Biotechnology Advances 73 (2024) 108378

Contents lists available at ScienceDirect

Biotechnology Advances

journal homepage: www.elsevier.com/locate/biotechadv

Research review paper

A roadmap for model-based bioprocess development

Khadija Mu'azzam [a,b], Francisco Vitor Santos da Silva [a], Jason Murtagh [b], Maria Jose Sousa Gallagher [a,*]

[a] Process & Chemical Engineering, School of Engineering & Architecture, University College Cork, Ireland
[b] DPS Group Cork, Arcadis, Netherlands

ARTICLE INFO

Keywords:
Quality by design
Design space
Industry 4.0
Industry 5.0
Digital twins
Process modelling
Process simulation
Biopharmaceutical manufacturing

ABSTRACT

The bioprocessing industry is undergoing a significant transformation in its approach to quality assurance, shifting from the traditional Quality by Testing (QbT) to Quality by Design (QbD). QbD, a systematic approach to quality in process development, integrates quality into process design and control, guided by regulatory frameworks. This paradigm shift enables increased operational efficiencies, reduced market time, and ensures product consistency. The implementation of QbD is framed around key elements such as defining the Quality Target Product Profile (QTPPs), identifying Critical Quality Attributes (CQAs), developing Design Spaces (DS), establishing Control Strategies (CS), and maintaining continual improvement. The present critical analysis delves into the intricacies of each element, emphasizing their role in ensuring consistent product quality and regulatory compliance.

The integration of Industry 4.0 and 5.0 technologies, including Artificial Intelligence (AI), Machine Learning (ML), Internet of Things (IoT), and Digital Twins (DTs), is significantly transforming the bioprocessing industry. These innovations enable real-time data analysis, predictive modelling, and process optimization, which are crucial elements in QbD implementation. Among these, the concept of DTs is notable for its ability to facilitate bi-directional data communication and enable real-time adjustments and therefore optimize processes. DTs, however, face implementation challenges such as system integration, data security, and hardware-software compatibility. These challenges are being addressed through advancements in AI, Virtual Reality/ Augmented Reality (VR/AR), and improved communication technologies.

Central to the functioning of DTs is the development and application of various models of differing types – mechanistic, empirical, and hybrid. These models serve as the intellectual backbone of DTs, providing a framework for interpreting and predicting the behaviour of their physical counterparts. The choice and development of these models are vital for the accuracy and efficacy of DTs, enabling them to mirror and predict the real-time dynamics of bioprocessing systems. Complementing these models, advancements in data collection technologies, such as free-floating wireless sensors and spectroscopic sensors, enhance the monitoring and control capabilities of DTs, providing a more comprehensive and nuanced understanding of the bioprocessing environment.

This review offers a critical analysis of the prevailing trends in model-based bioprocessing development within the sector.

1. Introduction

Maintaining quality is vital within the biopharmaceutical industry to ensure the production of therapeutic products with the required safety and efficacy to treat patients. As such, regulatory bodies such as the FDA and EMA enforce stringent quality control measures, the violations of which potentially lead to significant penalties, product recalls and loss of reputation (Jagschies, 2018; A. Rathore and Winkle, 2009). As global competition for bioprocesses intensifies, companies face increased pressure to improve operational efficiency, reduce time to market and improve the overall quality of their products (Grangeia et al., 2020). Traditional methods of quality assurance, Quality by Testing (QbT), within the bioprocess industry were reactive; processes were designed and executed, followed by sampling end products against established

* Corresponding author.
E-mail address: m.desousagallagher@ucc.ie (M.J. Sousa Gallagher).

https://doi.org/10.1016/j.biotechadv.2024.108378
Received 19 December 2023; Received in revised form 10 May 2024; Accepted 12 May 2024
Available online 15 May 2024

quality standards. Samples that fail to meet the required specifications are discarded, the root cause analysis is carried out, and the process is repeated to achieve a product within specifications. This approach, while functional, is time-consuming and costly, resulting in lost batches and loss of materials/labour (Alt et al., 2016; Benyahia et al., 2020; Horst et al., 2021). Quality by design (QbD) has emerged as an effective strategy for improving performance within the bioprocess industry. QbD is a systematic approach to process development based upon predefined quality objectives and a high level of understanding of both the product and process. In contrast to QbT, QbD is a proactive approach, wherein quality is built into the process design and control to ensure the production of a consistently high-quality product (Alt et al., 2016; Benyahia et al., 2020; Horst et al., 2021).

The International Council for Harmonisation (ICH) has established a series of guidelines (ICH Q8 – Q12) for the implementation of QbD within the biopharmaceutical/pharmaceutical industry (Benyahia et al., 2020). QbD implementation (Fig. 1) involves the definition of a Quality Target Product Profile (QTPP), identification of Critical Quality Attributes (CQAs), definition of the Design Space, development of a Control Strategy and maintaining process capability and continual improvement (ICH, 2009). The QTPP is a predetermined profile outlining the desired quality characteristics of the final product, including its intended use, route of administration, dosage form, dosage strength, and safety and efficacy profiles. The QTPP serves as a roadmap during product development, offering a clear vision of its intended function, safety, efficacy requirements and development framework for outlining CQAs. A well-defined QTPP supports risk management by aiding in the identification of risks associated with its use. This approach results in optimised support control strategies that result in products that consistently meet their intended performance characteristics (Benyahia et al., 2020; Schlindwein and Gibson, 2018).

CQAs refer to those physical, chemical or biological properties that are required to be kept within an acceptable limit, range or distribution to ensure product quality. Unlike the QTPP, CQAs are typically easy to regularly assess during process development. Early identification of CQAs is critical to facilitate an understanding of which aspects of the product have an impact on safety and efficacy. Identification of CQAs typically relies on a risk evaluation process, which may consider prior product knowledge, experimental data, and literature. For example, CQAs for a monoclonal antibody would include parameters such as protein content, osmolality, pH and excipient content (Alt et al., 2016). Once identified, these CQAs serve to guide product and process development efforts. Manufacturers can ensure consistent product quality by understanding and controlling sources of variability affecting these attributes (Benyahia et al., 2020; Schlindwein and Gibson, 2018).

