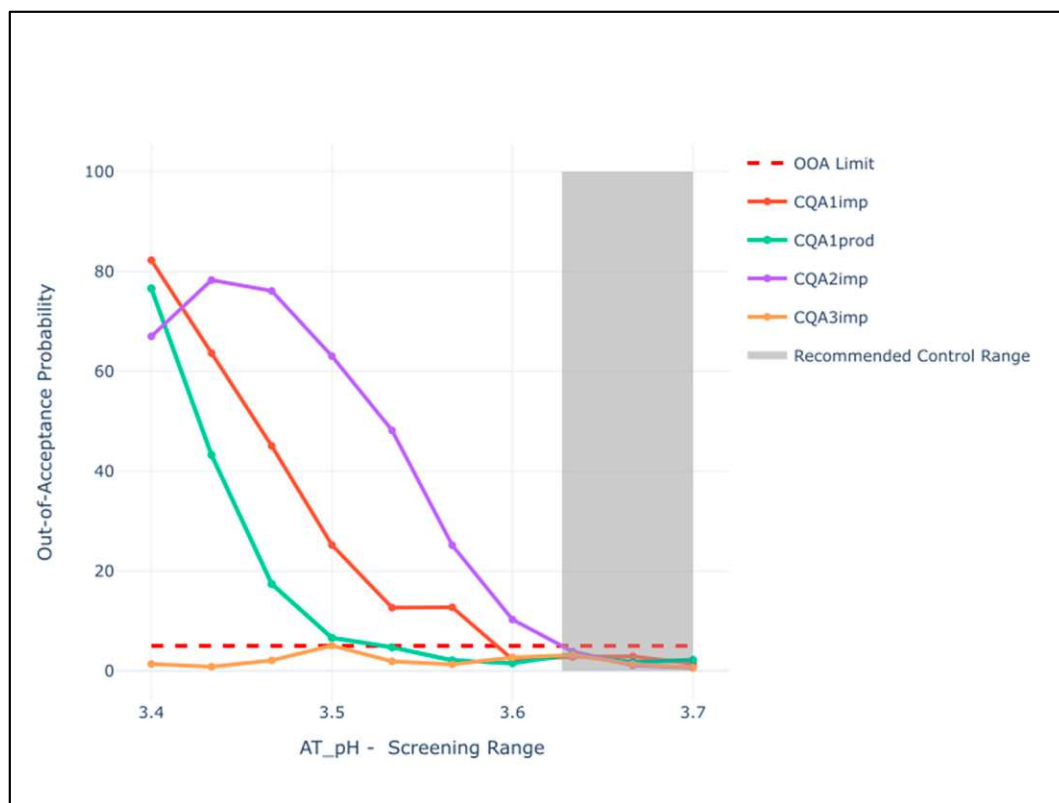


Figure 6. PSA plot for AT_pH. The impact of the 4 CQAs is depicted as a behavior across the screening range. The proposed control range is highlighted in grey. Note that AT is the second unit operation in process order, but the plots depict the impact on DS results.

Table 3 presents the generalized results for the PSA study, with regards to setting up a control strategy. The initially proposed manufacturing control range was either acceptable across the entire screening range, or only a subset of the original manufacturing control range was found to be acceptable; that is, a restriction on either higher or lower end is proposed. In this case, the impacted CQA is listed and plotted in the supplementary information (Figures S7–S9).

4. Discussion

The novel linkage of development data to the PCS QbD milestones of risk assessment severity ranking and preliminary control strategy establishment was proven feasible within the case study data provided by Boehringer Ingelheim. These models now improve characterization workflows by reducing FMEA subjectivity and decreasing required PCS resources as described below.

4.1. Automated FMEA Severity Ranking Linearization

In comparing the novel IPM FMEA result to the current best practice methods of FMEA ranking via process expertise, there are two aspects of the ranking that are of interest to compare:

- The overall FMEA severity ranking;
- The CQA(s) most likely to be impacted by the pCPP.

As can be seen in Figure 4, an identical overall CPP designation was generated between the IPM FMEA ranking and the current state-of-the-art experience-based FMEA ranking. Both evaluations assign critical and non-critical designations in the same manner for both process parameters. The FMEA team ranked the CAP_Residence Time as more critical than did the PSA algorithm, but both agreed that it should not be identified as a CPP. An

advantage of the novel procedure is that it produced similar results, but is not dependent on long discussions or the availability of experts, nor is it subject to other challenges discussed in the introduction. Nonetheless, targeted discussions may be used if there is uncertainty about the PSA ranking, while still reducing the time consumed and focusing discussions on key findings.

The IPM method additionally generates risk rankings for all CQAs for each pCPP. Though this leads to the same ultimate conclusion as other methods (as in any case, the worst-case CQA ranking is taken as the final ranking), evaluating each process parameter with the individual CQAs has the advantage that it provides a more detailed and distinguished picture that adds to process knowledge. Here, it is of interest that two separate CQAs were severely impacted by AT_pH; this is considered to be a novel observation.

This granularity makes it so that, in the ensuing process characterization experiments, not all CQAs may need to be investigated at every unit operation, saving analytical costs. If a CPP is determined to not impact any investigated CQAs, it can be feasibly excluded from experiments in the PCS, reducing experimental cost. With the data-driven statistical evaluation of the parameters, there is a clear and easily describable justification for control of the parameter, and therefore the exclusion of the additional wet-lab work. Both provide a potentially substantial reduction in characterization effort.

Moreover, the IPM procedure provides a quantitative logic and result at the drug substance level that can be documented and visualized to demonstrate the data-driven impact of CPPs on CQAs. The ability to both quantify and visually display the full process relationships based on models significantly improves the understanding, justification of control, and ultimate acceptance by regulatory authorities in submission, while also removing some elements of subjectivity from risk-based decisions.

The analytical set here is small and therefore serves only as a proof of concept for the algorithm, and a full comparison of all investigated pCPPs would provide a better justification for the use of an automated severity risk assessment. Secondly, the IPM does not remove subjectivity completely. The screening range, as well as the predetermined critical limit (i.e., 5%), may include subjectivity.

Nonetheless, the ability to integrate the full process dataset into an automated ranking system provides improved detail, rationale, and less subjectivity, while potentially saving on future resources and increasing process control.

4.2. PSA Established Control Strategy

The IPM control strategy improves on current model-based PAR strategies by using a variation sampling technique that simulates the realities of manufacturing at scale, while concurrently establishing the PAR based on the results at drug substances, rather than at the individually modeled unit operations.

The simulation of the process parameters, sampled from within a normal distribution around the parameter set point, allows parameter settings to be simulated that are both more realistic (not simply worst case), and allows factors to interact in a multivariate space.

Furthermore, determining PAR on the results at drug substance also allows this variation to propagate throughout the process, delivering an accurate depiction of a process parameter's impact on the final product.

One additional underlying benefit is that the technology used here is still based on standard ordinary least square models, which are well established and already being used in the determination of PAR settings (Burdick et al. 2017), thus reducing the effort required to make the logical case for the concatenation of the models in an IPM as well as the adequacy of the resulting proposed PAR limits (which are also later verified by the process performance qualification) in regulatory submissions.

Through the effective use of development data and the implementation of the models described here, a primary control strategy can be established, which will allow the development team to focus characterization studies only on areas where significant variability exists or on parameters that show a high level of uncertainty, thus allowing for clear justifi-

cations of control for all parameters and a reduced workload for process characterization—normally the largest work package for Stage I process validation.

This methodology is of course only as good as the models upon which it is built, and they are in turn dependent on the experimental conditions in development. As time and costs are a major constraint in pre-PCS studies, this could be a limitation to the efficacy of the proposed methods (e.g., screening designs are used instead of response surface designs). These results may nonetheless provide the basis for an iterative development road map, leading to the prioritization of targeted CQAs and CPPs in process characterization.

In this feasibility study, fifteen parameters were successfully modelled in the IPM, of which two were removed from later PCS studies. Thirteen remaining parameters that were then committed to PCS experiments represented 24% of all experimentally assessed characterization parameters. Thus, without further detailed quantitative assessment, it can be stated broadly that up to 24% of the PCS study parameters could be saved by applying the above-discussed approach, instead of using additional resources. This cost savings, along with the data-driven and graphical justification of control, combine to form a powerful tool to both reduce costs and simultaneously increase process understanding—normally a paradigm that has the opposite correlation.

5. Conclusions

Leveraging development data to create in silico IPM models improves upon current best practices by enabling faster establishment of QbD deliverables of risk assessment rankings and preliminary control strategies, ultimately leading to less future experimental effort based on better understanding of the available data, thus leading to significantly better process understanding and control.

A promising next step in this research would be to attempt to automate other standard FMEA rankings (e.g., occurrence or detectability) using an adjusted concept. For example, it could be possible to simulate occurrence results by estimating the capability values of the IPM simulated distributions. The ultimate goal in this work would be to generate the entire FMEA by using the IPM in development.

These approaches are limited in that they require the presence of data. Further research may therefore investigate the establishment of a multiple-product encompassing ‘platform’ IPM, which can serve as a starting point for the first-iteration FMEAs, which could point to the most probable CPPs based on knowledge from previous projects.

For future applications such as process monitoring, the IPM can be updated with new in-process manufacturing data. The IPM can then be used to predict the out-of-specification probability of the currently running campaign. If this probability is undesirably high, the model can then be used to propose changes in process parameter set points to lower the out-of-specification probability to acceptable levels. This constant update with data and feedback into the system could turn the IPM into a category of digital twin [21].

With the continuous exploration of advanced in silico process models, development data should increasingly be seen as a vital basis for IPMs. The technology presented here fully demonstrates the power of applying statistical tools to maximize the knowledge gained from the available data and how focused and efficient knowledge management can be used to invert the paradigm of increased process understanding being associated with increased development costs.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/bioengineering8110156/s1>. Table S1: Overview of identified models based on DoE data. Table S2: Overview of specific clearance models. Figures S1–S3: Simulation trend plots. Figures S4–S6: PSA CQA plots. Figures S7–S9: PSA CPP plots.

Author Contributions: C.T. designed and implemented the linearization approach of the PSA FMEA rankings, conducted the case study, and drafted the manuscript. L.M. designed and implemented the PSA, designed the control strategy establishment methodology, and provided support in drafting the manuscript. C.T. and L.M. contributed equally to this project. M.K. provided key statistical and

process expertise and functioned as the liaison between the two partnering companies. M.R. and F.R. performed the case study experimental design and experiments and assisted in establishing all statistical models relevant to the process. B.P., J.V., T.Z., J.S. and C.H. assisted in the writing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. *FDA Process Validation: General Principles and Practices Guidance for Industry*; FDA, January 2011. Available online: <https://www.fda.gov/media/71021/download> (accessed on 10 August 2021).
2. Lim, L.P.L.; Garnsey, E.; Gregory, M. Product and process innovation in biopharmaceuticals: A new perspective on development. *R&D Manag.* **2006**, *36*, 27–36. [[CrossRef](#)]
3. Narayanan, H.; Luna, M.F.; Von Stosch, M.; Bournazou, M.N.C.; Polotti, G.; Morbidelli, M.; Butté, A.; Sokolov, M. Bioprocessing in the Digital Age: The Role of Process Models. *Biotechnol. J.* **2019**, *15*, e1900172. [[CrossRef](#)] [[PubMed](#)]
4. Bruynseels, K.; De Sio, F.S.; Hoven, J.V.D. Digital Twins in Health Care: Ethical Implications of an Emerging Engineering Paradigm. *Front. Genet.* **2018**, *9*, 31. [[CrossRef](#)] [[PubMed](#)]
5. Borchert, D.; Zahel, T.; Thomassen, Y.E.; Herwig, C.; Suarez-Zuluaga, D.A. Quantitative CPP Evaluation from Risk Assessment Using Integrated Process Modeling. *Bioengineering* **2019**, *6*, 114. [[CrossRef](#)] [[PubMed](#)]
6. Liu, H.-C.; Liu, L.; Liu, N. Risk evaluation approaches in failure mode and effects analysis: A literature review. *Expert Syst. Appl.* **2013**, *40*, 828–838. [[CrossRef](#)]
7. Sharma, K.D.; Srivastava, S. Failure Mode and Effect Analysis (FMEA) Implementation: A Literature Review. *J. Adv. Res. Aero Space Sci.* **2018**, *5*, 1–17.
8. Franceschini, F.; Galetto, M. A new approach for evaluation of risk priorities of failure modes in FMEA. *Int. J. Prod. Res.* **2001**, *39*, 2991–3002. [[CrossRef](#)]
9. Shahin, A. Integration of FMEA and the Kano model: An Exploratory Examination. *Int. J. Qual. Reliab. Manag.* **2004**, *21*, 731–746. [[CrossRef](#)]
10. Liu, H.-C.; Wang, L.-E.; Li, Z.; Hu, Y.-P. Improving Risk Evaluation in FMEA With Cloud Model and Hierarchical TOPSIS Method. *IEEE Trans. Fuzzy Syst.* **2018**, *27*, 84–95. [[CrossRef](#)]
11. ICH. Q8 (R2) *Pharmaceutical Development Q8 (R2)—Step 4*; The International Conference on Harmonisation: Geneva, Switzerland, 2009.
12. EMA/213746/2017 EMA-FDA Questions and Answers: Improving the Understanding of NORs, PARs, DS_p and Normal Variability of Process Parameters. Questions and Answers 4. Available online: https://www.ema.europa.eu/en/documents/scientific-guideline/questions-answers-improving-understanding-normal-operating-range-nor-proven-acceptable-range-par_en.pdf (accessed on 10 August 2021).
13. Djuris, J.; Djuric, Z. Modeling in the quality by design environment: Regulatory requirements and recommendations for design space and control strategy appointment. *Int. J. Pharm.* **2017**, *533*, 346–356. [[CrossRef](#)] [[PubMed](#)]
14. Wang, K.; 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. *Pharm. Technol.* **2016**, *40*, 36–44.
15. Glodek, M.; Liebowitz, S.; Squibb, B.-M.; McCarthy, R.; Sundararajan, M.; Vorkapich, R.; Healthcare, B.; Watts, C.; Millili, G. Process Robustness—A PQRI White Paper. *Pharm. Eng.* **2006**, *26*, 11.
16. Abu-Absi, S.F.; Yang, L.; Thompson, P.; Jiang, C.; Kandula, S.; Schilling, B.; Shukla, A.A. Defining process design space for monoclonal antibody cell culture. *Biotechnol. Bioeng.* **2010**, *106*, 894–905. [[CrossRef](#)] [[PubMed](#)]
17. Vukovinsky, K.E.; Li, F.; Hertz, D. Estimating Process Capability in Development & Low-Volume Manufacturing. Available online: <https://ispe.org/pharmaceutical-engineering/january-february-2017/estimating-process-capability-development-low> (accessed on 16 November 2020).
18. Eon-Duval, A.; Valax, P.; Solacroup, T.; Broly, H.; Gleixner, R.; Le Strat, C.; Sutter, J. Application of the quality by design approach to the drug substance manufacturing process of an Fc fusion protein: Towards a global multi-step design space. *J. Pharm. Sci.* **2012**, *101*, 3604–3618. [[CrossRef](#)] [[PubMed](#)]
19. 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. [[CrossRef](#)]

20. Burdick, R.; Coffey, T.; Gutka, H.; Gratzl, G.; Conlon, H.D.; Huang, C.-T.; Boyne, M.; Kuehne, H. Statistical Approaches to Assess Biosimilarity from Analytical Data. *AAPS J.* **2016**, *19*, 4–14. [[CrossRef](#)] [[PubMed](#)]
21. Zahel, T.; Hauer, S.; Mueller, E.M.; Murphy, P.; Abad, S.; Vasilieva, E.; Maurer, D.; Brocard, C.; Reinisch, D.; Sagmeister, P.; et al. Integrated Process Modeling—A Process Validation Life Cycle Companion. *Bioengineering* **2017**, *4*, 86. [[CrossRef](#)] [[PubMed](#)]

3.3 Specification-Driven Acceptance Criteria for validation of biopharmaceutical processes



OPEN ACCESS

EDITED BY
Matthias Rüdert,
HES-SO Valais-Wallis, Switzerland

REVIEWED BY
Cenk Undey,
Amgen, United States
Nick Whitelock,
Cytiva, United Kingdom

*CORRESPONDENCE
Christoph Herwig,
christoph.herwig@tuwien.ac.at

†These authors contributed equally to
this work and share first authorship

SPECIALTY SECTION
This article was submitted to Bioprocess
Engineering,
a section of the journal
Frontiers in Bioengineering and
Biotechnology

RECEIVED 03 August 2022
ACCEPTED 05 September 2022
PUBLISHED 23 September 2022

CITATION
Marschall L, Taylor C, Zahel T,
Kunzelmann M, Wiedenmann A,
Presser B, Studts J and Herwig C (2022),
Specification-driven acceptance criteria
for validation of
biopharmaceutical processes.
Front. Bioeng. Biotechnol. 10:1010583.
doi: 10.3389/fbioe.2022.1010583

COPYRIGHT
© 2022 Marschall, Taylor, Zahel,
Kunzelmann, Wiedenmann, Presser,
Studts and Herwig. This is an open-
access article distributed under the
terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which does
not comply with these terms.

Specification-driven acceptance criteria for validation of biopharmaceutical processes

Lukas Marschall^{1,2†}, Christopher Taylor^{1,2†}, Thomas Zahel¹,
Marco Kunzelmann³, Alexander Wiedenmann³, Beate Presser³,
Joey Studts³ and Christoph Herwig^{1,2*}

¹Körber Pharma Software, Vienna, Austria, ²TU Wien, Institute for Chemical, Environmental and Bioscience Engineering, Vienna, Austria, ³Boehringer Ingelheim Pharma GmbH & Co, Biberach an der Riss, Germany

Intermediate acceptance criteria are the foundation for developing control strategies in process validation stage 1 in the pharmaceutical industry. At drug substance or product level such intermediate acceptance criteria for quality are available and referred to as specification limits. However, it often remains a challenge to define acceptance criteria for intermediate process steps. Available guidelines underpin the importance of intermediate acceptance criteria, because they are an integral part for setting up a control strategy for the manufacturing process. The guidelines recommend to base the definition of acceptance criteria on the entirety of process knowledge. Nevertheless, the guidelines remain unclear on how to derive such limits. Within this contribution we aim to present a sound data science methodology for the definition of intermediate acceptance criteria by putting the guidelines recommendations into practice (ICH Q6B, 1999). By using an integrated process model approach, we leverage manufacturing data and experimental data from small scale to derive intermediate acceptance criteria. The novelty of this approach is that the acceptance criteria are based on pre-defined out-of-specification probabilities, while also considering manufacturing variability in process parameters. In a case study we compare this methodology to a conventional +/- 3 standard deviations (3SD) approach and demonstrate that the presented methodology is superior to conventional approaches and provides a solid line of reasoning for justifying them in audits and regulatory submission.

KEYWORDS

integrated process model, statistical modelling, bioprocess, control strategy, acceptance criteria, specification limits, process validation, DOE

1 Introduction

Process Validation for the pharmaceutical industry is “the collection and evaluation of data, from the process design stage throughout production, which establishes scientific evidence that a process is capable of consistently delivering quality products.” (FDA, 2011). This involves a series of activities taking place over the life

cycle of the product and process. The goal of process validation is to set-up and maintain a control strategy that enables the process to continuously deliver product quality. This desired quality is defined by the quality target profile (QTPP) of a product (ICH Q8, 2009, S. 8) and acceptable quality limits are defined by drug substance and drug product specification limits. The final gate keeper for the market release of product from a manufacturing process are the drug product specification limits for each of the individual attributes of the QTPP, referred to as Critical Quality Attributes (CQAs).

Amongst other goals, a control strategy aims to control 3 types of parameters: process parameters (CPPs), material attributes (CMAs) and the quality attributes themselves (Burdick et al., 2017). In process design, depicting phase 1 of process validation, process parameters and material attributes are assessed and investigated (FDA, 2011). Their impact on product quality and process performance is studied and quantified in experiments. Dependent on the observed effects on product quality, appropriate control ranges are defined for process parameters and quality attributes. Most commonly, each process step (or unit operation) is investigated individually. However, to define the control ranges of CPPs and CMAs, it is important to know which quality attribute levels are acceptable at each process step (Jiang et al., 2010).

In ICH Q6B, an acceptance criterion is defined as “An internal (in-house) value used to assess the consistency of the process at less critical steps.” (ICH Q6B, 1999, S. 6). Within this contribution, we focus on acceptance criteria for CQAs at intermediate process steps (Figure 1). Hence, we refer to these limits as intermediate acceptance criteria. They

describe which quality levels each unit operation has to deliver, whereas the drug substance or product specification limits describe, which quality levels the process has to ultimately deliver before product release.

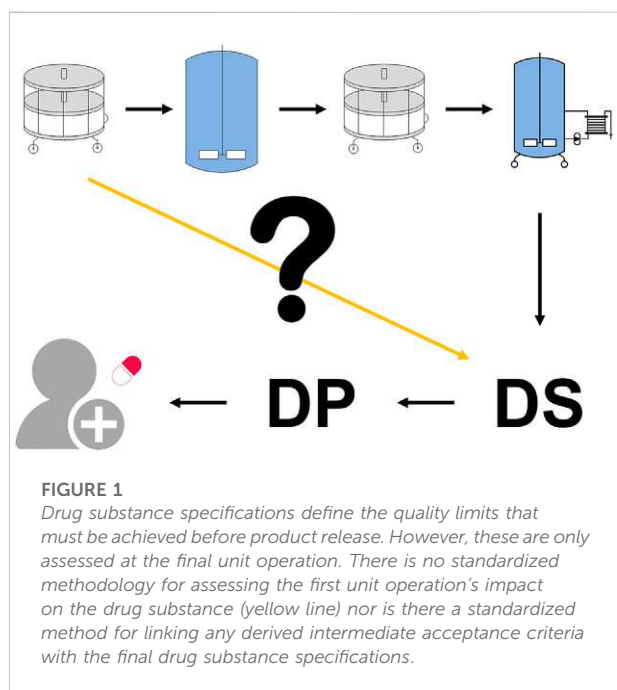
Without knowing which quality levels are acceptable at each process step, it is difficult to set up control ranges for CPPs and CMAs at the respective process steps. As managing the risk to quality is regarded to be the ultimate goal (ICH, 2005), deriving these limits is crucial for the success of a process validation project. EMA-FDA, also requires acceptance criteria for CPPs and CQAs to be part of the process validation protocol (European Commission, 2015). Per these guidelines, the acceptance criteria should be based on development data or documented process knowledge. If the measurement of quality attributes in the process are part of the control strategy (as in-process controls), intermediate acceptance criteria (iACs) are required and solid rationales should be provided for their establishment.

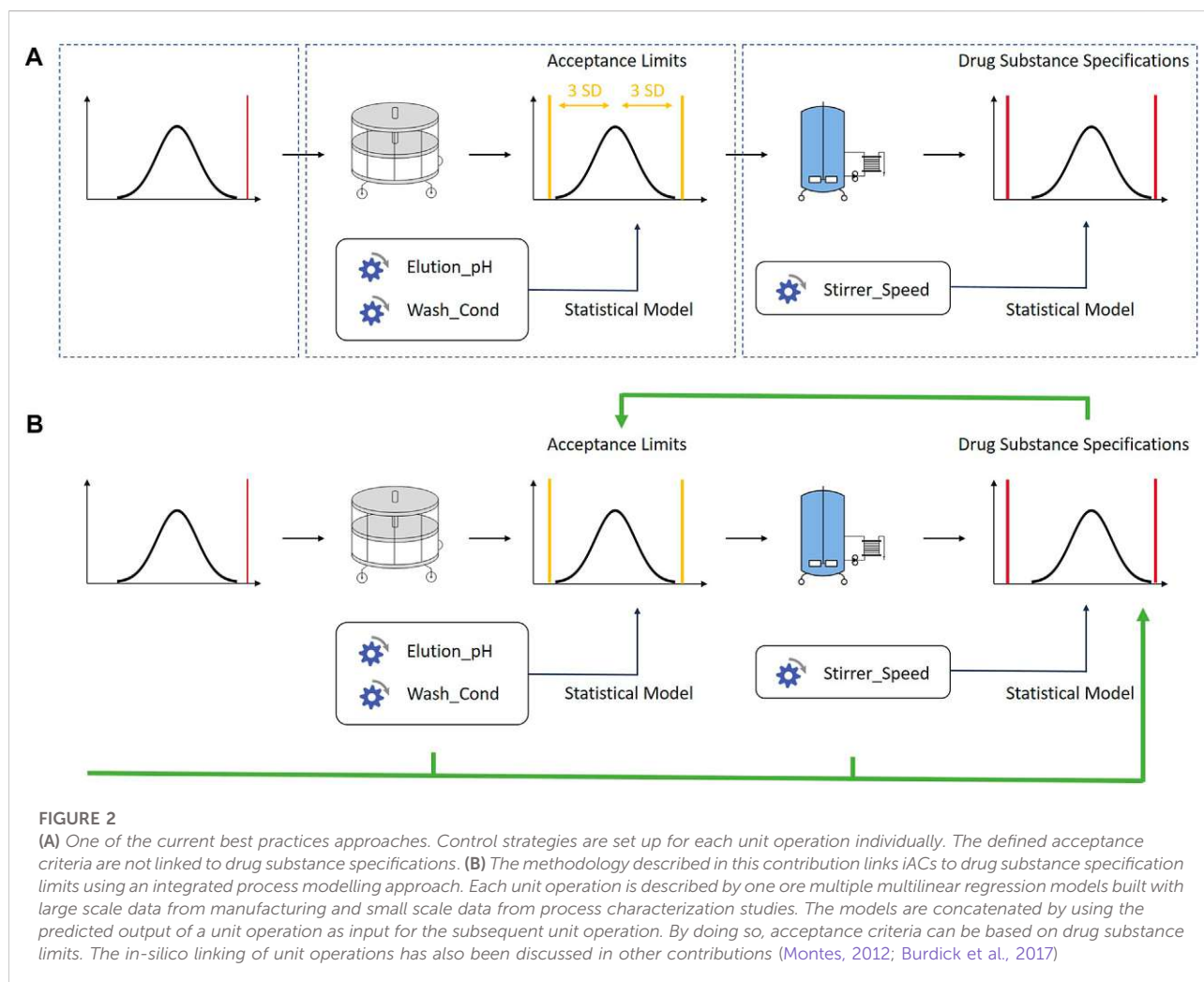
There are currently multiple methods to derive iACs for quality attributes.

One solution to define iACs is by performing wet-lab spiking studies. This is an approach commonly applied in virus clearance studies (Darling, 1993). EMA also explicitly mentions spiking experiments to demonstrate the clearance capacity of downstream unit operations for host-cell related impurities (EMA/CHMP/BWP/187338/2014, 2016). However, finding the correct spike material is difficult, as care has to be taken that the sample matrix is not completely altered by other components contained in the spiking material and correctly represents the material in the naturally occurring process.

Acceptance criteria may also be calculated based on data collected at set-point conditions. They can be calculated by applying ± 3 standard deviations (3SD) of the existing data, or statistical intervals based on an assumed underlying distribution (e.g. tolerance intervals). These approaches do not account for variability around process parameters and don't provide a linkage to drug substance specifications (Seely et al., 2003; Orchard, 2006; Wang et al., 2007). Moreover, both approaches heavily rely on the observed variance. Higher variation leads to wider acceptance criteria and lower variation to tighter limits. Both approaches reward poor process control and punish good process control. Moreover, the mentioned methods are focused on individual unit operations only.

Another approach linking knowledge across multiple unit operations is described by Montes (Montes, 2012). They compare methods to estimate the to-be-expected variance at each process step. One of the discussed approaches is to apply variance transmission. The variance for e.g. process step 3 is calculated by applying error propagation using the known regression models for process steps 1 to 3. The estimated variance is used to calculate tolerance intervals. The worst case side of the tolerance interval (in the case of a two-sided interval) is





then used as acceptance criteria for the respective step. This approach leverages the knowledge of known functional relationships. The defined acceptance criteria give information on the possible worst case of a process at the observed variance. However, they don't give any information how likely it is to meet drug substance criteria.

Monte Carlo approaches have been applied to the definition of specification limits (Burdick et al., 2017). Burdick et al. used the approach to calculate the final distribution of a drug product quality attribute after several storage steps and suggested to use the calculated distribution to derive specification limits.

Ideally, iACs share the following characteristics:

- iACs should provide a link to drug substance or product limits: the likelihood or probability of meeting drug substance specifications while staying within the intermediate acceptance criteria.
- iAC derivation should consider the uncertainty around process parameters and material attributes

Within this contribution, we build upon the concept on integrated process modelling as described by Zahel et al. (Zahel et al., 2017). In an integrated process model (IPM), each unit operation is described by a multilinear regression model where the performance (clearance or purification capability) is the dependent variable and the input of the previous unit operation as well as the process parameters act as independent variables. These models are built with large scale data from manufacturing and small scale data from process characterization studies.

The models are concatenated by using the predicted output of a unit operation as input for the subsequent unit operation. Using Monte Carlo simulation, random variability caused by process parameters can be incorporated into the modeled process (Zahel et al., 2017). IPMs can be used to predict the out-of-specification probability for a given set of process parameter set-points. Another application is to set up a control strategy for process parameters by defining proven acceptable ranges (Taylor et al., 2021).

TABLE 1 Available data sets, process parameters, and monitored critical quality attributes (CQAs) for each unit operation included in the IPM.

Unit Operation	Available datasets	PPs varied in DoEs	Monitored CQAs
Harvest	10 Manufacturing Runs (2 kl)	Load Pool Temperature	HCP ELISA
Capture Chromatography	5 OFAT Runs (3L), 10 Manufacturing Runs (2 kl)		HCP ELISA, UP-SEC Aggregates, UP-SEC Monomer
Virus Inactivation	5 OFAT Runs (3L), 10 Manufacturing Runs (2 kl)	Stirrer Speed	HCP ELISA, UP-SEC Aggregates, UP-SEC Monomer
Depth Filtration	10 Manufacturing Runs (2 kl)	-	HCP ELISA, UP-SEC Aggregates, UP-SEC Monomer
Anion Exchange (AEX) Chromatography	4 OFAT Runs (3L), 10 Manufacturing Runs (2 kl)	Equilibration_pH	HCP ELISA, UP-SEC Aggregates, UP-SEC Monomer
Cation Exchange (CEX) Chromatography	11 DoE Runs (3L), 10 Manufacturing Runs (2 kl)	Elutions buffer Cond, Elutions buffer pH	HCP ELISA, UP-SEC Aggregates, UP-SEC Monomer
Viral Filtration	10 Manufacturing Runs (2 kl)	-	UP-SEC Aggregates, UP-SEC Monomer
Hydrophobic Interaction (HIC) Chromatography	17 DoE Runs (3L), 10 Manufacturing Runs (2 kl)	Loading Pool_pH, Loading Pool_Conductivity, Loading Pool_Temperature	UP-SEC Aggregates, UP-SEC Monomer
Ultra- and Diafiltration	10 Manufacturing Runs (2 kl)	-	UP-SEC Aggregates, UP-SEC Monomer

Within this contribution, we aim to derive iACs that ensure a pre-defined out-of-specification probability. These specification-driven ranges enable the set up of a control strategy that prevents failed batches at highest possible manufacturing flexibility. The novelty of this approach is that the acceptance criteria are based on pre-defined out-of-specification probabilities, while also considering manufacturing variability in process parameters (Figure 2).

The manuscript is structured in two parts. First the developed method is described. In a second step the developed method is applied to a real world case study and compared to a conventional approach.

2 Methods and materials

2.1 Candidate process for case study

For the case study, a monoclonal antibody (mAb) production process in mammalian cell culture was provided by Boehringer Ingelheim in Biberach, Germany. The model depicts the downstream process segment of the drug substance manufacturing process.

The downstream process consists of 9 unit operations. The first step is the pool of the harvested fermentation broth (UO 1), the second step is a chromatographic capture step (UO 2), followed by a viral inactivation (UO 3), depth filtration (UO 4), two chromatographic steps (UO 5, UO 6), a viral filtration (UO 7), another chromatographic step (UO 8) and ultra-and diafiltration (UO 9).

Three quality attributes defined as CQAs were modelled. One product-related impurity (UP-SEC Aggregates) and one host-related impurity (HCP ELISA) that need to be cleared by the downstream process and one parameter, purity (UP-SEC Monomer), that should be increased.

2.2 Data for the integrated process model

For the capture chromatography, the virus inactivation and the anion exchange chromatography one-factor-at-a-time (OFAT) studies were performed. For the cation exchange chromatography 2 factors were investigated in a design of experiments (DoE) approach. One factor was varied in 5 levels and the second factor in 3 levels. One center-point was performed. The design is able to resolve main effects and quadratic effects. For the hydrophobic interaction chromatography 3 factors were investigated in a face-centered central composite design with 3 center points. The design is able to resolve main effects, two-factor interactions and quadratic effects. All experimental studies were performed in small scale.

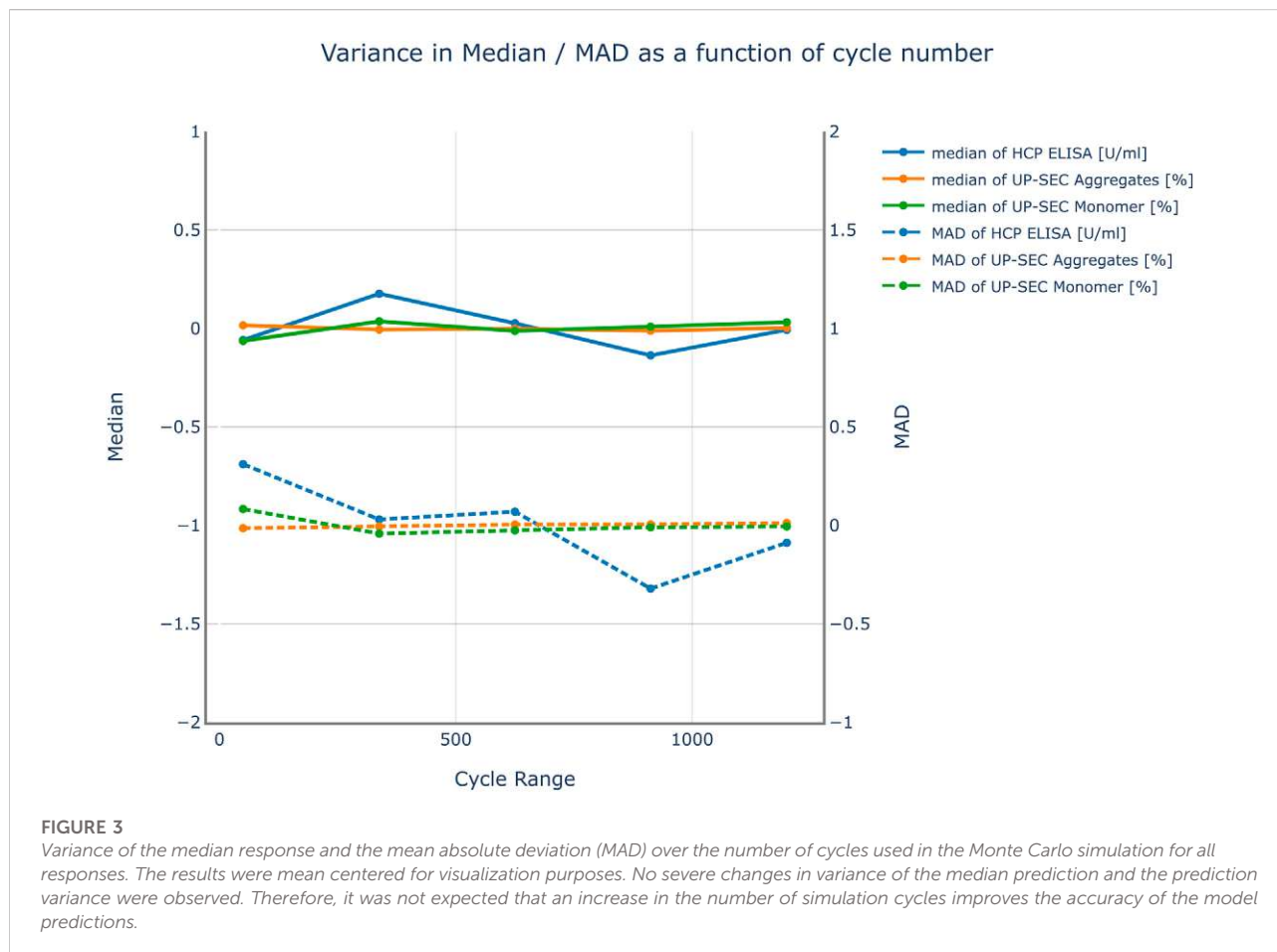
The available data for each unit operation is summarized in Table 1.

2.3 Calculation of performance indicators

Clearance parameters were calculated for each impurity (i) according to Eq. 1, where i is the specific impurity concentration, i.e. units per mg product, in load or pool of the respective process step.

$$SC_i = \text{Specific Clearance}_i = \frac{i_{\text{load}}}{i_{\text{pool}}}$$

For product quantity and purity attributes, yields were calculated according to Eq. 2, where i is the product amount or percentage of desired isoform in load or pool of the respective process step.



$$Y_i = Yield_i = \frac{i_{pool}}{i_{load}}$$

Eq. 2

2.4 Modelling the individual unit operations

Ordinary least squares (OLS) regression was used for statistical analysis. Scale was treated as fixed effect. As dependent variables, clearances and yields were used. A clearance represents the ratio of two assumed-to-be normally distributed random variables. Therefore, a clearance is not normally distributed. After analysis of the residuals it was decided to log-transformed the responses prior to modelling. All independent variables were scaled between -1 and 1 according to Goos and Jones (Goos & Jones, 2011). The independent variable (response) was neither scaled nor centered. For the analysis of the DoEs, a best subset variable selection was applied using a p -value threshold for the partial t-statistic of 0.1. The threshold of

0.1 was chosen as opposed to the commonly applied threshold of 0.05 to minimize the risk of overlooking potentially critical process parameters. A strong heredity principle was followed i.e. if a two-factor interaction is included in the model, the main effects of both factors involved in the interaction are included in the model as well (even if the main effects are not significant with the chosen threshold). To ensure model adequacy, a thorough analysis of the model residuals is performed to check whether any of the assumptions for regression analysis are violated. i.e. the model errors are statistically independent, of constant variance, and normally distributed.

The unit operations were described by the specific clearance or yield for a given quality attribute. Clearances were used to describe the performance of the respective unit operation. Specific clearances are clearances calculated from impurity concentrations that are normalized to the amount of total product. This harbors the advantage that the values are independent of the scale and total volume. The specific clearances and yields were described by OLS models or observed mean values (in the absence of OLS models) with their respective calculated uncertainty. Models describing the

TABLE 2 Summary of models used for modelling each unit operation and each CQA. Models describing the specific clearance as function of process parameters are termed “DoE Model”. Models describing the specification clearance as function of the input material are termed “SC Model”. If neither a functional relationship of specific clearance on process parameters nor on the input material was found, the unit operation was described by the specific clearance observed in manufacturing, termed “Manufacturing SC”.

HCCF	HCP ELISA	UP-SEC monomer	UP-SEC aggregates
Capture	DoE Model+SC Model	Manufacturing SC	Manufacturing SC
Virus Inactivation	Manufacturing SC		
Depth Filtration	SC Model	SC Model	SC Model
AEX	SC Model		SC Model
CEX	DoE Model	DoE Model+SC Model	DoE Model
Viral Filtration		Manufacturing SC	Manufacturing SC
HIC		SC Model	DoE Model
UFDf		Manufacturing SC	
Bulk		SC-Model	Manufacturing SC

specific clearance as a function of process parameters are termed “DoE Model” and were derived from small scale experiments. Models describing the specification clearance as function of the input material are termed “SC Model” and were derived from manufacturing data. If neither a DoE model nor a SC model was available, the specific clearances were described by fitting a normal distribution to the available manufacturing data.

If more than one OLS model was available for a unit operation, both models were used to describe the unit operation. As the specific impurity loading concentration was not included as a factor in the DoE, interaction effects between factors investigated in the DoE and the specific impurity loading concentration were assumed not to be expected.

The linkage of DoE models and specific clearance models was performed as described elsewhere (Zahel et al., 2017). The combination of DoE model and load model predictions was performed according to Eq. 3, where \widehat{SC}_i denotes the specific clearance predicted from DoE model, $\widehat{SC}_i(PP_i)$ denotes the specific clearance predicted from the process parameters, $\widehat{SC}(SLC_i)$ denotes the specific clearance predicted from the specific clearance model using the input concentration from the simulation (SLC_i) and $\widehat{SC}(SLC_{DoE})$ denotes the specific clearance predicted from the specific clearance model using the concentration of the starting material of the DoE (SLC_{DoE}). The runs of a DoE were performed with the same starting material. The DoE model is valid for the concentration of the starting material used in the DoE. Therefore, the change in specific clearance from the DoE start concentration to the simulation input concentration was used as correction factor.

$$\widehat{SC}_i = \widehat{SC}(PP_i) \cdot \frac{\widehat{SC}(SLC_i)}{\widehat{SC}(SLC_{DoE})}$$

Eq. 3

2.5 Linkage of unit operations using the integrated process modelling technology

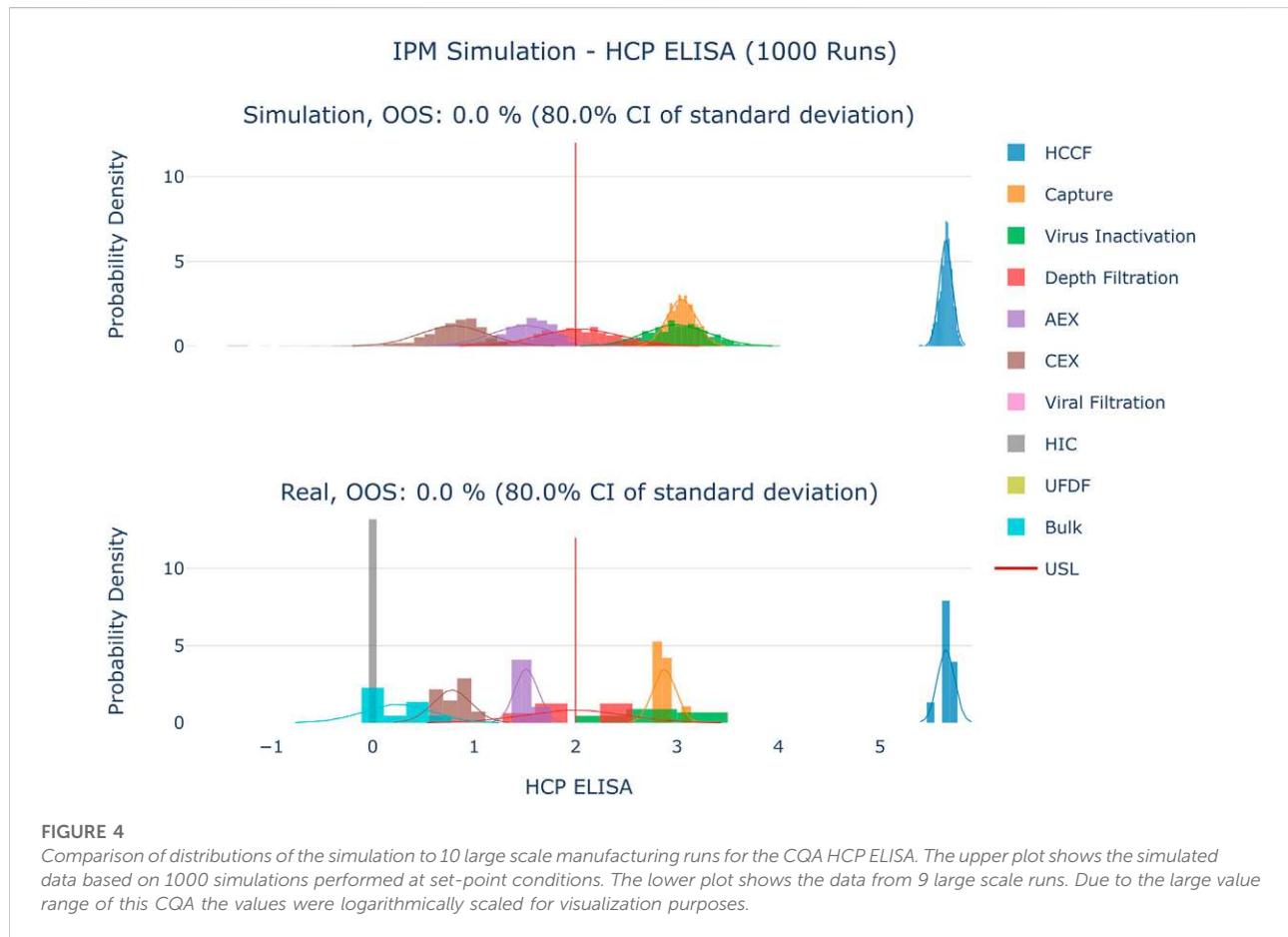
The IPM technology applied in this contribution is described in detail elsewhere (Zahel et al., 2017). The principle behind the IPM is to concatenate models describing the CQA values of individual unit operation together in order to predict the CQA distribution at each intermediate unit operation and ultimately at drug substance.

To account for error propagation during this concatenation, a Monte Carlo simulation is performed in the following way:

A pre-defined number of runs through 9 unit operations are simulated for each response, each using a set of different process parameter values drawn randomly from their normal operating range represented by a normal distribution. Only set-point values of process parameters were available at the time of analysis. Without loss of generality of the approach, the coefficient of variation of each parameter was assumed to be 3%. The technical realization of the normal operating range is given in the results section.

The impact of the number of Monte Carlo runs on the variance of the mean prediction and the prediction variance was investigated for all investigated responses. The results are shown in Figure 3. The impact of the number of simulation runs on the results was investigated in a range from 50 to 1200 simulations. No severe changes in variance of the median prediction and the prediction variance were observed. For that reason, 800 simulation runs were chosen for the subsequent parameter sensitivity analysis. This number leads to simulation cycles that can be conducted in a reasonable amount of time.

Unit operation performances are modelled as a function of process parameters (using OLS) and have some variance associated with them. Using this information, an uncertainty interval is defined around the mean prediction representing the uncertainty of the model prediction. Without loss of generality, 95% prediction intervals were chosen for the IPM. That is, for



each simulated run, a response value is drawn randomly from this uncertainty interval around the mean.

Using the predicted clearance of a unit operation and the available load concentration, the pool concentration is calculated.

Special consideration is given to the simulated load values that fall outside the range of the observed load values used to train the model. Similar to Zahel et al., no extrapolation of clearances outside of the observed load models was performed (Zahel et al.). If simulated load values outside fall outside of the range, the clearance of the unit operation was assumed to be constant. For impurities (CQAs that need to be decreased), this approach might underestimate the clearance for load values higher than the observed range used to fit the models. This is considered more conservative from a risk based approach. For load values lower than the observed range the clearance might be overestimated. For setting up acceptance criteria the load values are gradually increased for each unit operation. For that reason, this case was not observed. For purities (CQAs that need to be increased) the signs need to be reversed.

The overall result of the Monte Carlo simulation with varying process parameters is a distribution for a specific CQA in the pool of the last unit operation, i.e. in drug substance. This distribution may be used to verify OOS event probabilities, given process parameter and model variability.

2.6 Calculation of OOS events

The number of out-of-specification events was calculated according to Taylor et al. (Taylor et al., 2021). A normal distribution was fit to the data. The OOS probability was defined by the area under the curve that lies beyond the drug substance specification limit. The parameters of the normal distribution were the arithmetic mean and the upper 80% confidence interval of the standard deviation. The upper confidence of the standard deviation was used to provide a fair comparison between the simulated runs and the real manufacturing runs, because of the large difference in sample sizes (800 simulated runs vs. 10 manufacturing runs).

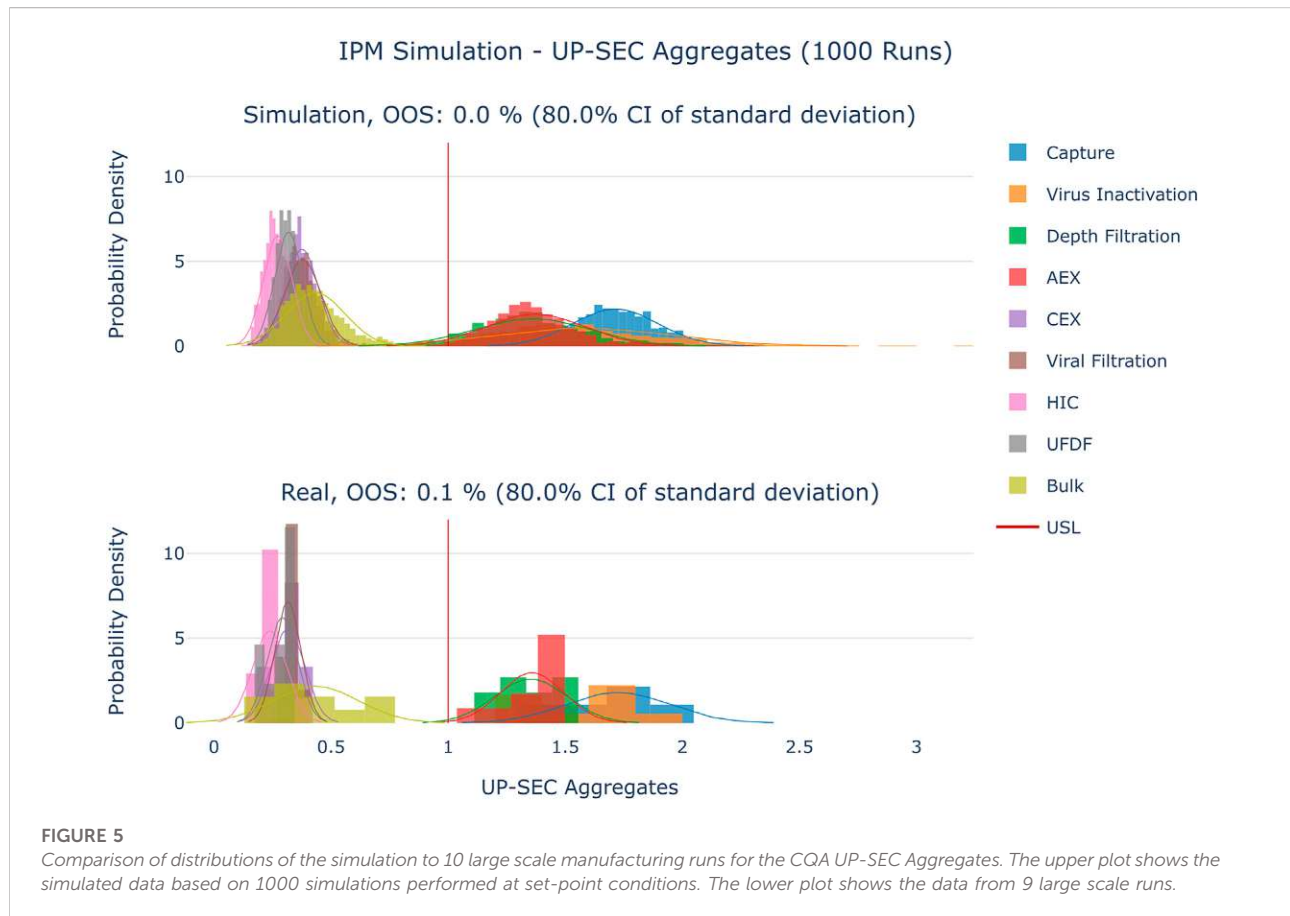


FIGURE 5

Comparison of distributions of the simulation to 10 large scale manufacturing runs for the CQA UP-SEC Aggregates. The upper plot shows the simulated data based on 1000 simulations performed at set-point conditions. The lower plot shows the data from 9 large scale runs.