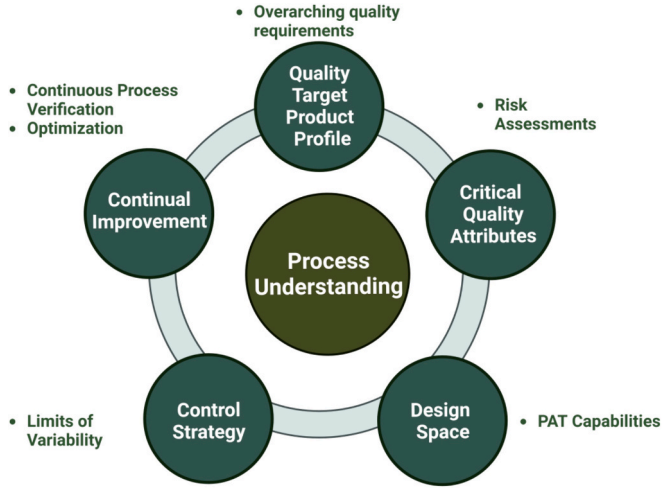

Fig. 1. Quality by Design implementation process.

Understanding the relationship between process parameters and CQAs enables design space mapping. Additionally, monitoring CQAs is advantageous during bioprocessing, real-time quality control and real-time quality management and can significantly reduce batch failure (Benyahia et al., 2020; Schlindwein and Gibson, 2018).

The Design Space (DS) refers to a multidimensional collection of process parameters and their interactions, which are required to ensure quality. The ICH defines the DS as *"the multidimensional combination and interaction of input variables (e.g. material attributes) and process parameters that have been proven to assure quality"* (ICH, 2009). Design spaces are developed through the analysis of the potential interactions between the variations of process parameters and material attributes on a given product's CQA. Identification of a DS involves applying statistical and risk management methods, such as the Design of Experiments (DoE), to investigate relationships among various variables and product quality. This approach allows manufacturers to identify optimal operating conditions and set boundaries within which product quality is not negatively impacted. This allows increased flexibility for process intensification and optimisation efforts, as any modifications within an approved DS are not considered a change to the approved process from a regulatory perspective (Benyahia et al., 2020; Schlindwein and Gibson, 2018). Correct identification of the design space is crucial ~~to~~ for understanding the relationship between CPPs and CQAs. However, despite its significance, several challenges exist in accurately defining the design space due to the complexities involved in the accurate modelling of biological processes. Additionally, the requirement for extensive experimentation, data collection, and analysis adds to the time and resource demands involved in DS identification. Addressing these challenges requires the implementation of a systematic framework and advanced analytical techniques. Contemporary literature (Geremia et al., 2024; Sachio et al., 2023) offers a method for identifying the DS. A systematic approach was proposed by the authors, involving several key steps; firstly, the necessity of a comprehensive understanding of the CPPs and CQAs, in line with QbD paradigm (Nayak et al., 2019). This knowledge is utilised to determine the applicability of statistical techniques, such as Design of Experiments (DoE), in order to systematically explore the effects of varying process parameters on product quality. (Geremia et al., 2024) A host of model-based tools can be employed to delineate the DS and observe the behaviour of selected unit operations across various process conditions.

Various approaches exist, including empirical modelling with manufacturing data, response surface methodologies, and first-principle models, each selected based on available literature, experience, and process knowledge/history (Metta et al., 2019). Predictive models offer the advantage of simulating processes under diverse conditions, often utilising Monte-Carlo simulations for scenario testing, while feasibility analysis supports the operational range of input factors (Kornecki and Strube, 2019; Metta et al., 2019; Geremia et al., 2024). However, challenges arise in incorporating process constraints into optimization problems, limiting applicability to processes with closed-form and differentiable constraints. In handling non-closed-form and/or non-differentiable constraints, such as those found in complex processes, traditional methods face computational challenges. However, a novel approach utilising α-shape surface reconstruction (Geremia et al., 2024; Sachio et al., 2023), which geometrically represents complex nonconvex problems, even with disjointed feasible regions has emerged in recent times to address this. Alternatively, surrogate-based feasibility analysis methods, relying on surrogate models to approximate feasibility functions, have shown promise in supporting DS identification, particularly for computationally expensive models and black-box constraints (Geremia et al., 2024).

This approach facilitates the identification of robust operating conditions and provides insights into process optimization. By employing various mathematical tools, the approach effectively incorporates the influence of manipulated variables on DS identification, crucial for real industrial environments where control strategies significantly affect DS

size. Additionally, it facilitates regulatory compliance by providing a structured framework for defining acceptable process ranges. However, there are limitations to consider, the process of design space determination can be time-consuming and resource-intensive, requiring extensive experimentation and data analysis. Furthermore, uncertainties inherent in biopharmaceutical manufacturing, such as variations in raw materials and equipment performance, may pose challenges in accurately defining the DS, however this can be addressed by implementing a stochastic approach to quantify the probability of satisfying all process constraints (Geremia et al., 2024; Sachio et al., 2023). Despite these limitations, the model-based approach to DS identification offers valuable insights and opportunities for streamlining the implementation of QbD within the biopharmaceutical space. Essentially, this framework utilises a suite of model-based tools to generate, categorize, and quantify sets of design and operating conditions capable of producing products and processes within specified criteria. It streamlines DS identification by connecting operating variables, design parameters, and system disturbances to CPPs in a unified approach.