3 Results

3.1 Description of the integrated process model

A pre-requisite of setting up an IPM is that the quality attributes to be modelled are measured both as input and output of the unit operations under investigation. Due to data availability, the IPM for HCP ELISA was set up from unit operation 1 to unit operation 6. For UP-SEC Aggregates and UP-SEC Monomer the integrated process model was set up from unit operation 2 to unit operation 9. [Table 2](#) outlines how the unit operations were modelled for each CQA.

3.2 Definition of the NOR

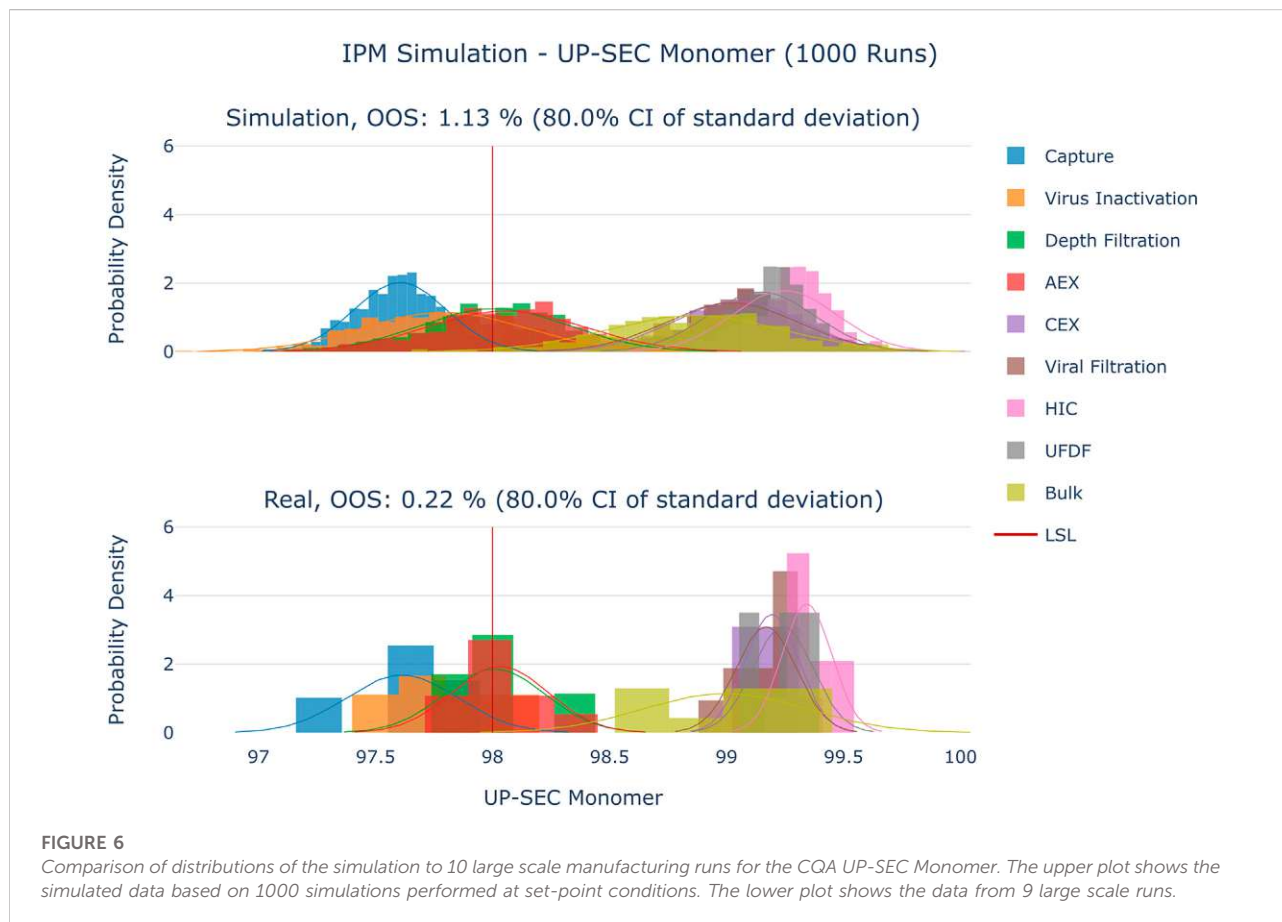
For modeling the process parameters, the definition for the NOR as outlined by FDA and EMA is followed.

“The NOR describes a region around the target operating conditions that contain common operational variability (variability that can’t always be controlled)” (EMA/213746/2017, o. J.).

For the purpose of the ensuing analysis, we aim to provide a technical realization of this definition. To our knowledge no mathematical description of the normal operating range has been given so far.

Without loss of generality, this operational variability is assumed to be caused by experimental errors stemming from several independent, uncontrollable sources. Therefore, it is sufficient to assume that continuous process parameter values follow a normal distribution (with the target operating value (set-point) being the most probable one (mean of the distribution)). This holds true for any targeted continuous process parameter value. For parameters that are controlled in such a way the NOR follows a normal distribution described by two parameters (mean and standard deviation). For parameters that don’t need to meet a target, but are allowed to stay within a range according to manufacturing batch records other distributions might be applicable (such as uniform distributions, poisson distributions or truncated normal distributions).

Each process parameter value has a certain probability of being observed associated with it; the set-point is the most probable value. The process parameter distribution follows a normal distribution around the set-point. The normal operating



range (NOR) of a process parameter is then defined as the lower and upper boundary of the distribution covering a pre-defined area under the curve (e.g. ± 3 standard deviations around the set-point). The values within the NOR are normally distributed (and not uniformly distributed). Following this definition, the normal operating range is a function of the applied set-points and is subject to change in the case the process parameter set-point is changed.

As a consequence, the results of the integrated process model are only valid if the process is controlled at target conditions including the uncertainty (NOR) around it, that is, whereat all PPs are kept at set-point and the process variability (i.e. standard deviation) does not increase.

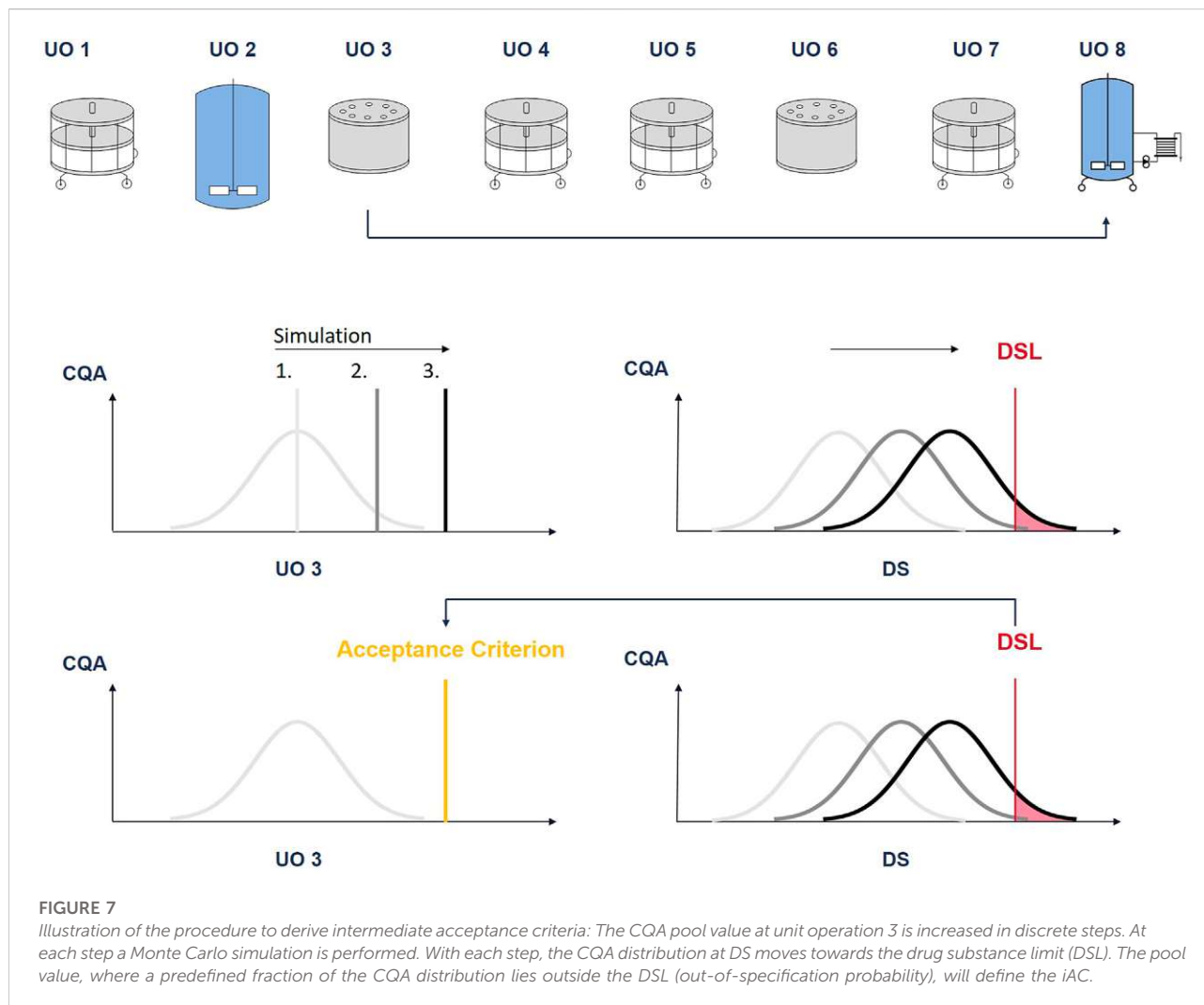
3.3 Plausibility check of the integrated process model

Each individual OLS model was assessed individually based on model statistics R^2 , Q^2 , p -values, and RMSE as described in the material method section. The quality of the simulation with the concatenated models was assessed by comparing the

predictions of the IPM with actually performed manufacturing runs at target conditions. Additionally, the predicted OOS rate was compared to the OOS rate calculated from the manufacturing runs.

The results are shown in Figure 4–Figure 6. The span of the bar in the histograms was normalized in a way that the sum of all bin areas equals 1 (i.e., the area of each bar corresponds to the probability that an event falls into that bin). For that reason, the height of the bars (i.e., the probability densities) between the simulated values and the real data might differ, but the integrals equal 1. Therefore, the y values in these plots are not relevant for comparing the simulation with the real data. For all investigated CQAs, the simulated distributions fit quite well to the available manufacturing data. The predicted OOS probabilities (given in the plot titles) are in the same range as the OOS calculated from the manufacturing data. Based on these results the set-up model framework is regarded as fit for the application of setting up acceptance criteria.

The definition of the Acceptance Criteria in ICH Q6B was followed. The statement “considered acceptable for its intended use” is interpreted in the following way: The drug substance material is considered acceptable for its intended use, if it conforms to the drug substance specification limits.



3.4 Definition and calculation of intermediate criteria

Following the outlined definition, the intermediate acceptance criteria will be defined by performing a parameter sensitivity analysis (PSA) within the IPM simulation framework. It will be assessed how a change in CQA load values in an intermediate unit operation affects out-of-specification (OOS) events at drug substance level.

For each CQA, the PSA was conducted as follows:

- 1) The screening range for the PSA was calculated from available manufacturing data. A range of plus/minus 10 standard deviations around the observed mean in the pool of the unit operation was calculated. The screening range was divided into 15 equidistant segments. If this resulted in negative values, the screening range was decreased by limiting it to positive values only.
- 2) The CQA's pool value of the UO, for which the acceptance criteria are calculated, (= load value of the next UO) is set to a fixed value.
- 3) An IPM Monte Carlo simulation consisting of 800 simulated runs was performed according to the description in section 3.1.2, where all process parameters are randomly drawn from their normal operating range.
- 4) The number of OOS results for the CQA and a corresponding OOS probability is calculated.
- 5) Steps 2-5 are repeated for each of the screening range segments defined in step 1.
- 6) The intermediate acceptance criteria is then defined by the CQA pool concentration that results in the pre-defined OOS probability.

The procedure is then repeated for each CQA in each UO. An illustration of this procedure is given in Figure 7.

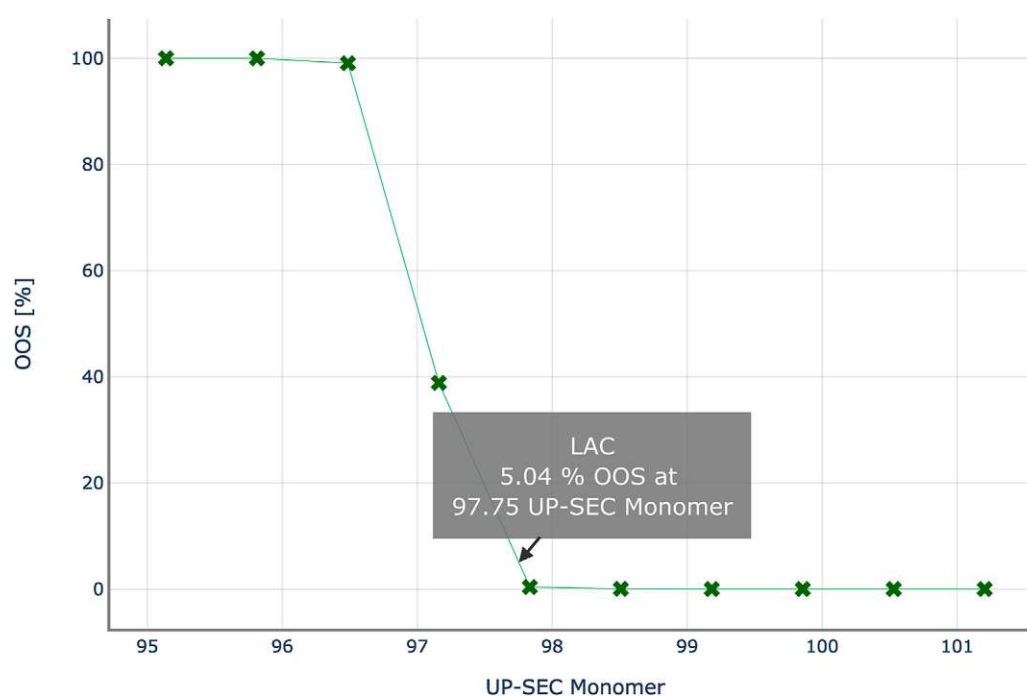


FIGURE 8

Out-of-specification probability at drug substance for various theoretical UP-SEC Monomer pool values in unit operation 2. Intermediate acceptance criteria are defined, as the CQA's pool value, for which the out-of-specification probability equals a critical threshold, here 5%.

For the case study an OOS probability of 5% was defined as threshold.

3.5 Case study—Comparison of approaches for setting up acceptance criteria

Figure 8 shows the results of the PSA to determine the intermediate acceptance criteria or UP-SEC Monomer in unit operation 2. For each data point, i.e. for a specific CQA pool value, CQA distributions at DS are predicted, and the probability to generate an out of-specification (OOS) limit is calculated. The OOS probability is then plotted as a function of the pool value. With each step the CQA distribution at drug substance moves towards the specification limit, increasing the risk of OOS events. At a 5% OOS probability, the proposed upper iAC for UP-SEC Monomer at unit operation 2 is 96.71% for the lower specification limit of 98%.

This procedure was followed for all unit operations and all CQAs under investigation. The corresponding plots are provided in the appendix.

In the case study the IPM derived acceptance criteria were compared to acceptance criteria based on \pm three standard deviations. For all three responses only 1-sided specification

criteria were defined. For this reason, the IPM derived acceptance criteria are also 1-sided. For impurities (HCP ELISA and UP-SEC Aggregates) an upper limit was defined and for purities (UP-SEC Monomer) a lower limit was defined.

Due to the large value range of HCP ELISA, the values were logarithmically scaled for visualization purposes (Figure 9). For HCP ELISA, the IPM derived acceptance criteria were higher than the upper three standard deviation limits in all investigated unit operations. Especially in the first four unit operations the three standard deviation derived limits are much tighter than the IPM derived limits. Runs that fall outside the 3SD limit might still exhibit an acceptable out-of-specification probability. If these 3SD limits are applied, it might lead to the issue that alerts are raised unnecessarily. Except for unit operation 7 at the last five unit operations no data was available for HCP ELISA. At unit operation 7 CQA measurements are available, however they all represent one value: the limit of quantification. For that reason no standard deviation could be calculated and integrated process modeling could not be applied. The intermediate acceptance criteria were therefore set equal to the drug substance specification limits. This approach relies on the assumption that the impurity does not increase in these unit operations.

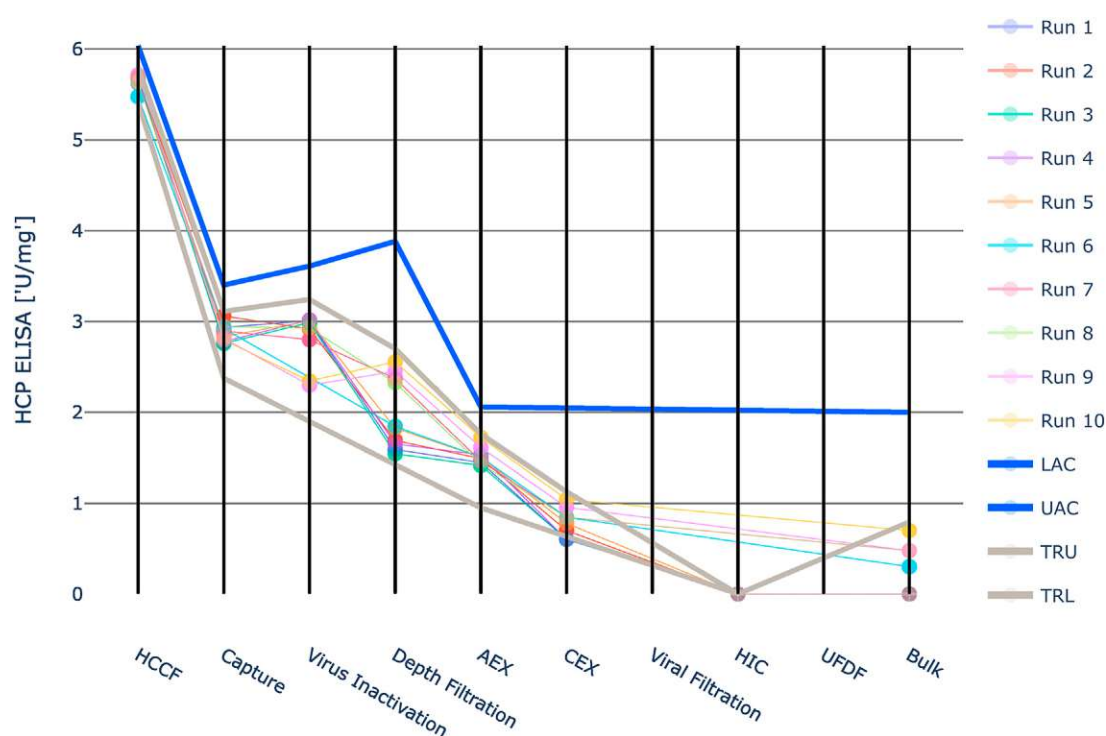


FIGURE 9
 Graphical representation of the intermediate acceptance criteria (blue line) across the entire downstream process for the response HCP ELISA. Available large scale, manufacturing data per batch (circles) and three standard deviation ranges (grey lines) are given as well. The shown iACs at DS are the DS specification limits. Due to the large value range of this HCP ELISA the values were logarithmically scaled for visualization purposes. Except for unit operation 7 at the last five unit operations the intermediate acceptance criteria were set equal to the drug substance specification limits. At unit operation 7 CQA measurements are available, however they all represent one value: the limit of quantification. For that reason no standard deviation could be calculated and integrated process modeling could not be applied.

For UP-SEC Aggregates the IPM derived acceptance criteria were higher than the upper three standard deviation limits in all investigated unit operations (Figure 10). As described for the previous CQA if 3SD limits are applied, it might lead to the issue that alerts are raised unnecessarily.

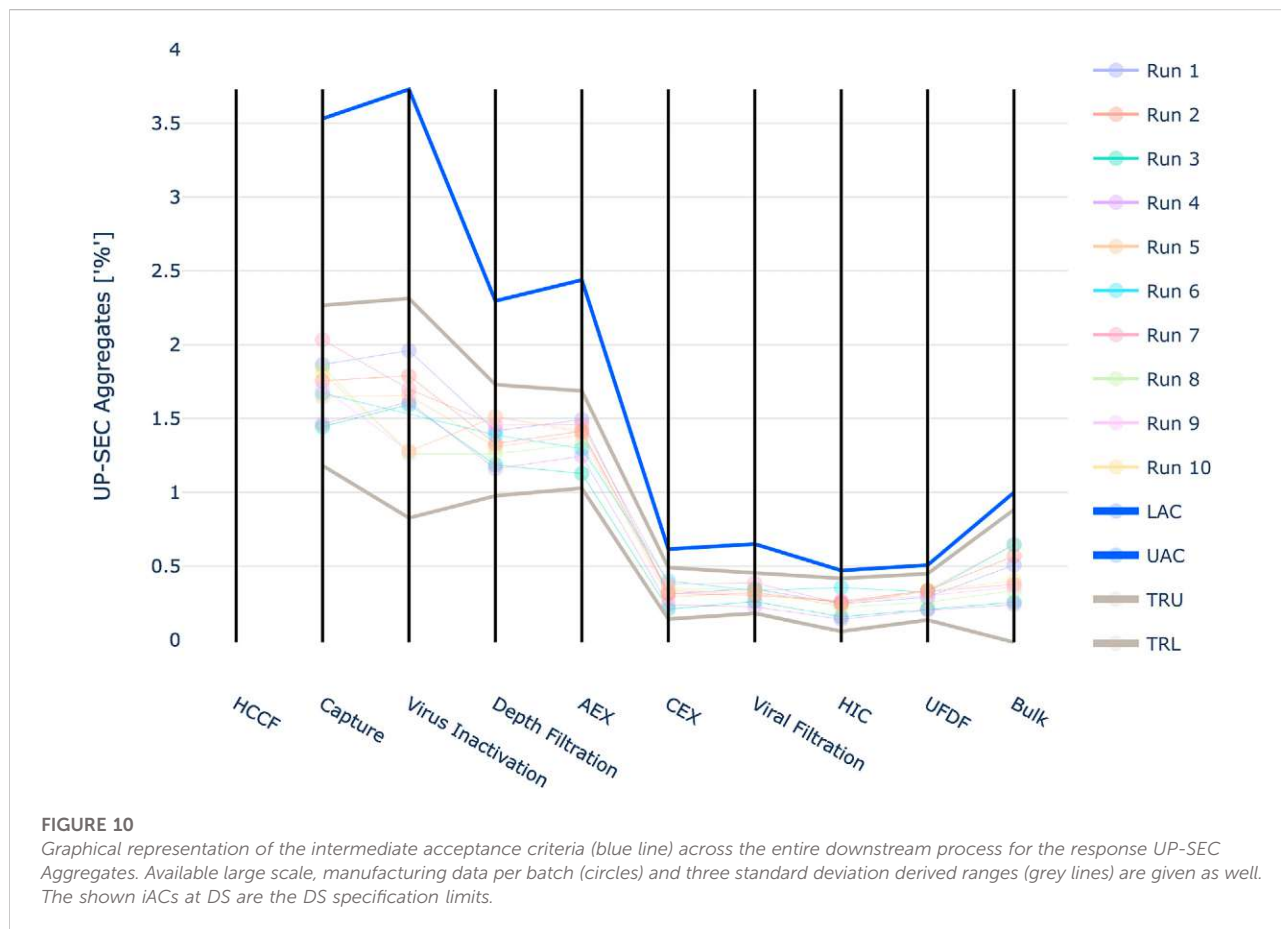
For UP-SEC Monomer the IPM derived acceptance criteria lie close to the observed manufacturing values in unit operations 1 to 4 (Figure 11). For unit operation 2 a manufacturing run falls even below the acceptance criteria, although it still meets the final drug substance specification limit. The definition of the intermediate acceptance criteria is based on a probabilistic approach, i.e. at the intermediate acceptance criterion, there is a certain probability (here 5%) that the CQA does not meet drug substance specification limits. Consequently, even if a manufacturing run is close to the proposed intermediate acceptance criteria, this does not necessarily lead to the run being out of specification at DS. If it lies exactly at the intermediate acceptance limit, there is still a 95% probability that the run is within the specification limit.

Additionally, the lower limit of the three standard deviation derived ranges is lower than the IPM derived acceptance criteria

for the first four unit operations. Based on the results of the IPM this means that the out of specification probability is larger than 5% at these limits. For unit operation 3 the lower 3 SD limit is 97.1%. At this value the IPM yields a 14.9% out of specification probability. If 100 runs were close to the lower 3 SD limit 14.9 runs would not meet the specification criteria at drug substance.