A Control Strategy (CS) is defined as a planned set of controls that ensure process performance and product quality (ICH, 2009). The CS ensures that elements relating to product quality, such as process parameters, components, and materials, are maintained within the limits defined in the Design Space. Developing a CS typically begins with identifying the Critical Process Parameters (CPPs) and Critical Material Attributes (CMAs). CPPs are process parameters that significantly impact quality. At the same time, CMAs are the attributes of raw materials or intermediates that affect the product's CQAs, both of which are identified through risk assessment methods such as Failure Mode and Effects Analysis (FMEA) and experimental studies (Benyahia et al., 2020; Schlindwein and Gibson, 2018). Once identified, acceptable ranges for each parameter and attribute must be specified to ensure the process stays within its DS. Suitable controls must be selected once the CPPs, and CMAs have been identified. Process monitoring and control techniques, such as Process Analytical Technology (PAT) tools, and material controls, such as specifications for raw materials or intermediates, are used to achieve this (Benyahia et al., 2020; Schlindwein and Gibson, 2018).

A robust control strategy is highly dependent upon a deep level of process understanding as well as the accuracy of the process models developed during the creation of the DS. Inaccurate process models have been associated with controller instability. Seven frequently employed control strategies are: open-loop control, closed-loop control, cascade control, fuzzy logic-based control, artificial neural network-based control, model predictive control, and model-based control (Moser et al., 2021; A. S. Rathore et al., 2021a). This review will focus upon model predictive control due its growing popularity within the control strategy space (Chopda et al., 2022).

Model Predictive Controllers (MPCs) offer the potential to optimize complex bioprocesses with multiple influencing factors. However, implementing MPCs in bioprocesses is challenging due to the highly complex and nonlinear nature of biological systems, necessitating robust and accurate predictive process models (Chopda et al., 2022; Jelsch et al., 2021; Luo et al., 2021). Current MPCs in bioprocesses primarily employ unstructured models, simulating the process in advance and using current process conditions for subsequent optimization.

Additionally, to address the complexities and variabilities of mammalian bioprocesses, nonlinear MPCs with dynamic models has been effective (A. S. Rathore et al., 2021b), despite longer computational times, which are acceptable given the lengthy process duration. Significant improvements in performance have been seen with the combinations of nonlinear MPCs coupled with min-max optimization and Kalman filters to account for uncertainties and sensor unavailability (Chopda et al., 2022; A. S. Rathore et al., 2021b).

The optimization of MPCs involves adjusting parameters allowing for multivariable control and optimization (Jelsch et al., 2021; A. S. Rathore et al., 2021b). The final stage in the implementation of QbD is the commitment to continuous improvement (ICH, 2009). Continual improvement within the QbD paradigm is akin to traditional methods employed within the bioprocess industry, utilising data to identify opportunities for process improvement in a systematic manner using quality management tools. The capability of the process is assessed quantitatively using a statistical measure of the inherent process variability. This is generally measured in terms of a process capability index, which compares the allowable process spread (as defined by the limits outlined within the DS) with the actual process spread, which is defined by the process standard deviation (Benyahia et al., 2020; Bleisinger et al., 2022; Schlindwein and Gibson, 2018). A higher process capability index indicates a more capable process, a process that is better able to produce output within the required limits. The primary difference between the QbD and the traditional approach is that changes are not required to undergo regulatory approval provided they stay within the limits of the design space (Benyahia et al., 2020; Bleisinger et al., 2022; Schlindwein and Gibson, 2018).

Fundamentally, QbD may be described as a model-based approach to bioprocess development. The implementation relies heavily upon developing a well-defined design space coupled with mathematical and statistical models to predict the interactions between process parameters and product attributes.

2. From theory to reality: Industry 4.0, 5.0 & the design space

Remaining competitive and efficient has always required industries to embrace new technologies. The fourth industrial revolution, referred to as Industry 4.0 or Pharma 4.0 when adopted into the pharmaceutical industry, further automated bioprocessing through the melding of the physical and digital realms (Sharma et al., 2023). These digital and physical systems interact in real-time, allowing for increased efficiency, productivity, and flexibility in the industrial processes. As a result of Pharma 4.0, several advanced technologies such as machine learning (ML), big data analytics, the Internet of Things (IoT) cloud computing and digital twins are increasingly being utilised in the bioprocess industry, leading to significant advancements (Ding, 2018; Sharma et al., 2023). AI and ML are used to extract patterns from large datasets, facilitating predictive modelling and real-time decision-making. These technologies can aid in defining the design space in QbD by identifying the combinations of input variables that ensure optimal product quality. The IoT is a network of interconnected devices that can reduce manual labour and create a system of human-computer interaction. Comprising both hardware and software architecture, IoT plays a critical role in the bioprocess industry by streamlining operations and facilitating easier connectivity between people and processes. This results in increased efficiency, reduced costs, and enhanced security. IoT also supports pharmaceutical healthcare services, including diagnostic therapies and small-scale drug production. Together, AI and IoT are revolutionising the bioprocess industry by introducing new ways of operating and optimising processes (Sharma et al., 2023).

In contrast, the emerging concept of Industry 5.0 prioritises the collaboration of human expertise with advanced machinery. It emphasises the return of human intervention in bioprocessing, merging machine accuracy with human critical thinking. Industry 5.0 focuses on large-scale customisation, allowing consumers to select products adapted to their individual requirements. This new paradigm fosters a deeper level of human-machine collaboration, thereby increasing efficiency. While repetitive tasks are allocated to machines, critical thinking tasks are reserved for humans, promoting more skilled job opportunities. In essence, while Industry 4.0 emphasises connectivity through cyber-physical systems, Industry 5.0 augments this by fostering relationships with collaborative robots (cobots) (Maddikunta et al., 2022). A significant advantage of Industry 5.0 is its commitment to environmental sustainability, setting it apart from earlier industrial stages that lacked focus (Maddikunta et al., 2022). Industry 5.0 seeks to make more precise decisions by leveraging predictive analytics and operational

intelligence. The majority of its production process will be automated, using real-time data from machines in tandem with specialized experts (Maddikunta et al., 2022).

2.1. Defining the digital twin

A Digital Twin (Fig. 2) can be defined as a virtual replica of a physical system capable of bi-directional data communication. A functional DT should provide an accurate picture of the physical system, changing in real-time with any changes of its counterpart. A DT comprises three main integrated components: a physical system, a digital infrastructure and a virtual system (model) (Udugama et al., 2021).

The physical system is a tangible infrastructure upon which digital twins are built. This encompasses all aspects of the system, such as a bioreactor or the various instruments and sensors supporting it. A fully functioning DT requires the ability of sensors to accurately reflect the current state of the physical system (Udugama et al., 2021). Data acquisition from physical processes is crucial for developing a Digital Twin (DT). Critical process parameters (CPPs) can be sourced either manually from the human-machine interface (HMI) or via machine-machine interfaces (MMIs) like OPC, OPC DA, OPC UA, and Modbus, with OPC UA being the contemporary standard due to its enhanced features. Data transmission can employ various protocols, including MQTT and TCP/IP. Critical quality attributes (CQAs) are gauged using soft sensors, which merge hardware sensors with software models to offer insights into processes (Chen et al., 2020). These soft sensors have been applied in diverse sectors, ranging from freeze-drying to the distillation of crude oil. With vast data streams, a structured system like a data historian becomes essential for maintaining records. Recent studies have integrated local data historians for synchronised data collection and utilised communication protocols aligning with Industry 4.0 standards to foster smart factories and innovative solutions (Chen et al., 2020; Reyes et al., 2022).

The second component, the digitalisation infrastructure, links the virtual and physical aspects of the DT. This includes the automation in the physical system coupled with the virtual system's data communication links. Contemporary bioprocessing plants typically automate the production process using SCADA or DCS systems, with PI historians serving as the memory of the system. Modern DCS generations may support Digital Twins, though a dedicated server and information management system might be needed. Despite ongoing cybersecurity concerns, bioprocessing companies have increasingly turned to the cloud for data storage (Chen et al., 2020; Udugama et al., 2021). Various model types, including mechanistic, data-driven, and hybrid models, are available for use in DTs. Constructing a comprehensive DT that mirrors its physical counterpart whilst incorporating real-time analysis requires computational demand that local desktops and regular Central Processing Units (CPUs) are unable to satisfy. High-performance computing (HPC) becomes essential for running data-intensive models in real or near real-time. Numerous commercial modelling platforms, such as MATLAB and Simulink, COMSOL Multiphysics, gPROMS Formulated Products, aspenONE products, and STAR-CCM+, are prevalent in the process industries, providing a vast array of models and tools to replicate actual processes. Some of these platforms are also developing cloud-based solutions to facilitate DT construction, integration, and data management. In contrast, others prioritise compatibility with standard data management systems and Internet of Things (IoT) platforms (Chen et al., 2020).

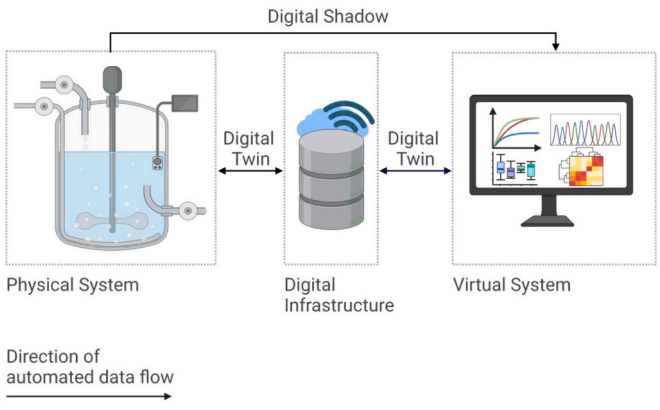

Fig. 2. Elements of a digital twin.

The third component of the DT is a high-fidelity, validated, mathematical model of the physical system that operates and mirrors it in real time. The model mirrors the physical system and can make predictive adjustments based on real-time data (both input and output) obtained via the digitalisation infrastructure (Udugama et al., 2021). Commercial IoT Platforms as a Service (PaaS) such as Predix, Mindsphere, and TIBCO Cloud, among others, have been developed to offer extensive tools for data management on cloud servers. Major cloud service providers such as Amazon Web Services, Microsoft Azure, and Google Cloud provide versatile platforms with tools for distributed computing, data analysis, and cybersecurity. However, software and data format disparities across manufacturers and services often obstruct data integration. Although cloud services offer integration options through application program interfaces, a universal file format is essential for enhancing cross-platform collaboration. The World Wide Web Consortium (W3C) introduced languages such as XML and RDF to model information. SQL databases, backed by the American National Standards Institute, are standard on cloud servers. Still, challenges in scalability paved the way for the emergence of NoSQL databases, optimised for scalability and hosted on cloud servers visualisation (Chen et al., 2020). Beyond storage, cloud servers provide significant computational power for data analysis and simulations and can host web services for real-time data visualisation (Chen et al., 2020).