4 Discussion

Many contributions elaborate on methods to set up control strategies for process parameters (e.g. design space) (Abu-Absi et al., 2010; Jiang et al., 2010). A prerequisite for that is the knowledge, which levels of quality attributes are acceptable. Acceptance criteria serve as backbone for a proper control strategy on process parameters and material attributes. Often irrespective of the control strategy methodology, the reader is left alone in setting up acceptance criteria. Additionally EMA requires acceptance criteria for CPPs and CQAs to be part of the process validation protocol, which should be based on



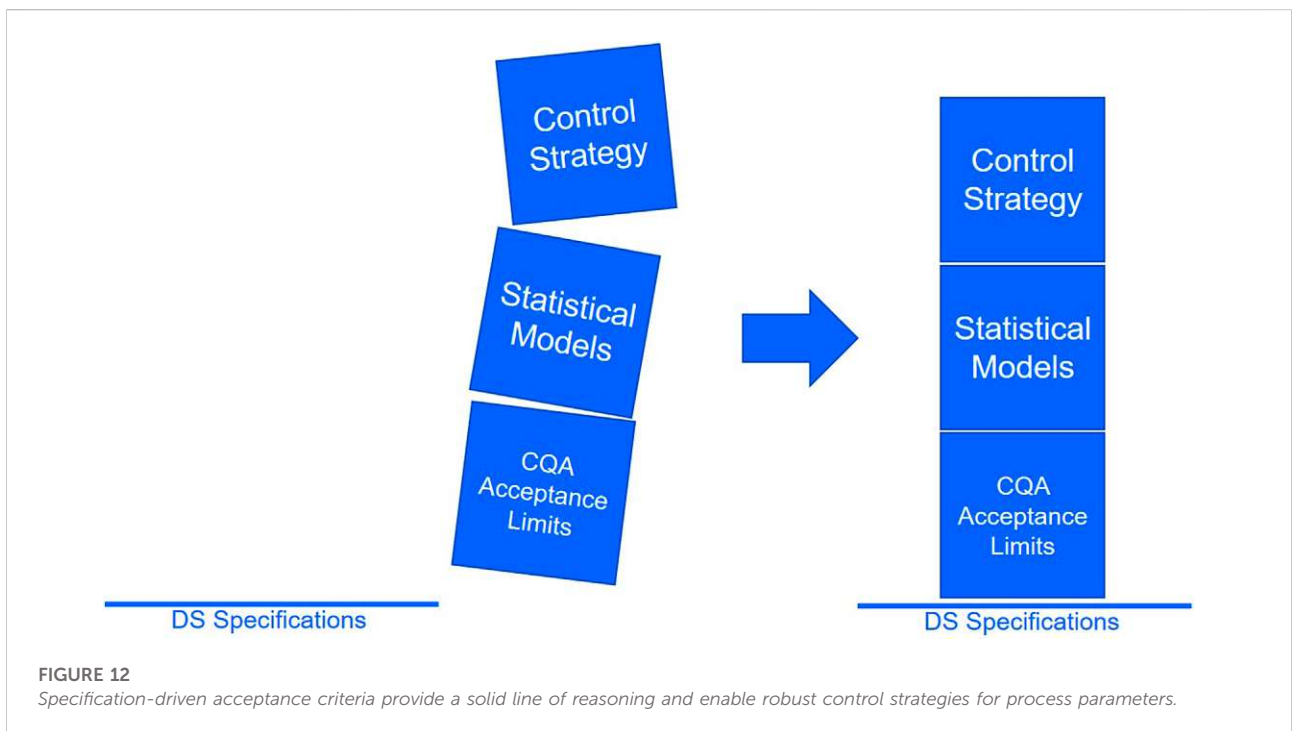
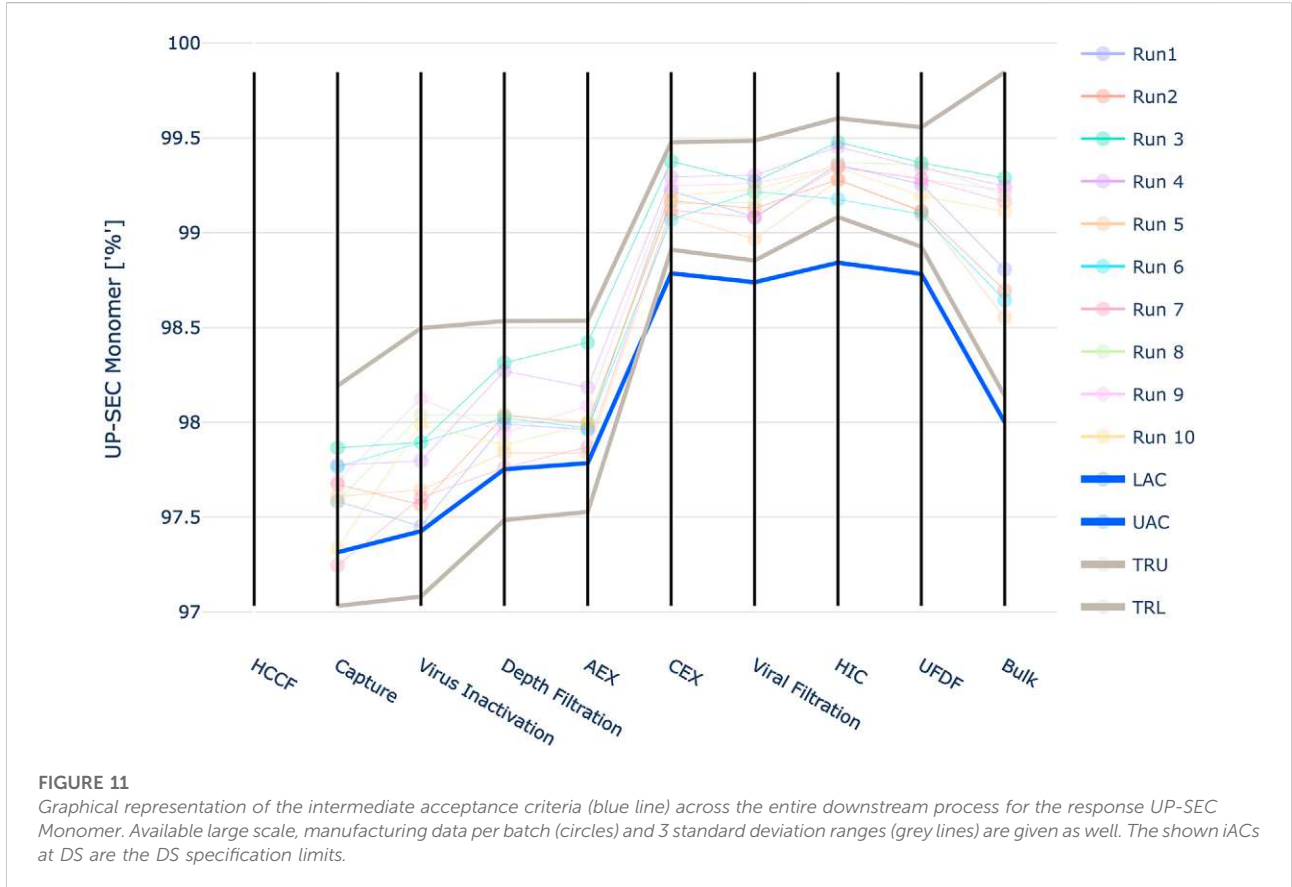
development data or documented process knowledge (European Commission 2015). However, here no specific guidance is provided to derive such limits.

Within this contribution we refined the definition of acceptance criteria by ICH Q6B by further specifying the term “for intended use” to having a link to final specification limits used for drug substance release (ICH Q6B, 1999). Additionally, we presented a methodology to calculate intermediate acceptance criteria based on drug substance specification limits and considers uncertainty around process parameters.

It should be emphasized that IPM derived acceptance criteria are only valid for a defined set of process parameter conditions. This means that if acceptance criteria were defined based on manufacturing runs at set-point conditions, they are only valid for runs that are performed at set-point. In the case of process changes, intermediate acceptance criteria need to be revised. This not only applies to the method presented in this contribution, but also applies to other approaches that rely on historic manufacturing data such as approaches that rely on min-max ranges, +/- 3 standard deviations, or statistical intervals (Seely et al., 2003; Orchard, 2006; Wang et al., 2007). Approaches that include data where variance was purposefully introduced into process parameters, as done in process development or process

characterization studies, offer the advantage that the established models can easily be used to calculate acceptance criteria for the new process set-points without the need of acquiring new data (Montes, 2012; Burdick et al., 2017). Updating the acceptance criteria is in line with ICH Q8, which states that acceptance criteria can be updated in the case new process knowledge is available (ICH Q8 (R2), 2009, S. 8). Whereas ICH Q8 states that they should be updated in the case new process knowledge is available, we want to emphasize that they also need to be updated if process changes are implemented (e.g. process parameter set-points).

In this contribution we used OLS regression models to describe the individual unit operations. At the time of the case study the experimental work has already been conducted. The performed OFATs and DoEs were designed to be analyzed using OLS regression. This technique is the standard method for the analysis of DoEs. Care has to be taken, when extrapolation beyond the training range is performed. However, the described methodology for setting up acceptance criteria is not limited to OLS models. If mechanistic models are available model-based DoE approaches could be applied and the functional relationship between quality attributes and process parameters could be



described by purely mechanistic or hybrid models (Kroll et al., 2017; Nold et al., 2021). For model-based approaches capturing the prediction uncertainty is not straight-forward and novel methods to do so are discussed in scientific literature (Briskot et al., 2019). However, an in-depth comparison of modelling approaches is beyond the scope of this contribution.

In addition to suitable data, the presented method requires knowledge in programming or scripting languages to concatenate the individual OLS models and perform the Monte Carlo simulations. In contrast, to that the 3SD approach can easily be applied in table calculation programs like MS Excel. Despite the complex knowledge required, we believe that the benefit of being able to leverage all available process knowledge in the form of statistical models in the integrated process model outweighs the increased analysis effort. Additionally, setting up integrated process models can be automated dependent on the digital maturity of the companies. If quality data and process parameter values are automatically collected in a centralized system the process of setting up an integrated process model can be facilitated.

The available guidelines encourage basing the definition of limits on the entirety of process knowledge. ICH Q6E states “In this respect, limits are justified based on critical information gained from the entire process spanning the period from early development through commercial scale production.” (ICH Q6B, 1999). ICH Q8 further emphasizes the fact that it should be justified how in-process controls contribute to the final product quality (ICH Q8 (R2), 2009, S. 8). ICH Q11 states that links between process and quality is needed (ICH, 2012, S. 11). The above approach puts the guidelines recommendations into practice. It combines the knowledge from small scale studies and manufacturing runs. Functional relationships of quality and process parameters are included. The results are based on drug substance specification criteria. Following the principle of the control strategy lifecycle as outlined in ICH Q8, acceptance criteria can be updated using the IPM as new knowledge is available (ICH Q8 (R2), 2009, S. 8).

The presented IPM approach models independently from each other. Hence, it relies on the assumption that there are no interactions between the studies quality attributes. This could be addressed by studying various CQA starting concentrations in wet-lab experiments and modelling CQAs as function of other CQAs. In that way multivariate range can be set up that not only consider multivariate dependencies on process parameters but also on other CQAs.

Currently most specifications are based on process variability and not patient-driven. We'd like to see future work that focuses on how to define drug substance/product specifications that are based on patient response (safety and efficacy). In order to achieve this, manufacturing data should be linked to data from the clinic. Additionally, the quantity and quality of the data is important.

The aforementioned aspects of the IPM derived acceptance criteria provide a solid line of reasoning for justification in audits as they are built on the total amount of available evidence, while using already well established modelling techniques (i.e. OLS). The described methodology enables the definition of acceptance criteria based on the probability of reaching the specification limits. We therefore firmly support using specification-driven acceptance criteria form a solid base for activities in setting up control strategies (Figure 12). The IPM derived acceptance criteria may prove to be an excellent foundation for the establishment of patient centric specifications as correlations between product attributes and clinical outcomes are made.

Data availability statement

The datasets presented in this article are not readily available because the data used in this study was generated for a commercial manufacturing process. The rights to the dataset are owned by the company. Requests to access the datasets should be directed to lukas.marschall@koerber.com.

Author contributions

LM designed and implemented the acceptance criteria functionality and the modelling framework, conducted the case study and drafted the manuscript. CT performed the statistical modelling, designed the acceptance criteria functionality, conducted the case study and drafted the manuscript. CT and LM contributed equally to this manuscript. TZ had the idea for the methodology and implemented core functionalities in the code framework and assisted in writing the manuscript. MK provided key statistical and process expertise and functioned as the liaison between the two partnering companies. BP, JV, JS and CH assisted in the writing and review of the manuscript. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Program.

Conflict of interest

Authors MK, AW, BP, and JS were employed by Boehringer Ingelheim Pharma GmbH. Authors LM, CT, and TZ were employed by Koerber Pharma Software. This work was

funded by Boehringer Ingelheim Pharma GmbH in the course of a project with Koerber Pharma Software.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fbioe.2022.1010583/full#supplementary-material>

References

- Abu-Absi, S. F., Yang, L., Thompson, P., Jiang, C., Kandula, S., Schilling, B., et al. (2010). Defining process design space for monoclonal antibody cell culture. *Biotechnol. Bioeng.* 106 (6), 894–905. doi:10.1002/bit.22764
- Briskot, T., Stückler, F., Wittkopp, F., Williams, C., Yang, J., Konrad, S., et al. (2019). Prediction uncertainty assessment of chromatography models using Bayesian inference. *J. Chromatogr. A* 1587, 101–110. doi:10.1016/j.chroma.2018.11.076
- Burdick, R. K., LeBlond, D. J., Pfahler, L. B., Quiroz, J., Sidor, L., Vukovinsky, K., et al. (2017). *Statistical applications for chemistry, manufacturing and controls (CMC) in the pharmaceutical industry*. New York: Springer International Publishing. doi:10.1007/978-3-319-50186-4
- Darling, A. J. (1993). Considerations in performing virus spiking experiments and process validation studies. *Dev. Biol. Stand.* 81, 221–229.
- EMA-FDA. EMA/213746/2017. (o. J.). EMA-FDA Questions and Answers: Improving the understanding of NORs, PARs, DSp and normal variability of process parameters. *Quest. Answers* 4.
- EMA/CHMP/BWP/187338/2014 (2016). *Process validation for the manufacture of biotechnology-derived active substances and data to be provided in the regulatory submission*.
- European Commission (2015). *Annex 15: Qualification and validation*.
- FDA (2011). Guidance for industry. Available at: <https://www.fda.gov/downloads/drugs/guidances/ucm070336.pdf>.
- Goos, P., and Jones, B. (2011). *Optimal design of experiments: A case study approach*. Wiley.
- ICH (2012). *Development and manufacture of drug substances (chemical entities and biotechnological/biological entities)—Q11—step 4*. Available at: http://sh.st/st/787f28ed3e745c14417e4aec27303038/www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q11/Q11_Step_4.pdf.Q11 (step 4)
- ICH Q6B (1999). *Specifications: Test procedures and acceptance criteria for biotechnological/biological products—Q6B - current Step 4 Version*.
- ICH Q8 (R2) (2009). *Pharmaceutical development Q8 (R2)*.
- ICH (2005). *Quality risk management Q9—step 4*.
- Jiang, C., Flansburg, L., Ghose, S., Jorjorian, P., and Shukla, A. A. (2010). Defining process design space for a hydrophobic interaction chromatography (HIC) purification step: Application of quality by design (QbD) principles. *Biotechnol. Bioeng.* 107 (6), 985–997. doi:10.1002/bit.22894
- Kroll, P., Hofer, A., Ulonska, S., Kager, J., and Herwig, C. (2017). Model-based methods in the biopharmaceutical process lifecycle. *Pharm. Res.* 34 (12), 2596–2613. doi:10.1007/s11095-017-2308-y
- Montes, R. O. (2012). Variation transmission model for setting acceptance criteria in a multi-staged pharmaceutical manufacturing process. *AAPS PharmSciTech* 13 (1), 193–201. doi:10.1208/s12249-011-9734-7
- Nold, V., Junghans, L., Bisgen, L., Drerup, R., Presser, B., Gorr, I., et al. (2021). Applying intensified design of experiments to mammalian cell culture processes. *Eng. Life Sci.*, 202100123. doi:10.1002/elsc.202100123
- Orchard, T. (2006). Specification setting: Setting acceptance criteria from statistics of the data. *Biopharm. Int.* 19 (11), 40–45.
- Seely, R. J., Munyakazi, L., and Haury, J. (2003). Statistical tools for setting in-process acceptance criteria. *Dev. Biol.* 113, 17–25.
- Taylor, C., Marschall, L., Kunzelmann, M., Richter, M., Rudolph, F., Vajda, J., et al. (2021). Integrated process model applications linking bioprocess development to quality by design milestones. *Bioengineering* 8 (11), 156. doi:10.3390/bioengineering8110156
- Wang, X., Germansderfer, A., Harms, J., and Rathore, A. S. (2007). Using statistical analysis for setting process validation acceptance criteria for biotech products. *Biotechnol. Prog.* 23 (1), 55–60. doi:10.1021/bp060359c
- Zahel, T., Hauer, S., Mueller, E. M., Murphy, P., Abad, S., Vasilieva, E., et al. (2017). Integrated process modeling—a process validation life cycle companion. *Bioengineering* 4 (4), 86. doi:10.3390/bioengineering4040086

3.4 Architectural & Technological Improvements to Integrated Bioprocess Models towards Real-Time Applications

Article

Architectural and Technological Improvements to Integrated Bioprocess Models towards Real-Time Applications

Christopher Taylor ^{1,2,†} , Barbara Pretzner ^{1,2,†} , Thomas Zahel ¹ and Christoph Herwig ^{1,2,3,*} ¹ Körber Pharma Austria GmbH, 1070 Vienna, Austria² Research Area Biochemical Engineering, Vienna University of Technology, Gumpendorferstrasse 1a, 1060 Vienna, Austria³ Competence Center CHASE GmbH, Altenbergerstraße 69, 4040 Linz, Austria

* Correspondence: christoph.herwig@tuwien.ac.at; Tel.: +43-676-473-7217

† These authors contributed equally to this work.

Abstract: Integrated or holistic process models may serve as the engine of a digital asset in a multistep-process digital twin. Concatenated individual-unit operation models are effective at propagating errors over an entire process, but are nonetheless limited in certain aspects of recent applications that prevent their deployment as a plausible digital asset, particularly regarding bioprocess development requirements. Sequential critical quality attribute tests along the process chain that form output–input (i.e., pool-to-load) relationships, are impacted by nonaligned design spaces at different scales and by simulation distribution challenges. Limited development experiments also inhibit the exploration of the overall design space, particularly regarding the propagation of extreme noncontrolled parameter values. In this contribution, bioprocess requirements are used as the framework to improve integrated process models by introducing a simplified data model for multiunit operation processes, increasing statistical robustness, adding a new simulation flow for scale-dependent variables, and describing a novel algorithm for extrapolation in a data-driven environment. Lastly, architectural and procedural requirements for a deployed digital twin are described, and a real-time workflow is proposed, thus providing a final framework for a digital asset in bioprocessing along the full product life cycle.

Keywords: integrated process model; digital twin; Pharma 4.0; bioprocess; control strategy; upstream; downstream; real time; holistic model; data science



Citation: Taylor, C.; Pretzner, B.; Zahel, T.; Herwig, C. Architectural and Technological Improvements to Integrated Bioprocess Models towards Real-Time Applications. *Bioengineering* **2022**, *9*, 534. <https://doi.org/10.3390/bioengineering9100534>

Academic Editors: Ulrich Kruhne and Carina L. Gargalo

Received: 6 September 2022

Accepted: 6 October 2022

Published: 9 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Background of Integrated Process Models

In recent years, bioprocess research & development has been seeking to speed up the time to market through the advanced analytical modeling of development data. Of particular focus is the ability to predict final drug quality with minimal data input. One promising technology is integrated process models (IPMs, also referred to as holistic models). These are in silico model frameworks of multistep processes used to perform simulations that predict the behavior and outcome of a full process chain [1,2]. A digital twin (DT) is effectively an extension of this technology, which feeds the resulting output data back into the model in real time [3]. The key components in building a DT are the physical asset (i.e., the process), the digital asset (DA, i.e., the model), and the bidirectional connectivity between them to exchange data and enable a control loop [4]. This concept was mentioned as early as 2003, but has been receiving increasing attention in industry in recent years, not least in the pharmaceutical and biotechnology sectors [1,5–9]; extensive descriptions can be read elsewhere [10]. With IPMs serving as the DA component to a DT, the industrial potential is clear. By leveraging a digital copy of the process where simulations replace physical experiments limited only by computational power, process success can be maximized, and failures may be swiftly mitigated. For bioprocesses and

bioproduct lifecycles, an IPM can substantially shorten development and improve quality both in terms of speed to market and manufacturing success rates [1,11,12].

Modeling individual process steps or unit operations (UOs) in succession have a long history, starting from simple linkage studies [13–15]. However, until recently, few comprehensive frameworks had been established in biopharma development. In 2017, a baseline IPM technology was proposed (IPM 1.0) to serve as a bioprocess ‘life cycle companion’ with the potential to be a DA. In this framework, the bioprocess is constructed by concatenating individual UO models in a central repository that statistically depicts the entire process in the correct process order [16]. Each model represents a single UO with process parameters (PPs) as input factors and critical quality attributes (CQAs) as responses. Once established, the model serves as a “mirror” to the physical asset [10]. Monte Carlo (MC) applications are then leveraged to simulate the propagation of error across the process on the basis of the variation in input factors and subsequent responses. The final simulation result is obtained at drug substance.

IPM 1.0 was trained primarily on specific clearance measurements in characterization data at a small scale – usually performed within a Design of Experiment (DoE) – and in the limited available large-scale (LS) manufacturing data, though the framework also accepted mechanistic and hybrid models. The two scales were fitted into separate matrices and combined to create a single output prediction per variable. This two-matrix system has the disadvantage that the two models require a secondary mathematical step to combine the results. This both leaves any scale offset unaddressed and results in a non-normally distributed result during simulation due to the multiplication and division of the random variables. Equation (1) defines the j -th CQA’s predicted specific clearance (\hat{SC}) as a ratio of the SLC_l (large-scale) and mean SLC_{DoE} (small-scale) results, at a given process parameter setting ($\hat{SC}(PP_i)$).

$$\hat{SC}_j = \hat{SC}(PP_i) \times \frac{\hat{SC}(SLC_l)}{\hat{SC}(SLC_{DoE})} \quad (1)$$

In the population of simulated values where these terms are both normal distributions, the resulting simulated SC_j distribution is Cauchy distribution. Furthermore, this ratio is multiplied by the predicted PPs specific clearance ($\hat{SC}(PP_i)$). The final distribution is, therefore, a product distribution that is proportional to, but not per se, a normal distribution. This relationship can potentially give a less precise estimator of the final resulting simulated distributions and be biased versus a normally distributed predicted result [17].

IPM 1.0 also addressed only non-scale-dependent variables, such as those representing specific clearances, as mentioned above. This is useful in establishing the technology, as all the responses are easily linked with identical units across all UOs. This method also circumvents the issue of modeling volumes that are usually difficult to model since they are controlled by manufacturing and organization considerations. As a consequence, however, this limits the modeling of key process attributes such as *Yield* or *Product Amount*, which is of particular business interest.

1.2. State of the Art for Holistic Bioprocess Models

Since the introduction of IPM 1.0, additional MC applications have been introduced that target specific regulatory deliverables. These include estimating out-of-specification (OOS) results, defining control strategy elements such as proven acceptable ranges (PARs), and linking sensitivity analyses to quality-by-design (QbD) milestones [11].

Recent alternative approaches have also been studied with the goal of both comprehensively describing the process chain and meeting regulatory submission requirements.

Flowsheet models have been proposed for small-molecule pharmaceuticals that, while very similar to the IPM 1.0, differ in the selection of linkage variables used to concatenate the UOs. In one recent case study, using models based on first principles, output responses were directly translated into input variables for the subsequent UO’s mechanistic model [18]. This approach has the flexibility that response variables do not necessarily need to be

simulated across all UOs. Indeed, certain responses may be modeled only for use as an input factor in a different response's model, with all pathways leading to a potentially different final output CQA. This permits modeling flexibility, where each response is not necessarily assayed across all UOs. In bioprocess applications, it is of particular importance to consider a mechanism for estimating parameter and model uncertainty in the prediction, as heightened variation is inherent to the biosystem and significantly impacts the precision of the predictions [19].

Toolboxes of hybrid modeling techniques have recently been proposed that allow for maximally parsing relevant values at different scales. In one instance, an upstream UO was assessed at four different scales, with a sequential procedure for analyzing the multidimensional data to proceed with each subsequent experiment in a pathway optimized to reduce experimental load. This harbors the advantage of directly addressing the quantitative and qualitative differences of scale, and works towards a holistic process evaluation. It is efficient to combine the scales (as opposed to modeling them separately) for three reasons: the degrees of freedom increase, the manufacturing design space is more accurately represented, and the scale offsets can be measured directly. Nonetheless, the framework still needs to offer a linkage between the different UOs in order to address the ultimate impact on CQAs at drug substance [20].