Contemporary literature (Helgers et al., 2021; Juarez et al., 2021; Udugama et al., 2021) describes three levels of digital twin implementation as categorised by the control and automation community: the digital model, the digital shadow, and finally, a digital twin. Each of the three implementation levels requires an accurate, validated representation of the physical system (i.e., the digital model) - an accurate digital representation of an object with the differentiator based upon the flow direction of the communication links between the physical and virtual systems. A digital model can be described as an accurate virtual representation of a physical system with no real-time communication. Despite the lack of a communication line, data obtained from the physical system can be utilised to generate and validate digital models. A digital shadow extends the capability of the digital model by receiving real-time unidirectional data from the physical system. Lastly, a digital twin further extends the digital shadow by incorporating bidirectional communication with the physical system and additionally possesses the capability to control the physical system in real-time. In practice, these terms are used interchangeably, as many models that are classified as digital twins are, in fact, digital models or shadows (Helgers et al., 2021; Juarez et al., 2021; Udugama et al., 2021).

2.2. Building a digital model: Model types

In order to ensure the creation of a useful and effective model, various factors should be considered, namely, intended application, data availability and process understanding. A predictive model might be selected for forecasting future process performance, while descriptive models might offer a greater understanding of relationships within a process (Tsopanoglou and Jiménez del Val, 2021; Sansana et al., 2021). Data availability is also a key concern; depending on the model type, there are limitations based on the quality and quantity of the data. For example, empirical models (Fig. 3) are best suited to processes with a large amount of experimental or observation data, for which

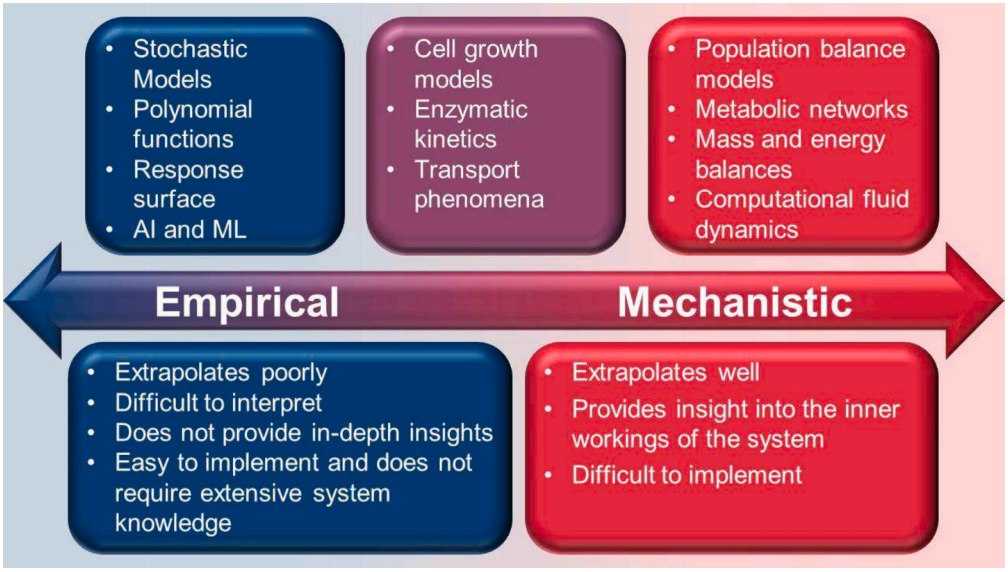

Fig. 3. Types of models.

mechanistic models consistently fail to predict outcomes; conversely, mechanistic models are derived from theoretical principles and are preferred when the data is lacking, coupled with a strong understanding of the mechanisms of the process, which quantitatively matches well reality (Tsopanoglou and Jiménez del Val, 2021; Sansana et al., 2021). In practice, every model is somewhere on the spectrum between purely empirical and mechanistic.

Empirical models are data-driven models created from observed relationships between process inputs and outputs. These models utilize statistical methods such as regression analysis for development. While empirical models do not require an in-depth understanding of the underlying mechanisms of a system, they still offer insight into process behaviour as well as provide insights into historical performance but tend to extrapolate poorly to untested conditions (Metta et al., 2019; Narayanan et al., 2023; Park et al., 2021).

Stochastic models are a class of empirical models that consider random events and variability within a system, thereby providing insight into and prediction of variability within bioprocesses such as gene expression, cell growth and protein production.

Mechanistic models are built upon fundamental principles and mechanisms underlying any process, detailing its physics, chemistry, and biology, which represent mass transfer, reaction kinetics, or other relevant phenomena that govern it (Moser et al., 2021). Mechanistic models provide an in-depth knowledge of the particular process, allowing for the optimisation of parameters or prediction of the process performance as well as setting thermodynamic, stoichiometric, and physical limits for process performance.

Population balance models are a type of mechanistic model used when processes involve the distribution of entities, such as cell populations in bioreactors that are bound by mass balances. They consider heterogeneity within a system as well as dynamics within its population - including growth, division and death - providing accurate predictions about cell behaviour as well as optimising process conditions (Alhuthali and Kontoravdi, 2022; Metta et al., 2019).

Mass balance models are a type of mechanistic model that describes the flow and distribution of materials throughout the process (Narayanan et al., 2022). They aid in the understanding the movement of raw materials, intermediates, and products, as well as the utilisation and conversion of materials at different process stages while setting upper production yields.

Kinetic models are usually derived from first principles and describe the rate at which reactions occur during the process. They provide an understanding of the reaction mechanisms, reaction rates, and the effect of process conditions on reaction kinetics. Kinetic models are significant for processes involving cell growth, protein synthesis, enzymatic reactions, and other biochemical reactions, which tend to have longer processing times than other chemical processes (Liu, 2020). Transport models are semi-mechanistic models that focus on the movement of fluids, heat, and mass transfer in the process. They aid in the optimisation of factors such as fluid dynamics, mixing, heat transfer, and mass transfer within bioreactors, purification columns, and other process equipment (Gupta et al., 2018).

Merging empirical with mechanistic strategies offers a myriad of advantages. In hybrid modelling, a mechanistic framework is used to describe the biological system (e.g., material balances). However, cell-specific rates are expressed as statistical expressions using input variables and parameters. Hybrid modelling traditionally falls into two main categories: serial and parallel architectures (Sansana et al., 2021; Narayanan et al., 2023). In the serial approach, elements of knowledge-based models are incorporated through the use of data-based models, signifying that the inputs for the former are derived from the latter. Conversely, in parallel architecture, knowledge-based and data-driven models operate simultaneously, with predictions from both being combined. Knowledge-based models use mathematical equations, often systems of differential-algebraic ones, to depict engineering expertise and process insights models (Narayanan et al., 2023; Sokolov et al., 2021; Sansana et al., 2021).

Data-based models, however, leverage statistics or machine learning to identify data patterns, bypassing the need for mechanistic understanding. Historically, many examples showcase the effective use of hybrid models in various bioprocesses, with a majority utilising serial architecture. This serial method is especially beneficial as it confines and separates the information that the data-based model interprets, resulting in sturdier models even when data is limited. In contrast, the parallel setup informs the data-based model without relying on existing knowledge, inheriting all the shortcomings inherent to solely data-driven models (Narayanan et al., 2023; Sokolov et al., 2021; Sansana et al., 2021). Hybrid modelling, when paired with the design of experiments, holds the promise of expediting process development. By deepening process understanding, this approach can decrease the number of necessary experiments, potentially reducing product/process development expenses. An alternative tactic involves using mechanistic models to produce training datasets for their statistical or ML surrogate models. Furthermore, hybrid models can be pivotal in devising experiments that

provide the most valuable information, further refining model efficacy (Narayanan et al., 2023; Sokolov et al., 2021).

2.3. Building a digital model: Data collection & processing

Data collection within the bioprocess industry has advanced rapidly in recent years, transitioning from observational data compiled by operators in front of the bioreactor to real-time data acquisition using wireless sensors, optical scanning, and image analysis (Gargalo et al., 2021). This progress is driven by the complexity and inherent variability of bioprocesses, as well as the gradual implementation of Process Analytical Technology (PAT) and QbD guidelines. Contemporary bioprocessing facilities are electronically supervised, generating process records that are well-characterized and less susceptible to human errors. Electronic batch records are used to document material input, process output, control actions, and physical parameters, ranging from the frozen cell vial to production scale bioreactors (Goos et al., 2021; Klein et al., 2020). Typically, there are three types of variables to measure and monitor: physical, chemical, and biological, with various sensing techniques available for each. Additionally, data can be sourced from established databases such as Laboratory Information Management Systems (LIMS) or bioTRAK (Gargalo et al., 2021; Goos et al., 2021; Klein et al., 2020).

Traditional measurement tools in bioprocessing are being complemented by novel techniques that use new instruments, data, and advanced computational tools. Free-floating wireless sensors, spectroscopic sensors, soft sensors, and chemometrics are among these developments, contributing to improved data collection and analysis (Reyes et al., 2022).

Soft sensors are sophisticated process monitoring systems that use algorithms to evaluate measurements online to generate information about a process state otherwise not measurable. Significant attention has been garnered by soft sensors as potential alternatives to conventional automated methods, particularly in the context of tracking critical variables in bioprocesses that cannot be directly measured (Brunner et al., 2021; Jiang et al., 2021). Occasionally, spectroscopic sensors are termed 'soft sensors' as they use software models to interpret spectroscopic data, providing outputs akin to traditional hardware sensors.

Soft sensors play a crucial role in sensor fusion by combining and estimating process variables based on available measurements. Sensor fusion, also known as data fusion (Kong et al., 2020; McAfee et al., 2022), is the process of integrating data from diverse sources to generate information with reduced uncertainty when compared to individual sources. Modern instrumentation systems use multiple sensors, each with its own function, and utilize specialized signal processing algorithms to merge their independent measurements into values thereby creating a multi-sensor system (Blasch et al., 2021; Kong et al., 2020; McAfee et al., 2022). Coordination among sensors is crucial in these systems, where data is fed to a data fusion layer for amalgamation using advanced signal processing techniques, while control applications provide necessary signals to actuators for effective system operation. Mathematical techniques as well as technical methodologies such as Kalman filters (discussed in this review), Bayesian networks, neural networks, and Gaussian processes are used for data fusion (Blasch et al., 2021; Kong et al., 2020; McAfee et al., 2022). Its advantages over individual source processing include enhanced detectability and reliability, expanded spatial-temporal perception and reduced inference ambiguity (Kong et al., 2020).

The three modelling methods for soft sensors (mechanistic, empirical and hybrid) each have distinct advantages and disadvantages relating to levels of accuracy, interpretability and ease of implementation (Fig. 4). Soft sensors based upon empirical modelling methods are characterized by their ease of implementation, as they do not necessitate in-depth knowledge of the process. They demonstrate a high degree of flexibility, capable of modelling complex, non-linear relationships. Additionally, they can adapt quickly to changes when trained on real-time

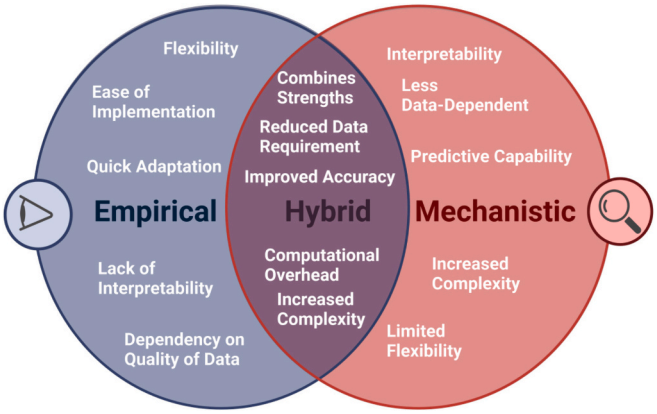

Fig. 4. Modelling approaches for soft sensors.

data. However, these models often lack interpretability, frequently described as "black boxes" due to the difficulty in understanding their decision-making processes. Another significant aspect is their dependency on the quality of data; poor or noisy data can result in inaccurate models. In contrast, soft sensors based upon mechanistic methods have enhanced interpretability as they are based upon underlying theories. However, this approach typically involved increased complexity, often necessitating sophisticated mathematical modelling. The model's predictive capability allows for the forecasting of future scenarios based on current inputs and states. Additionally, it is less data-dependent, not requiring extensive data for training. It is however noted that these sensors tend to exhibit limited flexibility, potentially struggling with unexpected, non-linear, or complex behaviours (Reyes et al., 2022).