This linkage has recently been assessed in a Bayesian framework for concatenating UOs. Here, the outcomes of potentially multiple models (or one model trained on bootstrapped data) are leveraged as uninformed prior distributions and, using Markov chain MC algorithms, are transformed into a posterior distribution. Random sampling from this distribution is used for the transfer to subsequent UOs. One advantage here is the combination of multiple models per UO, which may be useful in creating more robust predictive outcomes, especially in data-poor environments [21]. One consideration to add to this framework is the prediction of extreme model outputs. Such values are likely outside the training dataset range, but are probabilistically inevitable. This is particularly important for the variable with the most impact on the linkage between UO models. In case of an extreme linkage value, a de facto extrapolation occurs in the second UO, which is highly discouraged in data-driven environments. In any future manufacturing state, potentially extreme results and their impact on subsequent UOs should be considered on risk management grounds. This extrapolation is not performed at the moment in any data-driven holistic process model of which we are aware.

Lastly, to the best of our knowledge, none of these recent bioprocess use cases proposes an integrated real-time application, particularly in commercial manufacturing where the effects of PP deviations in an ongoing process can be simulated onto final drug substance specifications. Such a prediction would provide actionable information to optimize or mitigate process outcomes. Enabling this application would have the potential to increase the process success rate and shorten the time to market. The collection of these innovations would provide a robust platform on which to build a real-time simulation, prediction, and feedback loop. Such a technology would ultimately provide bioprocesses with a plausible DT.

1.3. Suggested Improvements

Each of the recent approaches has significant advantages within the context of bioprocess development requirements. This contribution aims to leverage them collectively to establish a novel IPM that solves numerous challenges in one framework:

- Simplification and improvement of the IPM 1.0 two-matrix procedure.
- Combination of manufacturing- and development-scale data.
- Establishment of scale-dependent variable procedure.
- Improvement of model uncertainty intervals.
- Creation of an extrapolation procedure for non-controllable parameters.
- Description of a real-time DA application.

This contribution proposes building the above improvements on the conceptual backbone of IPM 1.0. The resulting technology would lead to a plausible DA for a bioprocess development DT. Computational comparison with previous approaches is not within scope here, as the primary goal is to create a framework that combines all the above improvements.

Figure 1 compares the above-discussed limitations with the proposed innovations. The top model shows the structure of the IPM 1.0, whereas the lower model depicts the proposed innovations (IPM 2.0) to be discussed in this collaboration.

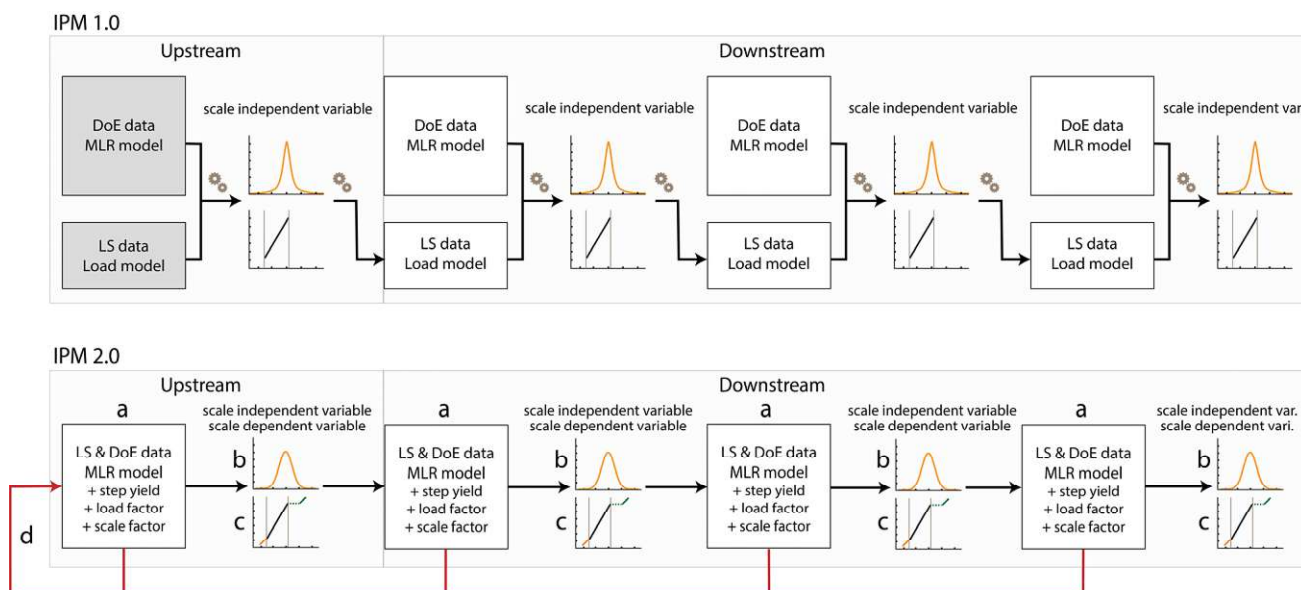


Figure 1. (top) Original IPM technological process flow (IPM 1.0). (bottom) Proposed collection of IPM innovations (IPM 2.0). IPM 2.0 differs from IPM 1.0 in the following improved areas: (a) robust and simplified data model, (b) addition of scale-dependent responses, (c) conservative extrapolation procedure for multiple linear regression (MLR) models, and (d) real-time feedback loop (depicted as a red line).

2. Materials and Methods

2.1. Software

The IPM was developed with commercially available software PAS-X Savvy 2022.01 (Körber Pharma Austria GmbH, Vienna, Austria). This software uses Python 3.79 as a base (Python Software Foundation, available online: <https://www.python.org/>, accessed on 20 January 2022). The procedures below were built onto the framework of the IPM 1.0.

2.2. Data

A case study was prepared with an industry partner to assess the proposed procedures as a proof of concept. A recombinant protein production process in a mammalian cell culture was provided that had been developed and characterized with a limited number of at-scale manufacturing runs. The model contains one primary upstream UO and seven downstream UOs, followed by final results at drug substance.

The downstream process consists of the following UOs: a chemostat bioreactor (UO1), followed by a filtration step (UO2), a concentration step (UO3), a virus inactivation step (UO4), a capture chromatography step (UO 5), filtration (UO6), and two polishing chromatography steps (UO7 and UO8).

The primary response for the case study is *Step Yield*, as it best leverages and displays the proposed innovations, further discussed in the Results section.

The available statistical models for each UO are summarized in Table 1 and are characterized in more detail in Table S1. PP is a model built upon process parameters,

but not including an input load value. A *Step Yield* model is only missing from UO4, as no statistically significant model was found. The raw data for *Step Yield* for each UO are described in Table S2.

Table 1. Model availability for *Step Yield*.

UO	<i>Step Yield</i>
UO1	Starting UO
UO2	PP
UO3	PP
UO4	No model found
UO5	PP
UO6	PP
UO7	PP
UO8	PP

2.3. IPM Data Model

This collaboration builds on the IPM 1.0 technology, adapting the general concept of combining a lab-scale model with manufacturing process data. The lab scale provides the bulk of the investigated design space, and the manufacturing data provides the primary UO linkage. IPM 1.0 proposes a two-matrix system based on scale-independent variables (specific clearances, SC, downstream).

The two required data matrices are the following: a standard $p \times n$ matrix (where p is the number of parameters and n is the number of runs) of small-scale DoE data that explore the investigated design space. Many DoEs, in our experience, are modeled in the individual UO and have no connection to the previous UO. The large-scale data matrix is a $1 \times n$ matrix with the only factor being the incoming specific load of a given CQA to be regressed against the output CQA. This depicts a de facto transfer function between UOs [2].

Equation (2) defines the specific load clearance model (SLC) of the j -th CQA as the i -th UO's pool values in percentage (%) divided with the i -th UO's load density (which itself is load divided by column volume CV).

$$SLC_j = \frac{\left(\frac{CQA_{j,i} \text{ load}}{CV} \right)}{CQA_{j,i} \text{ pool}} \quad (2)$$

The combination of the two models occurs only during the simulation phase and proceeds according to Equation (1).

3. Results

3.1. Data Model

A simpler and more robust data model can be established, given the availability of certain additional information about the scale and starting material. All scale data can be combined in a single matrix and subsequently fitted by a single model provided that two new columns are also added: *Scale* and CQA_{load} .

Scale is treated as a fixed categorical factor, thereby having the benefit of capturing any scale offsets within the model. In addition to providing this important scale comparison as a simple regression coefficient, the *Scale* | *Large* level can be selected as the prediction setting during the MC procedure, thus always simulating under manufacturing-scale conditions.

CQA_{load} refers to the pool value of the CQA from its precursor UO. That is, the starting material value for any given CQA is used to model its impact on the pool CQA (CQA_{pool}) in the current UO. This factor does not refer to *Load Concentration* (i.e., the desired molecule amount over volume) or *Load Density* (i.e., the desired molecule over resin volume/filter area) necessarily, but rather each CQA's own starting material. The upshot is the creation of an individual factor matrix X for each CQA, as seen sorted by color in Figure 2.

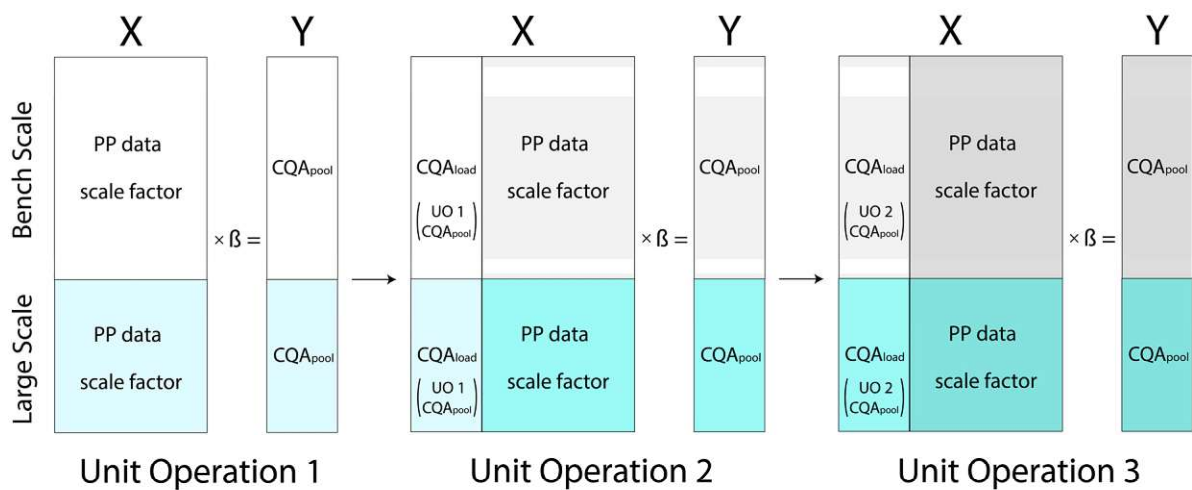


Figure 2. Proposed data model for the IPM. Small-scale DoE data (greyscale boxes, top) and large-scale manufacturing data (colored boxes, bottom) are in the same matrix with an added categorical scale factor. Each modeled CQA has a unique additional factor called CQA_{load} , which is the CQA_{pool} value from the previous UO.

The regression is now described in Equation (3): the predicted CQA (\hat{y}_i) is generated by model intercept β_0 plus all investigated factors (x_i) and their respective coefficients (β_i) plus the error term (ϵ). The two non-PP terms (CQA_{load} and $Scale$) provide the linkage of the model to both manufacturing scale and the subsequent UO. Architecturally, the process consists of statistical model objects representing the UO models. This permits a simple modular build-up of the full model and replacement upon refitting with new data.

$$\hat{y}_i = \beta_0 + \beta_{load}x_{load} + \beta_{scale}x_{scale} + \beta_1x_1 + \dots + \beta_nx_n + \epsilon \tag{3}$$

When performing MC simulations, CQA_{load} serves as the mathematical link between the precursor and current UOs. If the CQA_{load} is nonsignificant in the regression model, there is no mathematical link between the UOs for that CQA.

3.2. Extrapolation Procedure

In the above case, there is likely, nonetheless, a point at which the relationship is indeed quantifiable even if it is outside the investigated design space. Guarding against overlooking such a relationship requires extrapolation. As discussed, for data-driven models, extrapolation is discouraged in the absence of established first principles or process knowledge, since data alone is agnostic to behavior outside the observed data [22]. DoEs purposefully vary PPs outside typically observed manufacturing ranges. However, this space is limited by resources and knowledge. Additionally, not all PP can be specifically controlled, such as the CQA_{load} , which contains propagated variation from all previous UOs. It is generally assumed that CQA_{load} has a quantifiable influence on the CQA value in the following UO (CQA_{pool}), even if not detectable in the design space. Without a mechanism to account for this uncertainty, the DA can only predict within already observed data.

Naive extrapolation of a data-driven model is indeed associated with extreme statistical uncertainty [23], but extrapolation may be constrained by conservative process-based assumptions that allow for a reasonable worst-case assessment of the quantified relationships [24]. Specifically for bioprocesses, this constraint must be at least severe enough to satisfy risk management in bioprocess development. Therefore, a linear stepwise extrapolation strategy for the simulation of CQA_{load} values is proposed here. This strategy differs depending on whether the CQA is categorized as impurity or purity and whether the simulated value is below or above observed measured values as depicted in Figure 3.

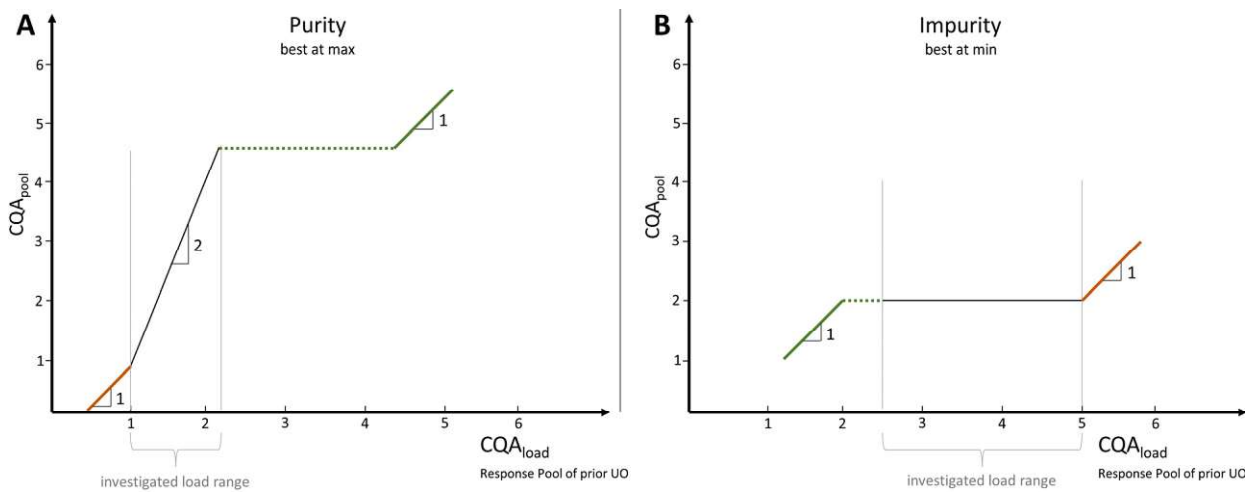


Figure 3. Visualization of CQA_{pool} value correction strategy for (A) purities and (B) impurities. (A) If the simulated CQA_{load} value is beyond the investigated load range (grey area), but below the maximal observed value of CQA_{pool} , then CQA_{load} is purified up to the maximal CQA_{pool} value at most, as depicted by the green dashed line. If, on the other hand, the simulated CQA_{load} value already exceeds the maximal observed CQA_{pool} value, no further purification takes place, and the CQA_{load} value equals the CQA_{pool} value, indicated by the green solid line. Conversely, if the simulated CQA_{load} value is below the investigated load range, then no purification takes place, and CQA_{load} corresponds to the CQA_{pool} value, visualized by the orange solid line. (B) The correction of the impurity CQA_{pool} values follows the same strategy as for the purities, only exactly reversed.

3.2.1. Purities (Best at Max)

Above the Observed Load Range

If the simulated CQA_{load} value ($lo\hat{a}d_i$) is above the observed load range ($\max(load)$) but below the observed maximal CQA_{pool} value (y_{max}), the resulting simulated CQA_{pool} value (\hat{y}_i) is corrected by the CQA_{load} value coefficient (β_{load}) multiplied by the offset between the maximal observed load range and the simulated CQA_{load} value, as shown in Equation (4), depicted as the green dashed line in Figure 3A. This is a conservative assumption that ensures that no further purification occurs when CQA_{load} values are purer than those in the pool. Thus, CQA_{pool} values are constrained to the maximal observed CQA_{pool} value.

$$\hat{y}_{i,corrected} = \hat{y}_i + \beta_{load} * (\max(load) - lo\hat{a}d_i) \tag{4}$$

If the simulated load value exceeds the observed CQA_{pool} value, the excess load is added to the corrected CQA_{pool} value ($\hat{y}_{i,corrected}$), as described in Equation (5). That is, the CQA_{load} value is simply passed through to the pool and no further clearance takes place, depicted as the green solid line in Figure 3A. It was assumed that the purity no longer decreased, and that a 1:1 propagation occurred.

$$\hat{y}_{i,corrected} = \hat{y}_{i,corrected} + |lo\hat{a}d_i| - |y_{max}| \tag{5}$$

Below the Observed Load Range

If the simulated CQA_{load} value is below the investigated load range, no purification takes place, as visualized by the orange solid line in Figure 3A. This conservative correction, as described in Equation (6), results in the CQA_{load} value not being purified, and the same concentration arriving in the pool.

$$\hat{y}_{i,corrected} = \hat{y}_i + (1 - \beta_{load}) * (lo\hat{a}d_i - \min(load)) \tag{6}$$

3.2.2. Impurities (Best at Min)

Above the Observed Load Range

For impurities, the conservative approach follows that no clearance occurs if the simulated CQA_{load} values are above the investigated load ranges, as described by Equation (7) and depicted by the orange solid line in Figure 3B.

$$\hat{y}_{i,corrected} = \hat{y}_i + (1 - \beta_{load}) * (\hat{load}_i - \max(load)) \quad (7)$$

Below the Observed Load Range

If the simulated CQA_{load} is below the investigated load range, but not below the observed minimal CQA_{pool} value, the simulated CQA_{pool} value is forced to the minimal observed CQA_{pool} value, as visualized as the green dashed line in Figure 3B.

$$\hat{y}_{i,corrected} = \hat{y}_i + \beta_{load} * (\min(load) - \hat{load}_i) \quad (8)$$

If the simulated CQA_{load} value is below the minimal observed CQA_{pool} value, the CQA_{load} value is passed to the pool without any clearance, as described and depicted as a green solid line in Figure 3B.

$$\hat{y}_{i,corrected} = \hat{y}_{i,corrected} + |\hat{load}_i| - |y_{max}| \quad (9)$$

3.3. Uncertainty Intervals

Where implemented, process models (including the IPM 1.0) tend to estimate uncertainty by sampling the confidence interval of the individual models. These intervals determine the uncertainty of the model mean, but are not optimized for predicting manufacturing data over many batches. Therefore, tolerance intervals were added as the default prediction setting for the IPM 2.0 data model on the basis of an established fixed-effect regression model implementation [25]. As such, both the confidence and the future coverage of the prediction are considered in the total variation, which, to the best of our knowledge, is not currently used in any equivalent integrated process model.

3.4. Scale-Dependent Variable Simulation Procedure

IPM 1.0 did not describe the modeling and simulation of responses other than specific clearances, which have scale-independent units that do not change over the UOs. To test the feasibility of an alternative pathway for nonspecific or scale-dependent variables, we propose modeling the product amount at the end of the upstream process (i.e., Harvest) and then adjusting via the individual UOs to simulate *Step Yields* without requiring the separate modeling of volumetric changes. This entails partially removing the response from the process model chain while still retaining the impact by process parameters. CQA_{load} is replaced by a variable that was only modified by the model output and is assayed through as much of the process as possible; in this case, *Global Yield*. The procedure is below and is generalizable to any variable that has a component (i.e., *Volume*) not described in the process models themselves.

As seen in Figure 4, the yield may be seen as a combination of *Step Yield* and *Global Yield*. The proposed procedure during the IPM MC simulations is as follows:

1. *Concentration* at harvest converted *Product Amount* to amount either by a known fixed volume or by sampling a distribution of feasible volumes.
2. *Product Amount* becomes the first downstream UO pool value.
3. *Step Yields* are fitted in the individual UO data, unconnected to the precursor UO, as per Equation (10).
4. *Step Yield* is multiplied by the current *Product Amount*, and a new *Product Amount* is calculated.

5. The new *Product Amount* remains outside the model loop and is adjusted by the subsequent UO *Step Yield* predictions.
6. In addition to modifying *Product Amount*, a new attribute is produced: *Global Yield*, which is the current UO's *Product Amount* divided by the original harvest *Product Amount* (Equation (11)).
7. The above process repeats until drug substance and a final *Global Yield* is produced, defined as the ratio of the final *Product Amount* to the original (max) *Product Amount*.

$$Step\ Yield_i = \frac{Amount_i}{Amount_{i-1}} \tag{10}$$

$$Global\ Yield_i = \frac{Amount_i}{Amount_0} \tag{11}$$

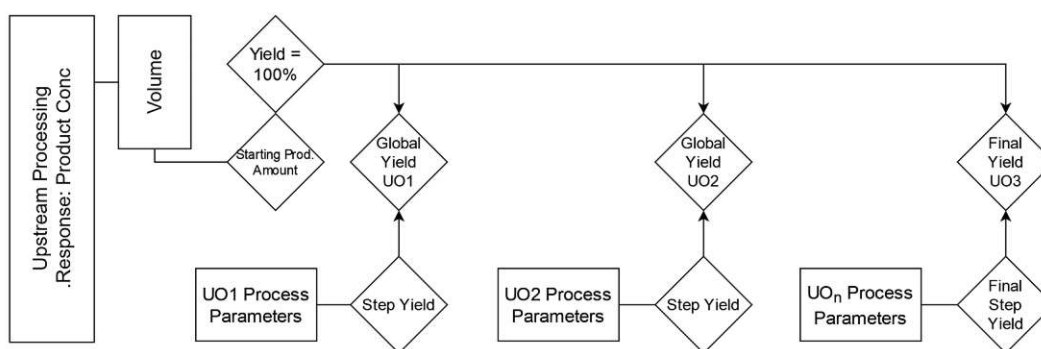


Figure 4. Step and Global Yield procedure as a model for scale-dependent variables. *Product Amount* is determined from upstream processing and considered to be 100%. All subsequent *Step Yields* may be removed from the linkage of UOs. *Step Yields* modify *Product Amount* by percentage recovery, which in turn modifies the *Global Yield*, which is updated after each UO towards a final *Global Yield* metric at DS.

Defining the acceptance criteria for *Global Yield* allows for the establishment of intermediate acceptance criteria for *Step Yields* via parameter sensitivity analysis. While this result does not produce a final *Product Concentration* per se, the final *Product Amount* may be modified by final volume adjustments as needed to arrive at a concentration.

3.5. Feasibility Case Study Results

A proof of concept was performed using the dataset shared by an industry partner described in the Methods section. This case study evaluates *Step Yield* to show the feasibility of the above-mentioned improvements. The *Step Yield* IPM was built successfully, containing all UO models. For each UO in the IPM chain, a *Step Yield_{load}* (i.e., the *Step Yield* from the precursor UO) design space was determined and divided into equidistant points called grids. The grid size covers a proposed range of likely *Step Yields* from the precursor UO, purposefully chosen to be outside the observed *Step Yield* ranges. The holistic process was then simulated at each grid per UO. With each simulation, the process was allowed to culminate at DS, and the final result was compared to a *Global Yield* OOS limit determined by a process expert. After repeating the simulation 200 times per grid size, a final %OOS value was obtained.

The results are shown in Figures 5 and 6. In Figure 5, the simulated *Step Yields* and their respective OOS results (%) are shown, which include extrapolated *Step Yields* (no OOS was observed in the data). For most UOs, there exists an incoming *Step Yield* at which the OOS rate starts to steeply rise, i.e., the *Global Yield* specification is no longer attainable. Process experts were then able to fix the *Step Yield* acceptance criteria to the point at which the OOS increase passes 5%.

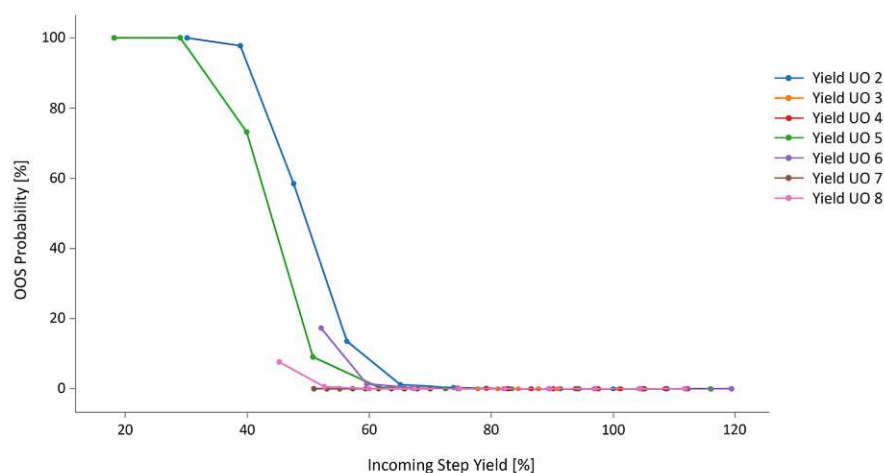


Figure 5. Parameter sensitivity analysis for *Step Yield* per UO. The y axis shows the proportion of OOS results based on the global yield drug substance specification, which is based on simulated incoming step yields per UO. Step yield results were extrapolated 10–20% outside the currently observed results to test the feasibility of the extrapolation procedure.