The integration of both approaches results in soft sensors comprising the benefits of both empirical and mechanistic approaches, thereby combining their strengths. This amalgamation typically results in improved accuracy, surpassing that of either approach used independently. Additionally, the inclusion of a knowledge-based component often reduces the overall data requirement. However, this integration tends to increase the complexity of the system, making it more challenging to develop and maintain. Furthermore, hybrid soft sensors may impose a higher computational overhead, necessitating more computational resources (Brunner et al., 2021).

Kalman filtering (KF) is a mathematical technique used to estimate the state of a dynamic system from a series of noisy measurements (Alsadik, 2019; Brunner et al., 2021). In practical application, Kalman filtering comprises two primary stages: prediction and update. Initially, in the prediction phase, the system's current state is forecasted by leveraging the previous state and a dynamic model of the system. This forecast is accompanied by an assessment of the margin of error associated with the prediction. In the update stage the anticipated state undergoes adjustment based on a fresh measurement of the system's state. Through the Kalman filter, a weighted mean of the anticipated and measured states is computed, where the weighting factors are determined by their respective uncertainties. This iterative process yields an enhanced estimation of the system's state, effectively integrating both the system's dynamics and the presence of noisy measurements (Khodarahmi and Maihami, 2023; Brunner et al., 2021). KF can be integrated into soft sensors that employ both mechanistic and empirical modelling methods to estimate process variables. In the case of mechanistic modelling, it can be used to improve the accuracy of the model predictions by incorporating noisy measurements into the state estimation process.

Mechanistic models describe the underlying physical principles governing the behaviour of a system, but they often rely on simplifications and assumptions that may not capture all aspects of the system dynamics. By adjusting the model predictions based on real-time measurements and compensating for modelling errors and uncertainties,

Kalman Filtering mitigates the limitations of mechanistic models. This integration enhances the reliability of the soft sensor by providing more accurate estimates of the process variables (Brunner et al., 2021; Reyes et al., 2022). Similarly, in empirical modelling approaches, Kalman filtering can be employed to refine the estimates obtained from data-driven models by incorporating dynamic information and accounting for measurement noise. Empirical models are constructed based on observed relationships between input and output variables without explicit knowledge of the underlying physics. However, they may suffer from limitations such as overfitting and sensitivity to noise in the data. KF mitigates these issues by dynamically updating the model predictions with measurements, resulting in more robust and adaptive soft sensors (Reyes et al., 2022).

Irrespective of whether the soft sensor relies on mechanistic or empirical modelling methods, Kalman filtering serves as a valuable tool for improving the accuracy and reliability of the estimates by integrating information from both models and measurements. This integration enables soft sensors to effectively capture the dynamics of the underlying processes and provide timely and accurate predictions for control, monitoring, and optimization purposes (Alsadik, 2019) thereby excelling in scenarios where systems undergo constant change. KFs primary benefit lies in their minimal memory requirements, as they solely store information regarding past states. Their rapid processing capabilities render them well-suited for addressing real-time concerns and integration into embedded systems (Khodarahmi and Maihami, 2023).

The expanding capabilities of computers, combined with advancements in AI and machine learning for signal processing, make soft sensors an attractive option for overseeing and controlling industrial production. Their potential benefits align well with the Process Analytical Technology (PAT) initiative and the broader trend of smart manufacturing. Further information on the application of soft sensors in bioprocessing can be found in the following referenced studies (Gargalo et al., 2020; Reyes et al., 2022; Jiang et al., 2021).

Free-floating wireless sensors are a novel type of non-invasive instrumented particles that can provide process data from within the harsh environment of an agitated bioreactor (Helgers et al., 2021; Juarez et al., 2021; Udugama et al., 2021). Similarly, spectroscopic sensors, which are based on the interaction of electromagnetic waves and molecular bonds, are gaining traction in bioprocess monitoring for in-line measurement of variables in the liquid phase. They collect process-relevant information on bioprocess variables in the spectral range from UV to mid-infrared (MIR) (Gerzon et al., 2022; Rösner et al., 2022). UV spectroscopy can be used to measure toxic intermediates, inhibitors, and cell debris, while MIR spectroscopy can monitor glucose, lactate, pH, etc. Coupling the spectra with chemometrics allows these sensors to detect several compounds instantaneously (Gargalo et al., 2021; Rösner et al., 2022). The new sensing methods in bioprocessing, such as free-floating wireless sensors and spectroscopic measurements, have limitations due to their technological immaturity, need for expert knowledge, and significant investment requirements. They also rely heavily on the collection and management of "big data," which presents its own set of challenges. Big data, characterized by its volume, variety, velocity, and value, requires extensive storage systems and analytical resources (Gargalo et al., 2021). The volume of data refers to the significant storage capacity required; the variety reflects the heterogeneity of data; velocity refers to the speed of data production and the need for high-throughput processing, whilst value refers to the necessity of efficient algorithms to extract useful information from the data. While many algorithms already exist (such as Support Vector Machine (SVM) (Rösner et al., 2022) or Artificial Neural Network (ANN)), their adaptation and integration into bioprocesses are still under development (Gargalo et al., 2021; Rösner et al., 2022).