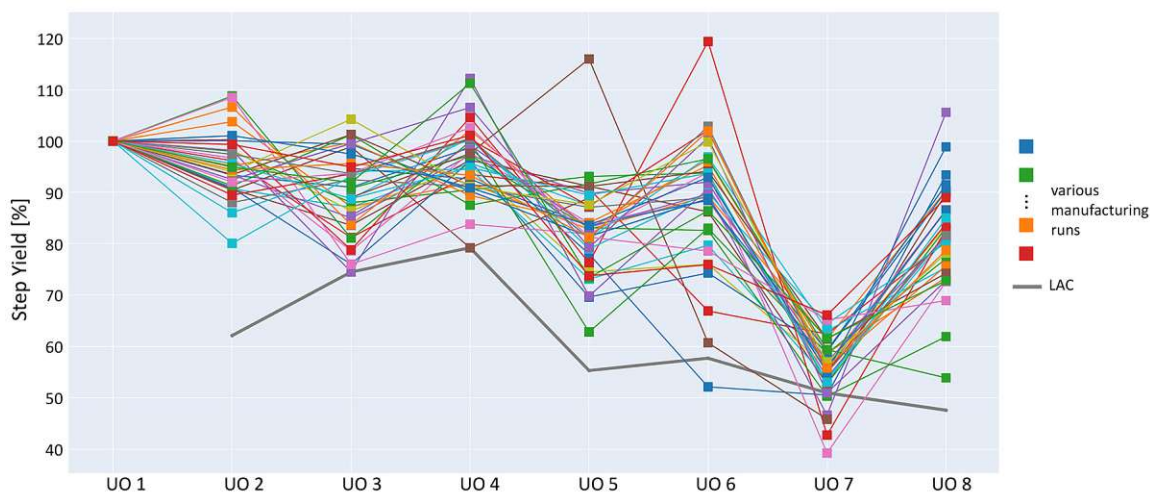


Figure 6. Parallel coordinate plot with results of parameter sensitivity analysis for the establishment of intermediate acceptance criteria for the *Step Yield*. Available manufacturing data are shown in various colors sampled from several campaigns/scales. The proposed intermediate acceptance criteria are marked in dark gray. Acceptance criteria were automatically generated across all intermediate UOs via the likelihood of meeting DS specifications predetermined by process experts.

In Figure 6, all observed data are plotted against the results of the above IPM-derived intermediate acceptance criteria. The acceptance criteria show a risk of increased OOS in the penultimate UO, which the process experts investigated and confirmed as a limitation of the current process. Subsequent actions were taken to adjust the process parameters to meet this new limit. The results were confirmed by process experts to be used in support of the final intermediate acceptance limit establishment.

There are instances of *Step Yields* with results >100%. Discussions were held with the subject-matter experts, and these artifacts stemmed from variation within the analytical method, i.e., variation in the load and pool values where both results were near 100%. The plot may also indicate high fluctuations of *Step Yields* between UOs. However, these results should be interpreted as independent of the precursor *Step Yields*. Since *Step Yields* always have a starting load amount of 100%, it is not unexpected to have large differences in mean yield in different UOs.

Thus, leveraging a *Global Yield* DS specification and all the improvements described above, plausible practical results were generated leading to adequate *Step Yield* intermediate acceptance criteria.

4. Discussion

4.1. Data Model

The simplification from the original two-matrix procedure into a single matrix aims to better meet the bioprocess development need of extracting data from differing scales. *Scale* is assayable directly in the same model where design space and UO linkage are fit. The manufacturing-based univariate model is replaced with a multivariate model, reducing the overall error term, since varying PPs are controlled.

The single matrix of course also reduces effort in compiling the data for a DA. Other than the addition of the individual CQA_{load} , the matrix requires only the necessary pre-processing for standard DoE-based regression analyses [26]. Moreover, it represents an improvement on classic linkage studies where multiple UOs must be modeled as one unit. Here, UOs may be modeled fully separately with no matrix overlap [27] while maintaining the CQA linkage. This simplicity also protects against data entry errors between the scales. Lastly, the data model provides normally distributed results since there is no longer a potential for product or Cauchy distributions due to the manipulation of the two models.

Newly arising higher-order terms may also be of interest, such as the interaction between *Scale* and PPs, which would give insight into the behavioral changes between scales rather than a simple offset. This could be used to significantly strengthen the conclusion of scale-down model qualifications, which are normally univariate. However, additional degrees of freedom are required for these terms, and, given the generally minor range of process parameter variables at a set point, the likelihood of an unfavorable correlation structure or even a singular matrix increases. Cost-benefit analysis should be undertaken before adding further terms.

There are further limitations to the current procedure that must be carefully considered. The two additional factors that were added to the matrix (i.e., *Scale* and CQA_{load}) are often not explored factors in original DoEs. Specifically, this information is often available, but was not included in the original design. The reassessment of appropriate design metrics (i.e., correlation, aliasing, power) is, therefore, required to ensure that the regression may still be performed. Less often, CQA_{load} is not tested at all. In this case, it is not possible for the data model to populate without additional context. Therefore, it is strongly advisable to include these factors a priori in statistically underpinned designs or minimally assess the data environment before beginning to fit models.

4.2. Extrapolation Procedure

The extrapolation procedure is a useful tool in bioprocess characterization since it allows for decision making within a risk management framework, even in the absence of data. The worst case defined in this procedure can allow for useful inferences about the edges of the system. Practically, it allows for conservative intermediate acceptance criteria and parameter limits to be provisionally established; these limits must otherwise be constrained within the current UO's observed data range.

Furthermore, this extrapolation procedure can be used as a stress test for subsequent UOs. Upon generating an extreme value, all subsequent UOs may process much more extreme input variables than those in their observed training data. Some of these UOs were physically designed to manage these unexpectedly high values and thus produce models that can easily purify excess material. Thus, one of two outcomes may be observed. Unexplored edges of the system show weaknesses in downstream steps. If worst-case results are easily managed in subsequent UOs, further experimental effort may be reduced as the risk of OOS is lessened.

The primary limitation is that the physical behavior of the process under extreme values is not known, and the system may react differently to the extrapolation assumptions.

While this procedure utilizes worst-case assumptions, thereby leaning on patient safety, these strict assumptions may nonetheless not hold upon fitting new data.

4.3. Scale-Dependent Variables

Simulating scale-dependent variables holistically over the process expands the application of the IPM to variables describing product quantity or process performance. The upshot is side stepping complexities arising in those variables having a volume component (or any component) that is controlled in a way that is not simple to define in a procedure or algorithm.

One operational disadvantage, in our experience, is that a lack of procedural strictness (such as in the case of a CQA dilution or concentration) is occasionally leveraged by operators towards increased manufacturing flexibility or buffer in achieving scale-independent results. In certain cases, this flexibility is preferred in operations; thus, the buy-in to this procedure may be dependent on the management's view of quality or yield outcome favorability.

4.4. Digital Environment and Real-Time Applications

Each complex step in bioprocess manufacturing potentially impacts the quality of the final product, yet state-of-the-art practices focus on the static outputs of individual UOs rather than on a holistic process model, particularly with regard to potential real-time applications [1,4,28–31]. Having so far discussed the innovative improvements to the IPM technology, it is now important to better define the framework for real-time DA deployment.

As previously discussed, by simplifying the data format, individual UO models can now easily be refitted by updating the single data matrix; thus, new predictions can be seamlessly conducted. With the physical process holistically depicted in silico and with a simple procedure to update the models, there needs only to be a framework for the feedback loop in real time.

Figure 7 shows a proposed graphical user interface for an IPM depicting the UOs in the upper half of the plot and the resulting predictions of the CQAs across the UOs in the lower half. A real-time workflow should proceed as follows:

The process begins at UO1 and ends at UO5, as shown in the upper half of Figure 7. At the start of the process, when no UO has been executed, the prediction of the resulting CQAs is based on sampling a most likely setting (i.e., normal distribution around the set point) of the PPs for each UO based on the variation of the large scale training dataset. This PP uncertainty maximally propagates through the prediction of the resulting CQAs. As the process progresses, however, and the actual PP settings are fed into the IPM (either manually or by automated import using API interfaces), these PP values become fixed points rather than distributions. Subsequent CQA predictions naturally become more accurate. By the last UO, the accuracy of the predictions should equal the accuracy of the individual UO model.

Figure 7 is, therefore, a snapshot of the process at a given time. The process is currently at UO3, and the uncertainty of the PP from UO1 to UO3 was set to 0, as the PP values are already known. These settings are immediately used to repredict the CQAs, creating a feedback loop and allowing for a reaction to the new conditions. If, for example, a PP is performed outside the normal operating ranges (shown as the orange bar at UO3 for PP4 in the plot), the effects of these PP settings are immediately shown in the lower half of the plot, where the new probability of the CQA conforming to drug substance specifications (depicted as red line) can be seen.



Figure 7. Proposed control panel for IPM use in a real-time environment. Process parameters may be controlled manually or through targeted APIs to either create a prediction around the process parameter (set point plus expected normal operating variation) or to bring in the discrete value when the PP setting is known. Predictions via MC simulations around the chance of specification conformity can be updated immediately. Refitting the models may also be performed in real time or at regular intervals.

This real-time prediction combined with the previously mentioned improvements allows for the probability of an OOS event to be calculated ahead of time and enables countermeasures to be taken as necessary. Furthermore, because predictions of scale-dependent process performance characteristics are now also included, the IPM can be used not only as a development tool for setting up an evaluating control strategy, but also as a manufacturing companion to optimize the process in terms of performance and quality.

5. Conclusions

The combined improvements of this IPM represent substantial progress in the development of a bioprocess DA. The original framework's conceptual advantages were kept while simplifying utilization, and expanding the scope, statistical rigor, applicability, and quality and business objectives.

As a real-time DA, the IPM allows for simulations during which PP settings can be quickly and seamlessly updated at the moment when new data are observed. Moreover, as further data become available, they may be immediately added via APIs from data sources to refit the model object. This provides the feedback loop both for observed parameter settings and model refitting, crucially enabling the IPM to function as a true DA within a DT concept.

Nonetheless, a substantial part of the improvements relies on the consistent testing of starting material CQAs, which is not universally performed. Thus, to gain benefits, more investment is needed in ensuring as comprehensive a testing plan as possible. While this does not need to be exhaustive, an adequate testing strategy should be built to provide sufficient CQA data at critical junctures to adequately profit from this procedure.

Further development should also be considered here. As this data model increasingly combines large- and small-scale data in the same data matrix, we see particular interest in the investigation of differences in scale behavior, offset, and variances where current

scale-down model qualifications are limited. The ease of comparing scales may motivate manufacturing managers to perform runs at the edge of normal operating ranges to gain insight into interaction effects with PPs while avoiding the risk of OOS results.

Moreover, the IPM technology could be used not only as a tool for control strategy development and deviation management, but also for planning experiments. For example, simulated spiking studies could be used to show which experiments would be needed to identify design space adaptations to decrease the OOS probability in a data-driven manner.

Ultimately, a holistic DA for a simple and robust bioprocess digital twin is eminently feasible and should continue to mature as an essential modeling tool in bioprocess development and manufacturing.

Supplementary Materials: The following are available online at: <https://www.mdpi.com/article/10.3390/bioengineering9100534/s1>. Table S1: overview of identified models based on DoE data; Table S2: step yield data sampled from different campaigns/scales.

Author Contributions: Conceptualization, C.T. and B.P.; methodology, B.P., C.T. and T.Z.; software, B.P. and T.Z.; formal analysis, B.P. and C.T.; investigation, B.P. and C.T.; data curation, C.T. and B.P.; writing—original draft preparation, B.P. and C.T.; writing—review and editing: C.T., B.P., T.Z. and C.H.; visualization, B.P. and C.T.; supervision, C.H. and T.Z.; project administration, C.T., C.T. and B.P. contributed equally to this project. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are not available due their proprietary nature with regards to the industry partner. Certain blinded data is available in the supplementary material.

Acknowledgments: This research was supported with a dataset for case study by our industry partners. We thank Martin Scholler, Thomas Posch, Michael Graninger, and Anne Tscheließnig for providing support, insights, and the data in supporting this effort. The authors acknowledge TU Wien Bibliothek for the financial support through its open-access funding program.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Smiatek, J.; Jung, A.; Bluhmki, E. Towards a Digital Bioprocess Replica: Computational Approaches in Biopharmaceutical Development and Manufacturing. *Trends Biotechnol.* **2020**, *38*, 1141–1153. [CrossRef] [PubMed]
- Zahel, T.; Hauer, S.; Mueller, E.M.; Murphy, P.; Abad, S.; Vasilieva, E.; Maurer, D.; Brocard, C.; Reinisch, D.; Sagmeister, P.; et al. Integrated Process Modeling—A Process Validation Life Cycle Companion. *Bioengineering* **2017**, *4*, 86. [CrossRef]
- Chen, Y.; Yang, O.; Sampat, C.; Bhalode, P.; Ramachandran, R.; Ierapetritou, M. Digital Twins in Pharmaceutical and Biopharmaceutical Manufacturing: A Literature Review. *Processes* **2020**, *33*, 1088. [CrossRef]
- Portela, R.M.C.; Varsakelis, C.; Richelle, A.; Giannelos, N.; Pence, J.; Dessoy, S.; von Stosch, M. When Is an In Silico Representation a Digital Twin? A Biopharmaceutical Industry Approach to the Digital Twin Concept. In *Digital Twins*; Herwig, C., Pörtner, R., Möller, J., Eds.; Advances in Biochemical Engineering/Biotechnology; Springer International Publishing: Cham, Switzerland, 2020; Volume 176, pp. 35–55. ISBN 978-3-030-71659-2.
- Piascik, R.; Vickers, J.; Lowry, D.; Scotti, S.; Stewart, J.; Calomino, A. *Technology Area 12: Materials, Structures, Mechanical Systems, and Manufacturing Road Map*; NASA Office of Chief Technologist: Washington, DC, USA, 2010.
- Bruynseels, K.; Santoni de Sio, F.; van den Hoven, J. Digital Twins in Health Care: Ethical Implications of an Emerging Engineering Paradigm. *Front. Genet.* **2018**, *9*, 31. [CrossRef] [PubMed]
- Geris, L.; Lambrechts, T.; Carlier, A.; Papantoniou, I. The Future Is Digital: In Silico Tissue Engineering. *Curr. Opin. Biomed. Eng.* **2018**, *6*, 92–98. [CrossRef]
- Tao, F.; Zhang, H.; Liu, A.; Nee, A.Y.C. Digital Twin in Industry: State-of-the-Art. *IEEE Trans. Ind. Inf.* **2019**, *15*, 2405–2415. [CrossRef]
- Grieves, M. Digital Twin: Manufacturing Excellence through Virtual Factory Replication. *White Pap.* **2014**, *1*, 1–7.
- Jiang, Y.; Yin, S.; Li, K.; Luo, H.; Kaynak, O. Industrial Applications of Digital Twins. *Phil. Trans. R. Soc. A* **2021**, *379*, 20200360. [CrossRef]

11. Taylor, C.; Marschall, L.; Kunzelmann, M.; Richter, M.; Rudolph, F.; Vajda, J.; Presser, B.; Zahel, T.; Studts, J.; Herwig, C. Integrated Process Model Applications Linking Bioprocess Development to Quality by Design Milestones. *Bioengineering* **2021**, *8*, 156. [[CrossRef](#)] [[PubMed](#)]
12. Borchert, D.; Zahel, T.; Thomassen, Y.E.; Herwig, C.; Suarez-Zuluaga, D.A. Quantitative CPP Evaluation from Risk Assessment Using Integrated Process Modeling. *Bioengineering* **2019**, *6*, 114. [[CrossRef](#)] [[PubMed](#)]
13. Hakemeyer, C.; McKnight, N.; John, R.S.; Meier, S.; Trexler-Schmidt, M.; Kelley, B.; Zettl, F.; Puskeiler, R.; Kleinjans, A.; Lim, F.; et al. Process Characterization and Design Space Definition. *Biologicals* **2016**, *44*, 306–318. [[CrossRef](#)] [[PubMed](#)]
14. Horvath, B.; Mun, M.; Laird, M.W. Characterization of a Monoclonal Antibody Cell Culture Production Process Using a Quality by Design Approach. *Mol. Biotechnol.* **2010**, *45*, 203–206. [[CrossRef](#)] [[PubMed](#)]
15. Agarabi, C.D.; Chavez, B.K.; Lute, S.C.; Read, E.K.; Rogstad, S.; Awotwe-Otoo, D.; Brown, M.R.; Boyne, M.T.; Brorson, K.A. Exploring the Linkage between Cell Culture Process Parameters and Downstream Processing Utilizing a Plackett-Burman Design for a Model Monoclonal Antibody. *Biotechnol. Progress* **2017**, *33*, 163–170. [[CrossRef](#)]
16. Zahel, T.; Marschall, L.; Abad, S.; Vasilieva, E.; Maurer, D.; Mueller, E.M.; Murphy, P.; Natschläger, T.; Brocard, C.; Reinisch, D.; et al. Workflow for Criticality Assessment Applied in Biopharmaceutical Process Validation Stage 1. *Bioengineering* **2017**, *4*, 85. [[CrossRef](#)]
17. Nadarajah, S.; Pogány, T.K. On the Distribution of the Product of Correlated Normal Random Variables. *Comptes Rendus Math.* **2016**, *354*, 201–204. [[CrossRef](#)]
18. Metta, N.; Ghijs, M.; Schäfer, E.; Kumar, A.; Cappuyns, P.; Assche, I.V.; Singh, R.; Ramachandran, R.; Beer, T.D.; Ierapetritou, M.; et al. Dynamic Flowsheet Model Development and Sensitivity Analysis of a Continuous Pharmaceutical Tablet Manufacturing Process Using the Wet Granulation Route. *Processes* **2019**, *7*, 234. [[CrossRef](#)]
19. Burdick, R.K.; LeBlond, D.J.; Pfahler, L.B.; Quiroz, J.; Sidor, L.; Vukovinsky, K.; Zhang, L. *Statistical Applications for Chemistry, Manufacturing and Controls (CMC) in the Pharmaceutical Industry*; Statistics for Biology and Health; Springer International Publishing: Cham, Switzerland, 2017; ISBN 978-3-319-50184-0.
20. Sokolov, M.; Morbidelli, M.; Butté, A.; Souquet, J.; Broly, H. Sequential Multivariate Cell Culture Modeling at Multiple Scales Supports Systematic Shaping of a Monoclonal Antibody Toward a Quality Target. *Biotechnol. J.* **2018**, *13*, 1700461. [[CrossRef](#)]
21. Montano Herrera, L.; Eilert, T.; Ho, I.-T.; Matysik, M.; Lausegger, M.; Guderlei, R.; Schrantz, B.; Jung, A.; Bluhmki, E.; Smiatek, J. Holistic Process Models: A Bayesian Predictive Ensemble Method for Single and Coupled Unit Operation Models. *Processes* **2022**, *10*, 662. [[CrossRef](#)]
22. Altman, D.G.; Bland, J.M. Statistics Notes: Generalisation and Extrapolation. *BMJ* **1998**, *317*, 409–410. [[CrossRef](#)]
23. Hahn, G.J. The Hazards of Extrapolation in Regression Analysis. *J. Qual. Technol.* **1977**, *9*, 159–165. [[CrossRef](#)]
24. Karpatne, A.; Atluri, G.; Faghmous, J.H.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; Kumar, V. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2318–2331. [[CrossRef](#)]
25. Young, D.S. Tolerance: An R Package for Estimating Tolerance Intervals. *J. Stat. Softw.* **2010**, *36*, 1–39. [[CrossRef](#)]
26. Montgomery, D.C. *Design and Analysis of Experiments*, 9th ed.; Wiley: Hoboken, NJ, USA, 2017.
27. Rathore, A.S.; Kateja, N.; Kumar, D. Process Integration and Control in Continuous Bioprocessing. *Curr. Opin. Chem. Eng.* **2018**, *22*, 18–25. [[CrossRef](#)]
28. Park, S.-Y.; Park, C.-H.; Choi, D.-H.; Hong, J.K.; Lee, D.-Y. Bioprocess Digital Twins of Mammalian Cell Culture for Advanced Biomanufacturing. *Curr. Opin. Chem. Eng.* **2021**, *33*, 100702. [[CrossRef](#)]
29. Narayanan, H.; Seidler, T.; Luna, M.F.; Sokolov, M.; Morbidelli, M.; Butté, A. Hybrid Models for the Simulation and Prediction of Chromatographic Processes for Protein Capture. *J. Chromatogr. A* **2021**, *1650*, 462248. [[CrossRef](#)]
30. Nargund, S.; Guenther, K.; Mauch, K. The Move toward Biopharma 4.0: Insilico Biotechnology Develops “Smart” Processes That Benefit Biomanufacturing through Digital Twins. *Genet. Eng. Biotechnol. News* **2019**, *39*, 53–55. [[CrossRef](#)]
31. Narayanan, H.; Luna, M.F.; Stosch, M.; Cruz Bournazou, M.N.; Polotti, G.; Morbidelli, M.; Butté, A.; Sokolov, M. Bioprocessing in the Digital Age: The Role of Process Models. *Biotechnol. J.* **2020**, *15*, 1900172. [[CrossRef](#)]

Part III

Conclusions

Conclusion

“ *We live in a world of frightful givens. It is given that you behave like this, given that you will care about that. No one thinks about the givens.*

— **Ian Malcolm**
Jurassic Park

“ *Assignable causes of variation may be found and eliminated*

— **Walter Shewhart**
-Postulate 3-

4.1 Summary

The innovations developed within this thesis broadly confirm that substantial progress towards originally defined QbD goals is realistic and, in fact, already integrable in the industry. Furthermore, there is generally adequate architecture in place in the industry that will enable rapid deployment of these tools in the form of an increasingly mature bioprocess digital twin technology.

To enable the above described innovations, wide spread availability of infrastructure within which the development scientist can establish modeling practices must be present. As mentioned in the background, most project partners already had access to standard statistical software such as SAS JMP or MODDE, which shows willingness to accept models as fundamental to development work. Significantly more important, however, is the wide availability and acceptance of statistical programming languages such as Python, in which the innovations described were developed. The acceptance of these tools allows a flexibility of innovation and ease of positioning the procedures within larger IT environments. That is, while a statistical software may produce models which are then described in a report, the model objects from a python class may be integrated into the back-end of an IT manufacturing solution, as described in 3.4.

Cultural acceptance, not only of the statistical models, but also of the level of abstraction required to interpret the Monte Carlo simulations running over the entire process model, led

to a buy-in of the results that may not have been present even a few years ago. Understanding that the models produced interpretable, inferential results per unit operation, which simply are transferred to the next model may have assisted in creating this buy-in, as in 3.2 and 3.3. Paradoxically, one could speculate that the proliferation of black-box machine learning algorithms, which very rarely allows insight into the inner working of the model, leads to an acceptance of the IPM, as long as the individual unit operation models are able to be understood inferentially. Furthermore, the proposed user interface for running the simulations as piloted in 3.4 along with reasonably user-friendly processing speeds, (whereby even computationally heavy Monte Carlo simulations were completed in a few seconds), lead to an ease of use for non-data scientists.

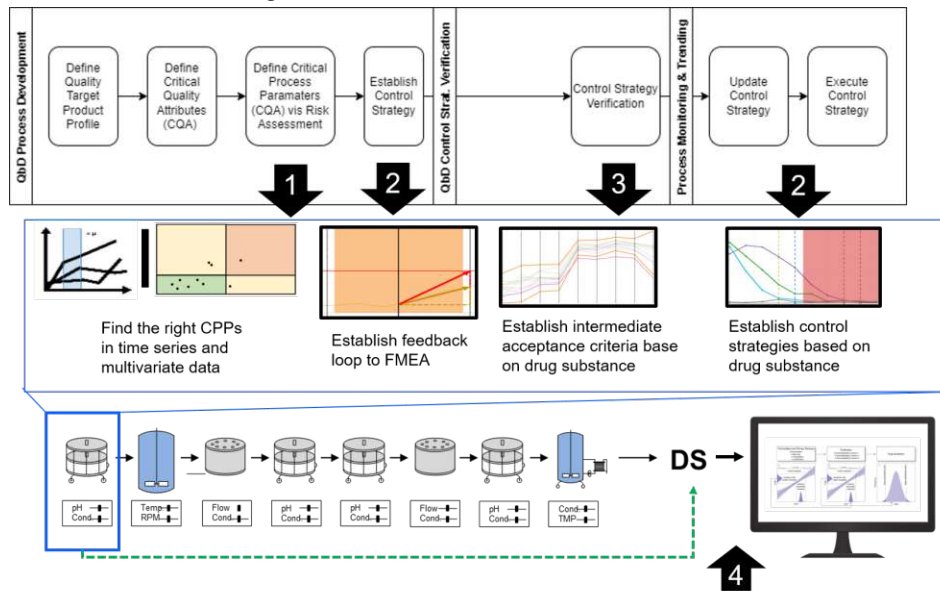
With these IT and cultural enablers in place, this thesis proves the following major advances:

The ability to successfully utilize algorithmic and procedural tools to increase both scope of variable type and data source type is eminently achievable. Where the current standard is to look for scalar data from direct or simple, derived measurements to serve as process inputs, it has now been shown that domain specific latent variables can also be easily and automatically extracted to be monitored as CPPs. These new variables can then be added directly to the risk management system. Furthermore, the importance of these variables can be detected through robust multivariate analyses like PCA, attached directly to the extraction procedure. Increased acceptance of latent, multivariate processes by our project partners as well as the ease of use of procedures such as ROBPCA greatly assists in the success of the innovation and leads to further systematic use in manufacturing.

The IPM can then build upon this more complete variable set and implement these into new applications that directly target QbD goals. Relationships to product specifications can now be mathematically defined. Using PSAs, this definition allows for a targeted simulation back-calculation to determine intermediate acceptance criteria. A similar PSA structure enables the determination of the PAR of the CPPs and additionally transforms this information into risk assessment severity rankings. Ultimately, the majority of QbD deliverables are improved via these procedures; risk rankings, intermediate acceptance criteria, and control strategies represent the bulk of the QbD workflow in Stage I process validation. The existing 2017 IPM framework was essential to the establishment of these applications, but also shined light on where the structure could be improved.

Finally, we show that we can centralize the above components into a repository that can be easily managed and updated. This places the totality of the IPM and its applications directly within an architecture that can be used in manufacturing as a digital asset of a digital twin. Ensuring consistent and real-time product quality results at drug substance, and then feeding these results back into the model, is at the heart of the potential of a digital twin. And with these incremental improvements, the IPM now may be seen as a mature digital asset in both development and manufacturing, to be further used for process optimization and risk mitigation, all in real-time.

Fig. 4.1.: Achievements of this dissertation: (1)uncovered latent variables within the process potentially having critical quality impact, (2) established applications of risk assessment rankings and control strategy deliverables within the integrated process model framework, and (3) developed a procedure to determine intermediate acceptance criteria within the same holistic framework. Finally, (4) improved the framework itself which connects the process holistically at each unit operation and bases all results on drug substance, while deploying this in a feedback loop in manufacturing



This thesis, therefore, shows that the industry can realistically take a large step toward the creation and leveraging of bioprocesses digital twins for the entire life cycle as defined in QbD.

Impact

The innovations in this thesis have two separate impacts on the different stages of the bioprocess lifecycle: the methods of development and the technology powering the methods.

The primary methods developed in the course of this thesis impact Stage 1 and Stage 3 of the FDA Validation Lifecycle. Bioprocess development (Stage 1) modeling is strengthened in its ability to produce QbD milestones based on drug substance via the IPM. That is, the current standard state-of-the-art statistically underpinned analysis centered on the unit operation can now be refocused as an integrated part of a centralized model. Such a mentality shift emphasizes that individual analyses will never be as useful as the analysis within the holistic process. Even the experimental designs may consider the integrated process model impact before any experimental run table is generated (see 6). Thus, from the earliest development stages, the scientist must consider the holistic model of the commercial process; exactly as QbD intended.

Commercial manufacturing (Stage 3), in turn, can now receive the completed IPM, (refit with any late-stage development and validation data) for use in process optimization and risk mitigation. All further experimental and manufacturing results should now be seen within the context of the IPM. This leads to another mentality shift: manufacturers can target and execute runs with specific manufacturing conditions that will statistically benefit the IPM to a maximum (see 6).

The technological impact is a shift in the way the multilinear regression matrices are considered and then built. That is, the IPM data structure has been simplified into a single data matrix, which now directly confronts scale issues and unit operation linkages with minimal additional complexity. Including the scale-linkage will allow scale-down model equivalence testing to potentially take place entirely within the existing factor matrix, with no need for additional runs. The addition of individual CQA load values helps contextualize holistic development decision-making, including where and when to test CQAs. For example, sampling plans should be better targeted to CQAs at their critical steps and precursors. This linkage impacts the general design strategy in development. If the modeling approaches are eminently utilizable, there is no reason not to design the characterization without the multi-step process in mind. This will ensure, even in the absence of a full IPM, a minimum of linkage to other unit operations in the process. Such a modeling structure will strengthen holistic results.

These innovations simplify the technical realization of the QbD milestones; how increasing use of simulations aligns mathematically and definitionally with QbD. That is, if the metrics required by regulatory authorities are able to be translated into mathematical definitions,

which are de facto described in the equations of the simulation applications, a virtual straight line can be made from the concept of QbD through the ICH and FDA guidelines to validation outputs.

These approaches then further establish the central use of Monte Carlo-based applications as a standard of integrated process simulation procedures. As bioprocesses are subject to high amounts of natural variation, it goes that a simulation procedure with focus on the propagation of error would be the primary tool used in risk management procedures. The successful use of several Monte Carlo methods within this dissertation should serve to reinforce this point.

It is our hope that both the methodological and technological impact of this thesis assists in forming a paradigm shift in how we reach QbD deliverables larger than the sum of the individual applications. Overall our design and technology in development now centers increasingly around what quality means to the patient impact.

Outlook

Much progress has been made in the applications of IPM technology; nonetheless, it is not yet the standard approach in bioprocess development. Significant effort is still required to make these approaches a mainstay in the industry. This thesis shines a light on the following principle areas that must still be investigated.

Discovering latent variables in multivariate datasets is becoming increasingly established. However, the design of the extraction algorithms and the dynamic process phases they rely on, are highly specific to the individual process. Improvements in platform processing, or minimally a generalized dynamic model of latent variable extraction, would help bring informative priors into new development projects. Dynamic phase robustness could be further improved upon by machine learning techniques such as random forest where detection of latency is more determined by prediction outcome rather than with inference.

Multivariate tools are increasingly common, but are still limited to use by advanced statistics users or pure statisticians. Ease of use of these methods is therefore critical. ROBPCA was used successfully in our contribution, but is largely only accessible to users with experience in the programming frameworks. Utilization of such tools may be straightforward, but still requires implementation in the architecture of choice with a simple user interface. The risk of course is that these tools go unused in the absence of statistical experts. Usability and error-proofing must be considered practically to gain maximum benefit from these advanced techniques.

Risk rankings were successfully evaluated using data-driven IPMs, but only for the severity ranking. QbD theory also suggests assessing frequency and detectability of failure modes. Further research could mathematically define these rankings and link them to the IPM. This brings up a further, larger point: there is an interesting gap that could also be filled between risk assessments and the outcomes of development data. In his work in 2021, Borchert described a rudimentary IPM that could establish models based purely on the risk rankings (i.e. severity and frequency) as defined by process experts. This is effectively a model built on the quantification of risk rankings alone [7]. Linking this work with the linear transformation of IPM data into risk rankings could provide an interesting feedback loop. In essence, one could start an IPM without any data (as is the case in most early development phases), which leads to improvements in the factor selection at a given unit operation in process characterization. Then, upon generating data, the rankings would be revised as in 3.2. A feedback loop would then iteratively replace each risk-based unit operation model in the IPM with the model based on a data model, once available, creating an entirely IPM based development. Such a feedback loop would allow a focus on drug substance specifications from the earliest stage of development through to validation.

Of course, the IPM is, by design, not limited to statistical models. Mechanistic models could certainly be added to the framework, similar to the flowsheet models described in 1.2.2. The framework allows for any model object that has inputs from the process as well as an output which can be fed into the subsequent unit operation model. In quality applications such as those described in 3.2 and 3.3, this is the load and pool quality attribute itself. Therefore, an interesting area of further development would be to describe this pool-load relationship mechanistically for quality attributes, of which we have seen relatively few. Of course, perhaps the most promising area is rather hybrid models, wherein a relationship is partially described by the data and partly by a mechanistic relationship. This could promisingly bring the best of both mechanistic and statistical worlds into the IPM.

A well-defined IPM would likely save experimental effort as well. We see that the IPM also has the potential to drive the selection of future experiments. The beginnings of research have started here, in what is called the integrated or holistic process design [35]. Such an algorithm would take available IPM data to optimize the next experiment or series of experiments. This could be seen as analogous to model-based DoE designs [28], but with a holistic perspective, always centering on drug substance specifications.

Concentrating on drug substance specifications: one final important area that developed through this work was the importance of understanding the true meaning of a drug substance specification. QbD theory well defines how this should be determined conceptually; specifications should be set to the point outside of which the CQAs have an impact on patient safety. In practice, however, there is very little literature in the field regarding the quantification of these limits vis-a-vis patient impact other than toxicology studies (which are not performed for every CQA). Moreover, even in the presence of toxicological studies, often significantly tighter limits are required by regulators since process control has a statistical meaning outside of pure patient pharmacology. An important field of research here would be to continue to establish specifications with true meaning for patient safety. This would create clearer and more interpretable meaning to the outcome of bioprocess development and further connect QbD with practical outcomes.

6.1 Conclusion

QbD deliverables, as defined by the ICH guidelines and FDA Validation guidelines, have been well established, but not used to the full extent of the definition of ensuring quality at the point of patient impact. As shown in this thesis, utilizing an improved integrated process model, QbD deliverables may be achieved to a more rigorous definitional standard and with increasing ease of implementation. Furthermore, once in place, these IPMs can support a digital twin and fulfill the potential, long sought, in the holistic control of biopharmaceuticals.

List of Figures

1.1	Simplified swim lane diagrams of (bottom) the FDA Validations 2011 Process Validation: General Principles and Practices document [36] and the (top) QbD methodology per ICQ Q8 [1]. Key milestones, such as the control strategy, discussed in this dissertation are shown in bold green boxes. Despite minor differences, the two are nearly identical approaches. Broadly summarized: first define the critical attributes of a process, then define the parameters that affect them and finally, build a control strategy for these parameters	4
1.2	An example of a univariate control strategy (PAR) plot. The model of the two factors (pH and conductivity) is shown against a response, including a tolerance interval (dotted line) and local intermediate acceptance criteria (red). Wherever the model plus tolerance interval intersects this threshold marks the end of the proven acceptable ranges. This definition is problematic in that it does not connect to downstream UOs and chooses acceptance criteria only relevant to the current UO and not the drug substance specifications	7
1.3	Levels of objectives within this dissertation:(A)Interpret QbD methodology maximally to patient interest, (B)establish innovations at key data-driven points in the QbD methodology (C) per unit operation, (D) based on results at drug substance, which are key to patient safety, and (E) produce a deployable digital twin to maintain all the above. The key innovations correspond to manuscripts 1-4, respectively in this dissertation)	13
4.1	Achievements of this dissertation: (1)uncovered latent variables within the process potentially having critical quality impact, (2) established applications of risk assessment rankings and control strategy deliverables within the integrated process model framework, and (3) developed a procedure to determine intermediate acceptance criteria within the same holistic framework. Finally, (4) improved the framework itself which connects the process holistically at each unit operation and bases all results on drug substance, while deploying this in a feedback loop in manufacturing	101

Part IV

Appendix

Appendix

A

The following index contains all supporting & supplementary information available for the research detailed within the manuscripts

A.1 A1 Supporting Information: Integrated Process Model Applications Linking Bioprocess Development to Quality by Design Milestones

Supplemental Data and Plots

Linking Bioprocess Development to Quality-by-Design Milestones via Digital Twin Applications

Model tables

Table S1: Overview of identified models based on DoE data. AIC was used as a primary identifier of selected model. R^2 , Q^2 , RMSE, and p -values, as well as residual analysis, were used alongside process expertise to determine acceptance of the model within the IPM. All residual variation is accounted for within the prediction interval of the simulated values within the IPM. (+) indicates positive coefficient and (-) indicates negative coefficient.

Unit Op	Blinded	Response	Model	R^2	Q^2	RMSE	P	Parameters
CAP	CQA1 _{prod}	Main SCX	Quadratic	0.45	0.34	1.119	0.019	(-) Load Density (+) Load Density ²
	CQA1 _{imp}	BPG	Linear	0.24	0.11	0.229	0.011	(-) Residence Time (+) Residence Time
	CQA2 _{imp}	HMW SEC	Quadratic	0.67	0.57	0.309	0.0001	(-) Load Density (+) Load Density ²
AT	CQA1 _{prod}	Main SCX	Interaction/Quadratic	0.85	0.55	0.016	0.0001	(-) Concentration (+) pH AT (-) Hold Time AT (+) Concentration * pH AT (-) Concentration * Hold AT (+) pH AT * Hold AT (-) pH AT ² (-) Concentration (+) pH AT (-) Hold Time AT (+) Neutr. pH
	CQA1 _{imp}	BPG	Interaction/Quadratic	0.94	0.92	0.04	0.0001	(+) Concentration * pH AT (-) Concentration * Neutr. pH (+) pH AT * Hold AT (-) pH AT ² (+) Hold AT ² (-) Concentration (+) pH AT (-) Hold Time AT
	CQA2 _{imp}	HMW SEC	Interaction/Quadratic	0.92	0.89	0.672	0.0001	(+) Neutr. Hold Time (-) pH AT * Neutr. Hold (+) Concentration ² (+) Hold AT ² (+) Neutr. Hold ² (-) Concentration (+) pH AT (-) Hold Time AT
	CQA3 _{imp}	LMW SEC	Interaction/Quadratic	0.43	0.12	0.031	0.0087	(+) Neutr. Hold Time (-) Neutr. pH (-) Concentration * pH AT (-) Hold AT ² (+) Neutr. pH ² (-) Concentration ²

Supplemental Data and Plots

Linking Bioprocess Development to Quality-by-Design Milestones via Digital Twin Applications

Unit Op	Blinded	Response	Model	R ²	Q ²	RMSE	P	Parameters
AEX	CQA1 _{prod}	Main SCX	Interaction	0.77	0.68	0.01	0.0001	(-) Load Density (-) Equil. pH (+) Load Density * Equil pH (-) Equil. pH ²
	CQA1 _{imp}	BPG	Quadratic	0.53	0.43	0.103	0.002	(+) Equil. pH (+) Equil. pH ² (+) Equil. pH
	CQA2 _{imp}	HMW SEC	Quadratic	0.65	0.17	0.067	0.0001	(-) Equil. Conductivity (+) Equil. pH ² (+) Equil. pH
	CQA3 _{imp}	LMW SEC	Quadratic	0.54	0.41	0.037	0.0005	(-) Equil. Conductivity (+) Equil. Conductivity ²
CEX	CQA1 _{prod}	Main SCX	Interaction / Quadratic	0.76	0.64	0.001	0.0001	(+) Elu. Conductivity (+) Load Density (+) Load Conductivity (+) Load / Elu. pH (-) Elu. Cond. * Load / Elu. pH (+) Load Cond. ² (-) Load / Elu. pH ² (-) Elu. Conductivity (-) Load Density
	CQA1 _{imp}	BPG	Interaction / Quadratic	0.77	0.65	0.068	0.0001	(-) Load / Elu. pH (+) Elu. Cond. * Load Density (+) Elu. Cond. * Load / Elu. pH (+) Elu Cond. ²
	CQA2 _{imp}	HMW SEC	Linear	0.33	0.22	0.598	0.0005	(-) Elu. Conductivity (-) Load / Elu. pH (-) Elu. Conductivity
	CQA3 _{imp}	LMW SEC	Quadratic	0.4	0.31	0.434	0.0007	(+) Load Conductivity (+) Load / Elu. pH (-) Load Cond. ²

Supplemental Data and Plots

Linking Bioprocess Development to Quality-by-Design Milestones via Digital Twin Applications

Table S2: Overview of models showing a correlation between specific CQA clearances and CQA load (load models). (+) indicates a positive regression line between clearance and load, whereas (-) indicates a negative regression line between clearance and load

Unit Operation	CQA	R ²	Effect
AT	CQA1 _{imp}	0.75	(+)
AEX	CQA2 _{imp}	0.76	(+)
	CQA3 _{imp}	0.70	(-)
DF	CQA3 _{imp}	0.76	(+)
CEX	CQA1 _{imp}	0.73	(-)
UFDF	CQA1 _{imp}	0.95	(-)
	CQA2 _{imp}	0.95	(+)

Supplemental Data and Plots

Linking Bioprocess Development to Quality-by-Design Milestones via Digital Twin Applications

Simulation Trend Plots

For all plots below, 1000 runs were simulated over the full process(top) and plotted against the sum of existing real manufacturing data (bottom) as well as the Out-of-Acceptance limit (red). Additionally, given only this information, a PPK result is given for OOA, that is, how likely given the normal distribution around the real and simulated data would be OOA at drug substance, which can be used as a comparative diagnostic of IPM quality.

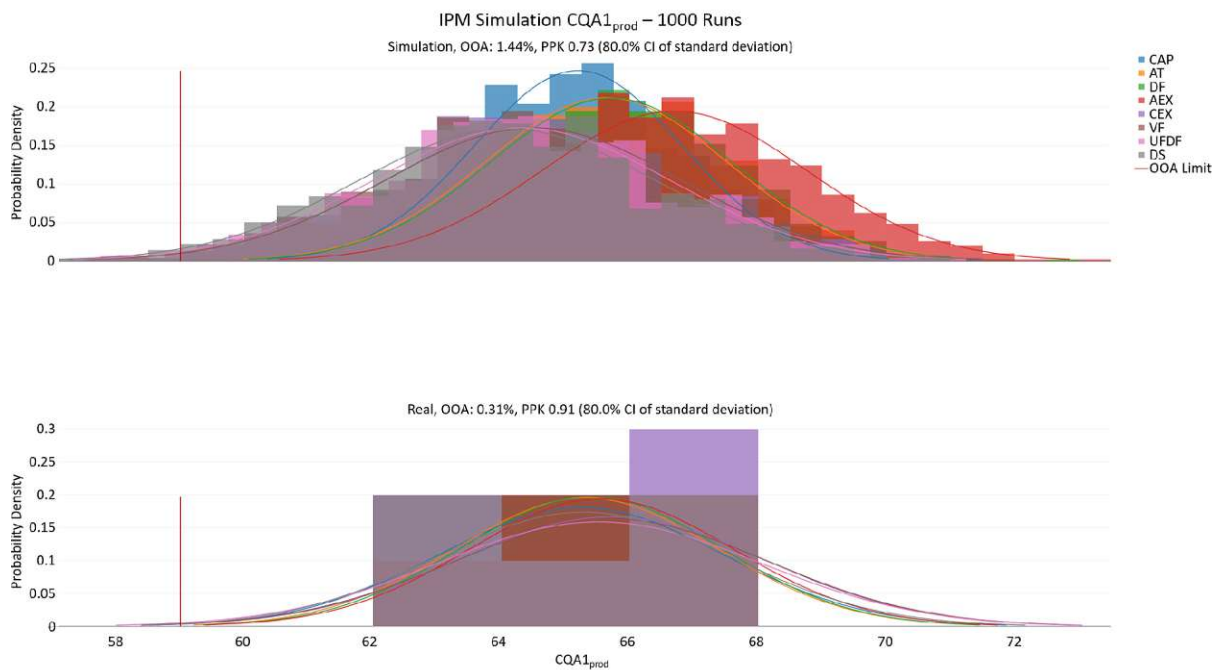


Figure S1: Simulation trend plot for CQA1_{prod}

Supplemental Data and Plots

Linking Bioprocess Development to Quality-by-Design Milestones via Digital Twin Applications

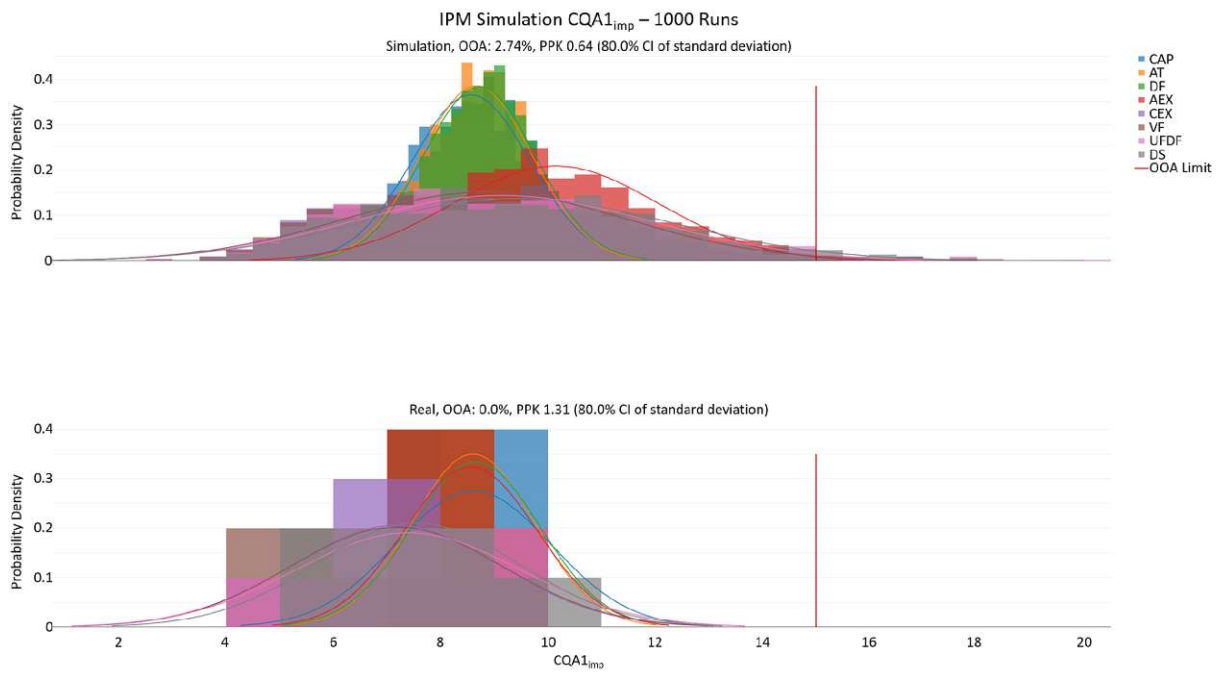


Figure S2: Simulation trend plot for CQA1_{imp}

Supplemental Data and Plots

Linking Bioprocess Development to Quality-by-Design Milestones via Digital Twin Applications

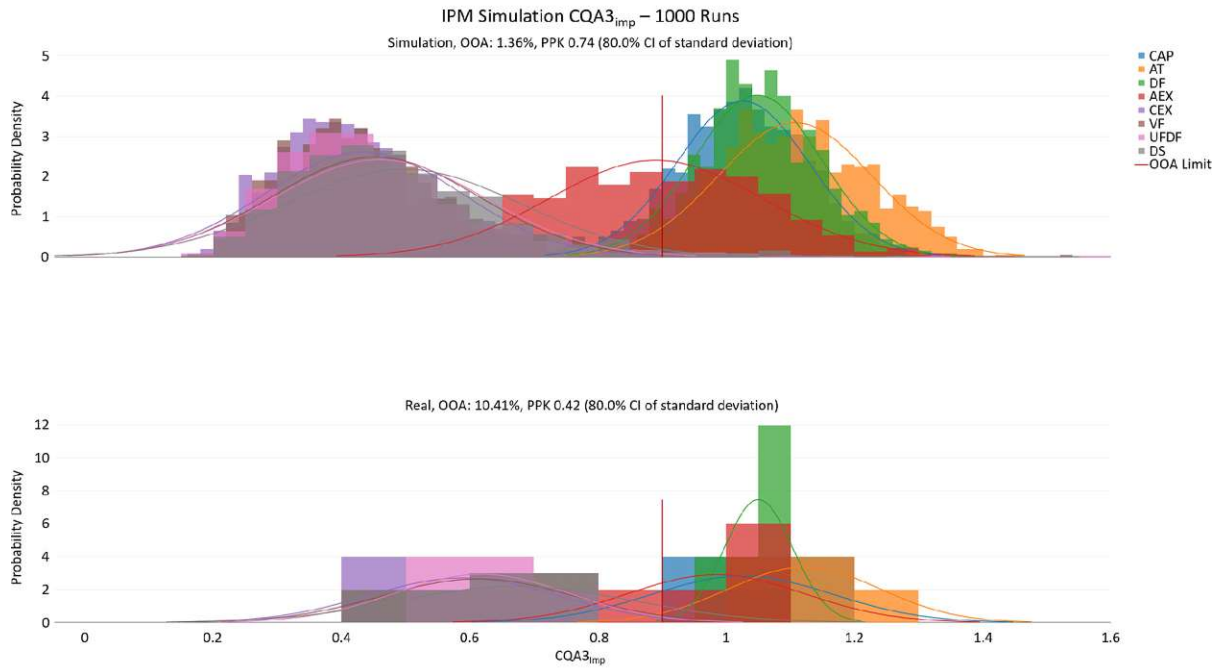


Figure S3: Simulation trend plot for CQA3_{imp}

Supplemental Data and Plots

Linking Bioprocess Development to Quality-by-Design Milestones via Digital Twin Applications

PSA CQA Plots

For all plots below, 1000 simulations per grid point were performed over the full process. The results in OOA are plotted across the relative screening range. The screening range is depicted such that 0 is always the set point condition and the remaining screening ranges are coded between -2, 2.

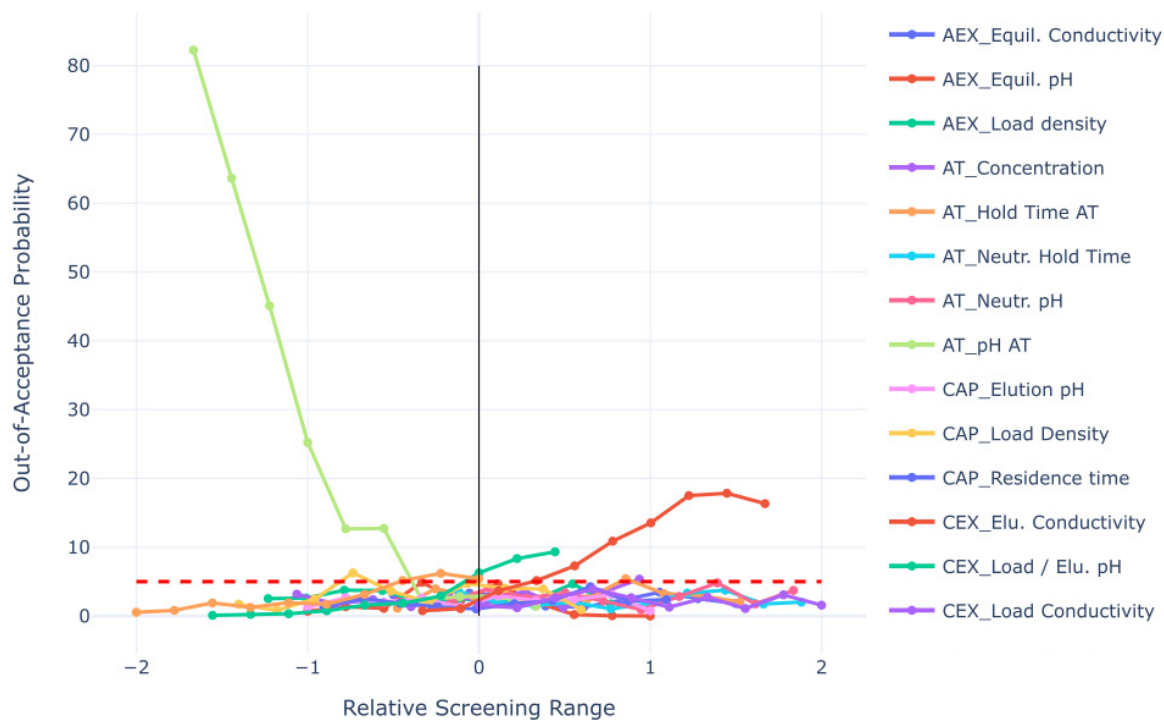


Figure S4: CQA1_{imp} CQA PSA plot

Supplemental Data and Plots

Linking Bioprocess Development to Quality-by-Design Milestones via Digital Twin Applications

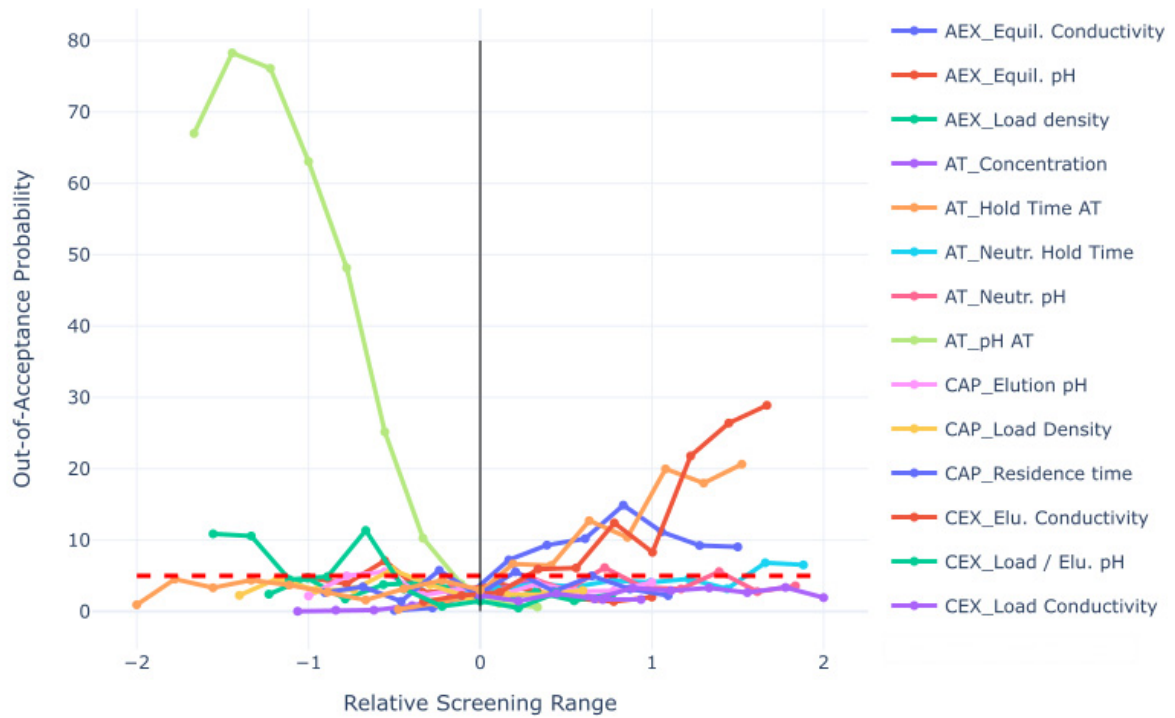


Figure S5: CQA2_{imp} CQA PSA plot

Supplemental Data and Plots

Linking Bioprocess Development to Quality-by-Design Milestones via Digital Twin Applications

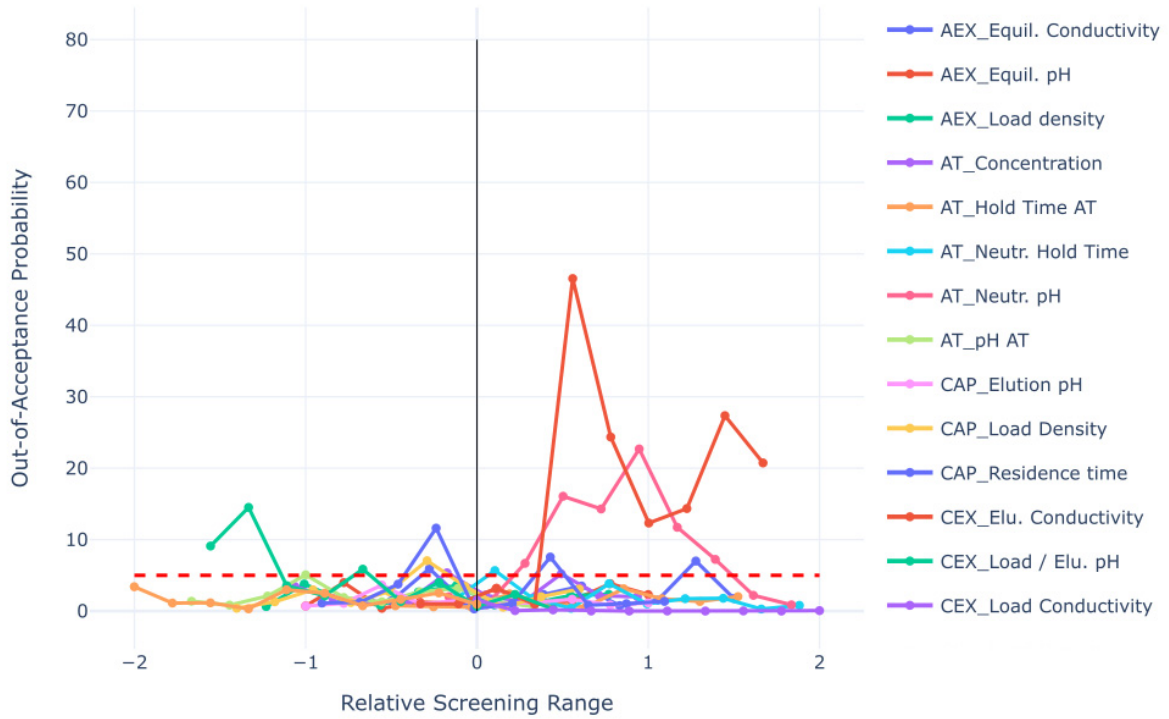


Figure S6: CQA3_{imp} CQA PSA plot

Supplemental Data and Plots

Linking Bioprocess Development to Quality-by-Design Milestones via Digital Twin Applications

PSA CPP Plots

For all plots below, 1000 simulations per grid point were performed over the full process. The results in OOA are plotted across the individual CPP range. The area is grey is excluded from the proposed manufacturing PAR range.

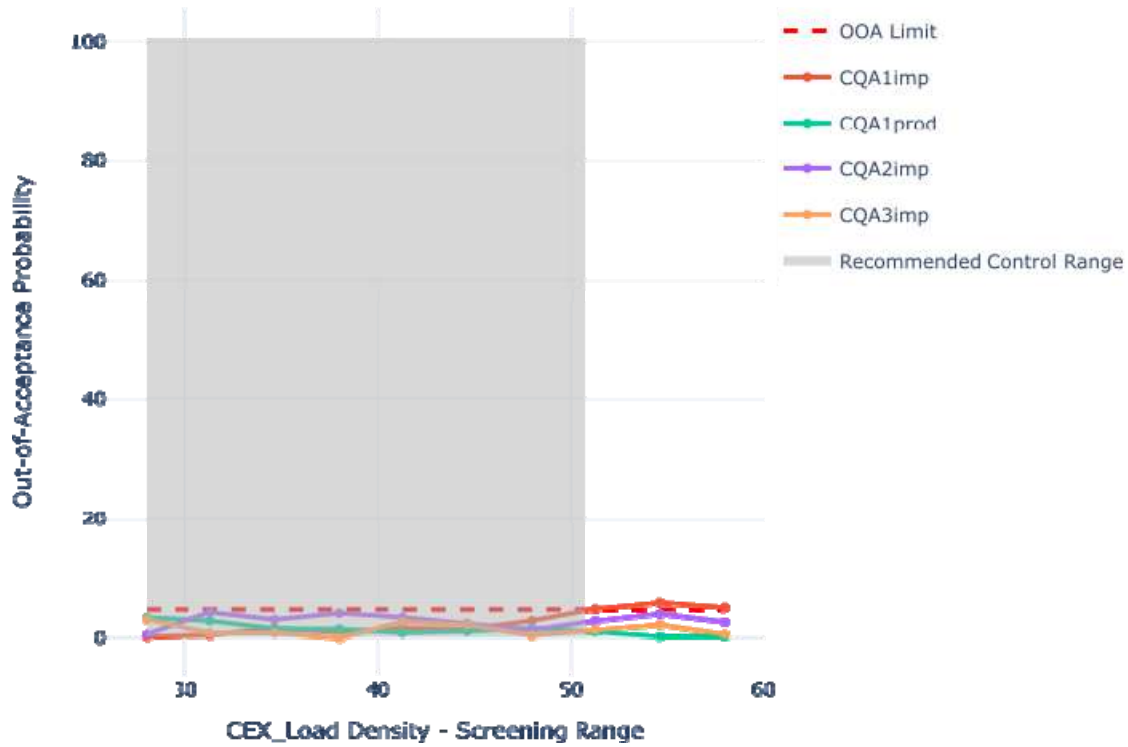


Figure S7: CEX_Load Density CPP PSA plot

Supplemental Data and Plots

Linking Bioprocess Development to Quality-by-Design Milestones via Digital Twin Applications

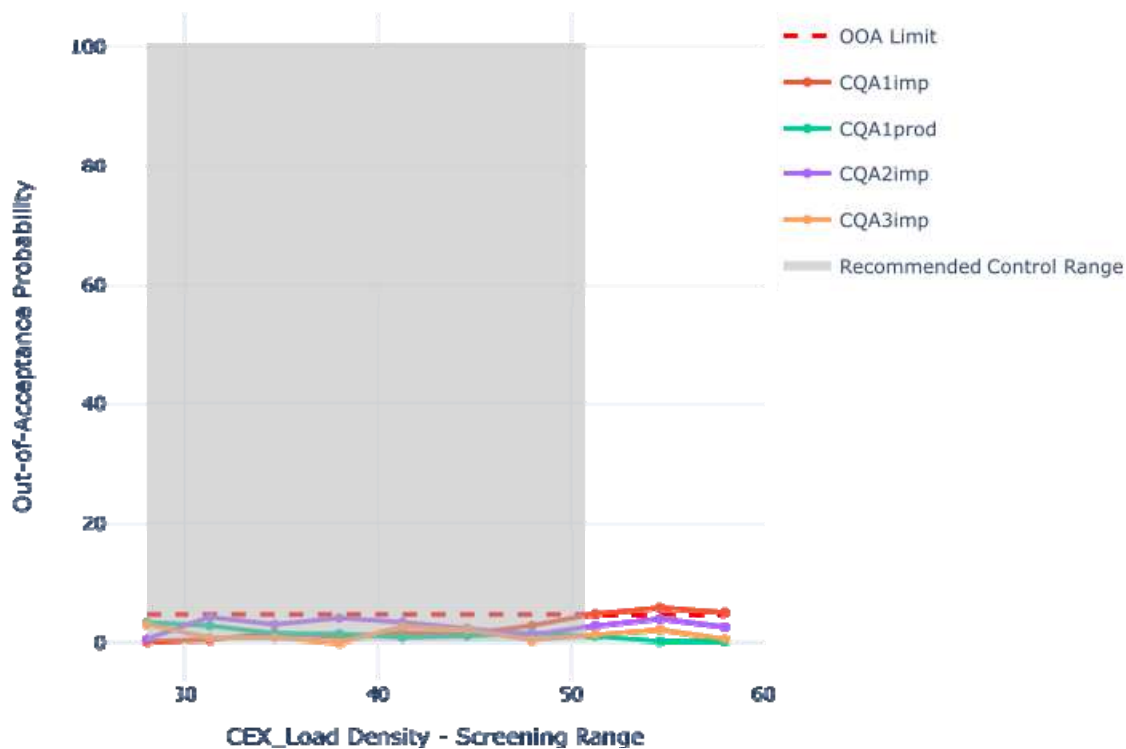


Figure S8: CEX_Load Density CPP PSA Plot

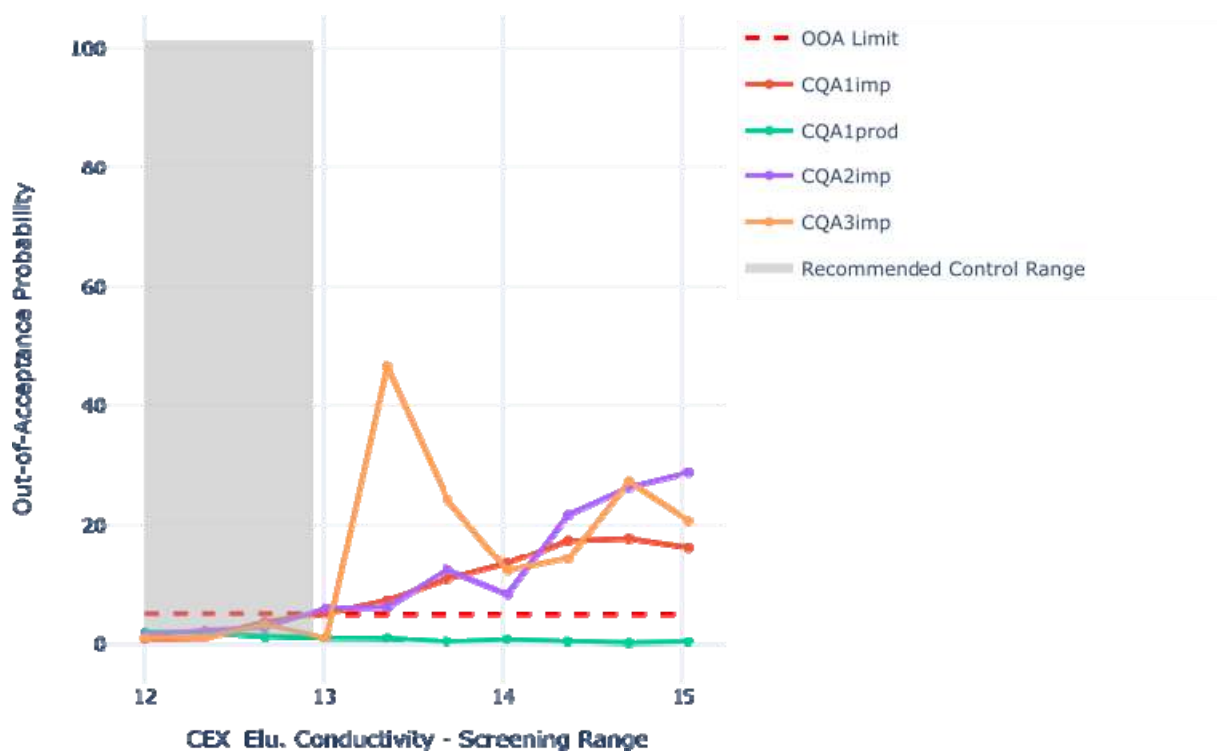


Figure S9: CEX_Elu. Conductivity CPP PSA plot

A.2 A2 Supporting Information: Specification-Driven Acceptance Criteria for validation of biopharmaceutical processes

Supplementary Material

Table 1: Overview of models describing the specific clearance as a function of process parameters. R^2 , Q^2 , and p-values, as well as residual analysis, were used alongside process expertise to determine acceptance of the model within the IPM.

Unit Operation	CQA	Adj. R2	Q2	p-value (F-Statistic)	No of Obs	Parameters
Capture	HCP ELISA	0.633	0.427	0.0674	5	Const (-1.30e-17), Load Pool Temperature (8.51e-01)
CEX	HCP ELISA	0.913	0.868	2.36E-05	11	Const (-2.637e-16), Elution Buffer pH (-8.397e-01), Elution Buffer Cond (-4.745e-01)
CEX	UP-SEC Aggregates	0.985	0.954	2.70E-07	11	Const (-1.665e-16), Elution Buffer pH (-7.616e-01), Elution Buffer Cond (-5.822e-01), Elution Buffer pH*Elution Buffer Cond (-2.521e-01)
CEX	UP-SEC Monomer	0.996	0.987	1.14E-06	11	Const (0.246), Elution Buffer pH (-0.644), Elution Buffer Cond (-0.571), Elution Buffer pH*Elution Buffer Cond (-0.444), Elution Buffer pH ² (-0.179), Elution Buffer Cond ² (-0.092)
HIC	UP-SEC Aggregates	0.453	0.321	0.00572	17	Const (-7.494e-16), Loading Pool pH (-5.001e-01), Loading Pool Temp (5.188e-01)

Table 2: Overview of models describing the specification clearance as function of the input material. R^2 , Q^2 , and p-values, as well as residual analysis, were used alongside process expertise to determine acceptance of the model within the IPM.

Unit Operation	CQA	Adj. R2	Q2	p-value (F-Statistic)	No of Obs	Parameters
Capture	HCP ELISA	0.424	0.427	0.0246	10	Const (4.921), HCP ELISA (0.000003)
Depth Filtration	HCP ELISA	0.892	0.869	7.78E-05	9	Const (-1.429), HCP ELISA (0.00438)
Depth Filtration	UP-SEC Aggregates	0.706	0.565	0.0028	9	Const (-0.973), UP-SEC Aggregate (0.712)
Depth Filtration	UP-SEC Monomer	0.487	0.352	0.0219	9	Const (-0.857), UP-SEC Monomer (-0.00874)
AEX	HCP ELISA	0.835	0.691	0.000136	10	Const (0.233), HCP ELISA (0.00614)
AEX	UP-SEC Aggregates	0.699	0.65	0.000119	14	Const (-0.368), UP-SEC Aggregate (0.268)
CEX	UP-SEC Monomer	0.845	0.812	2.10E-06	14	Const (0.358), UP-SEC Monomer (-0.00353)
HIC	UP-SEC Monomer	0.362	0.205	0.0387	10	Const (0.651), UP-SEC Monomer (-0.00655)
Bulk	UP-SEC Monomer	0.472	0.36	0.00397	14	Const (-1.033), UP-SEC Monomer (0.0104)

Table 3: Specific clearances were calculated by fitting a normal distribution to the available manufacturing data.

Unit Operation	CQA	Mean	Std	No of Obs
Virus Inactivation	HCP ELISA	0.131	0.636	9
Virus Inactivation	UP-SEC Aggregates	0.103	0.196	9
Virus Inactivation	UP-SEC Monomer	0.00198	0.0029	9
Viral Filtration	UP-SEC Aggregates	-0.00243	0.0932	14
Viral Filtration	UP-SEC Monomer	-0.000249	0.000769	14

UFDF	UP-SEC Monomer	-0.00103	0.000491	10
Bulk	UP-SEC Aggregates	-0.282	0.223	14

A.3 A3 Supporting Information: Architectural & Technological Improvements to Integrated Bioprocess Models towards Real-Time Applications

Supplemental Data and Plots

Integrated Bioprocess Model: Improvements, Case Study and Real Time Application

Supplementary Models & Data

Architectural & Technological Improvements to Integrated Bioprocess Models towards Real-Time Applications

Table S1: Overview of identified models based on DoE data. AIC was used as a primary identifier of selected model. R^2 , Q^2 , RMSE, and p -values, as well as residual analysis, were used alongside process expertise to determine acceptance of the model within the IPM. All residual variation is accounted for within the prediction interval of the simulated values within the IPM. (+) indicates positive coefficient and (-) indicates negative coefficient. (cat) indicates a categorical effect, for which there are multiple levels, with either (+) or (-) coefficients from the intercept

Response	Unit Op	Model	R^2	Q^2	RMSE	P	Parameters
Step Yield	UO1	Starting UO (100%)	-	-	-	-	-
	UO2	Linear	0.96	0.92	5.25	<0.0001	(cat) Campaign
	UO3	Linear	0.47	0.38	5.39	<0.0001	(cat) Campaign
	UO4	No model found	-	-	-	-	-
	UO5	Quadratic Interaction	0.72	0.59	8.44	<0.0001	(+) parameter1
							(-) parameter1 ²
							(+) parameter2
							(+) parameter2*parameter3
	UO6	Quadratic	0.26	0.15	10.98	0.0066	(cat) Campaign
							(+) parameter1 ²
	UO7	Quadratic	0.17	0.06	5.82	0.0193	(-) parameter1
							(+) parameter1 ²
	UO8	Quadratic	0.33	0.14	10.63	0.0007	(cat) Campaign
(-) parameter1							
(+) parameter2							
						(-) parameter2 ²	

Supplemental Data and Plots

Integrated Bioprocess Model: Improvements, Case Study and Real Time Application

Table S2: Step Yield Data sampled from different campaigns/scales. Missing data were discussed with process experts and confirmed before model fitting. DoE Data not included in order to compare data against expected manufacturing data.

Batch	Step Yield							
	UO1	UO2	UO3	UO4	UO5	UO6	UO7	UO8
C1 Batch1	100	90.7	87.1	95.7	87.5			
C1 Batch2	100	93.4	91.0	98.7	78.0	52.1	50.5	93.4
C1 Batch3	100	103.8	87.2	91.8	82.8	89.3	61.4	90.4
C1 Batch4	100	98.0	88.0	90.5	93.1	93.8	52.7	83.1
C1 Batch5	100	96.2	81.2	93.5	90.7	66.9	62.4	80.4
C1 Batch6	100	93.4	85.0	97.2	89.8	92.0	53.9	81.9
C1 Batch7	100	90.6	83.6	96.3	91.2	86.2	52.6	89.8
C1 Batch8	100	96.7	78.6	97.7	84.0	88.4	56.5	86.6
C1 Batch9	100	88.0	92.5	91.1	87.1	89.0	55.2	86.6
C1 Batch10	100	98.2	86.4	95.3	74.5	76.0	56.0	83.4
C1 Batch11	100	86.0	94.1	94.5	73.1	79.7	53.6	81.2
C1 Batch12	100	90.9	76.0	96.1	69.6	74.3	59.2	91.5
C2 Batch1	100	106.6	83.7	102.1	82.9	96.4	58.7	73.4
C2 Batch2	100	108.8	81.1	99.9	62.8	83.0	58.6	77.1
C2 Batch3	100	100.3	78.8	104.6	76.3	119.4	42.7	82.8
C2 Batch4	100	98.2	74.5	112.2	69.8	90.7	57.0	81.4
C2 Batch5	100	90.4	98.9	79.2	88.9			82.2
C2 Batch6	100	108.5	76.0	83.8	81.9	93.9	39.2	72.6
C3 Batch1	100		88.8	100.6	79.0	89.9	64.4	80.2
C3 Batch2	100		94.7	92.6	83.5	88.5	59.0	90.3
C3 Batch3	100		90.8	97.2	83.1	82.6	50.3	61.9
C3 Batch4	100		85.4	99.2	83.2	89.7	46.6	105.6
C3 Batch5	100		89.0	97.6	116.0	60.7	45.8	
C4 Batch1	100	80.1	93.2	100.2	81.0	96.9	63.1	75.4
C4 Batch2	100	100.1	99.4	90.3	81.4	88.8	56.0	86.5
C4 Batch3	100	95.1	99.7	89.4	84.1	95.3	55.9	75.7
C4 Batch4	100	91.3	101.2	87.5	91.6	96.5	59.3	53.9
C4 Batch5	100	89.5	93.6	100.4	87.3	102.1	55.0	83.9
C4 Batch6	100	92.5	99.5	106.5	79.3	100.7	50.9	72.8
C4 Batch7	100	93.5	101.3	91.3	91.1	94.2	61.4	74.3
C4 Batch8	100	92.0	93.7	102.6	81.3	78.6	65.2	68.9
C4 Batch9	100	97.4	93.7	100.5	81.0	102.9	54.9	81.6
C4 Batch10	100	94.0	104.3	91.2	87.6	99.8	56.9	78.2
C4 Batch11	100	95.6	88.8	94.9	89.5	93.7	53.1	85.0
C4 Batch12	100	101.1	97.5	90.9	83.6	92.9	54.8	98.9
C4 Batch13	100	94.4	95.6	93.3	81.2	101.9	55.8	78.8
C4 Batch14	100	94.9	91.9	111.3	73.4	86.4	61.6	72.8
C4 Batch15	100	99.4	94.9	101.1	73.8	75.9	66.1	89.0

Colophon

This thesis was typeset with \LaTeX 2 ϵ . It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at <http://cleanthesis.der-ric.de/>.

Declaration

I declare in lieu of oath, that I wrote this thesis and performed the associated research myself, using only literature cited in this volume. If text passages from sources are used literally, they are marked as such. I confirm that this work is original and has not been submitted elsewhere for any examination, nor is it currently under consideration for a thesis elsewhere.

Vienna, December 12, 2022

Christopher M. Taylor, MSc