Bioprocess data is typically multi-source, multi-scale, noisy, and heterogeneous, reflecting various modes of operation and equipment sizes (Gargalo et al., 2021). This diversity, along with differences in the frequency and type of measurements, can significantly impact the understanding of the data, resulting in the construction of an inaccurate model. The inherent variabilities of bioprocesses can result in interactions between process parameters, resulting in poorly understood effects on process responses. Most process parameters are dynamic, with time-related variations that complicate analysis and result in a complex data structure; parameters such as temperature or pH are typically recorded every second, while other parameters, such as end product concentration, are measured once or twice per day.

Digital Twins can be effective tools for identifying and dealing with faulty measurements. Sensors often present issues with stability, resulting in signal drift and noise. Drift, a long-term signal variation, is typically caused by gradual sensor degradation or slow changes in the system composition. Noise often results from interference factors such as temperature, pressure, or humidity. While distinguishing between drift and noise is challenging, as neither is deterministic, statistical techniques such as signal-to-noise ratio and SVM have been proposed within the literature to mitigate sensor drift and noise (Gargalo et al., 2021; Rösner et al., 2022). In order to deal with uncertainty, two approaches, the frequentist and Bayesian approaches, respectively, have been analysed within the literature (Gargalo et al., 2021). The frequentist approach utilises sampling distributions analysis, such as the Monte Carlo technique; however, due to the challenge associated with obtaining these distributions, asymptotic approximation is typically used to assess input uncertainty for large datasets. Contrarily, Bayesian methods, such as Bayesian modelling and flexible Bayesian, create distributions that quantify input uncertainty without requiring a large dataset or approximation for validation (Brunner et al., 2021). Although high in cost computationally, Bayesian methods are particularly useful at depicting variability from limited data as they can build the needed posterior predictive probabilities to test process reliability (Gargalo et al., 2021). It is noted that managing the uncertainties as a statistical process complicates model validation. It requires that a comparison of data with model results also estimates the likelihood of consistency between values (Gargalo et al., 2021; Brunner et al., 2021). In this manner, evaluating uncertainties improves the understanding of the reliability of model results. This is particularly vital for models which rely on correlated parameters instead of independent parameters. By utilising DTs, parameters and their interactions, which may not be directly measured, may be accurately estimated and subsequently optimised.

Due to the level of complexity and computational requirements, data processing represents a significant bottleneck in the development of DTs. Preparing data involves first extracting useful hidden information. Multivariate data analysis (MVDA), which involves the analysis of several variables and their possible interactions, is typically employed to identify useful information regarding the parameters. Other data-driven processing approaches include data pre-processing, dimensionality reduction and data mining. Data pre-processing consists of several sub-steps, such as data cleaning, normalisation and standardisation, transformation, de-noising, and missing value imputation. Noise is eliminated, valuable information is extracted using methods like discrete Fourier transformation or symbolic aggregate proximity (SAX), and time profiles are aligned (Gargalo et al., 2021). The second step involves the reduction of data dimensionality by decreasing the number of variables within the dataset. This simplifies later steps and reduces storage requirements. Principal Component Analysis (PCA) is often used for this purpose (Gargalo et al., 2021; Brunner et al., 2021; Reyes et al., 2022). Data mining utilises computational techniques to identify irregularities, trends, and correlations within the dataset to predict outcomes.

2.4. *Upstream modelling & defining the design space*

Cell culture modelling aims to understand and characterise the complex dynamics involved with cellular processes, thereby optimising the conditions within the reactor and improving productivity and quality while reducing bioprocessing costs. Key processing parameters

such as cell growth rates, densities, and reactions to environmental changes are affected by the type of expression system selected for culturing. Table 1 provides high-level summary of these differences.

Upstream processing can be divided into four operational modes: batch, fed-batch, perfusion, and continuous. Batch processing involves the addition of the media and inoculum at the beginning of the process. The only exchanges with the external environment are gases, pH solutions, antifoam, and some specific agents. Fed-batch processing starts similarly to batch processing but with the added advantage of introducing nutrient feeds during the culture to extend growth and boost productivity. The addition can be continuous or intermittent, with factors such as osmolality directly impacting the CQAs. Fed-batch cultures have increased productivity to batch (Lindskog, 2018). Perfusion processing maintains the bioreactor at constant volume by the continuous addition of fresh media and removal of spent media thereby providing a steady nutrient supply and reducing toxic by-product concentration. This method is effective for long durations of up to 6 months. Perfusion is rarely used for microbial cells due to their swift growth but has gained traction for stable products in mammalian cell bioprocessing. Continuous processing involves a steady flow of media through the reactor; however, in contrast to perfusion processing, there is no mechanism to maintain the biomass within the bioreactor. The process begins as a batch operation; when the culture reaches the exponential growth phase or becomes substrate-limited, a feed with fresh growth medium is charged with an equivalent volume of culture broth removed from the reactor (Lindskog, 2018).

Upstream processing in cell therapy presents challenges such as patient cellular material heterogeneity and variability in operator processing. There is an urgent need for advanced monitoring and control strategies, given the intricacies of cell therapy processing. Critical to this is the identification of critical quality attributes (CQAs) and critical process parameters (CPP), with multifactorial studies determining their direct impact on product quality. Process Analytical Technology (PAT) tools, including online, in-line, and at-line sensors, are crucial for real-time assessments of these CQAs and CPPs. These tools enhance process robustness and product quality, reduce variability, and offer insights into key process variables. Furthermore, they facilitate non-destructive monitoring, potentially reducing costs by minimising manual labour requirements in certain unit operations (Aifuwa, 2022).

To cater to the specific needs of cell therapy PAT, bioprocess suppliers either adapt biologics processing systems or create novel systems tailored for cell therapy. These PAT systems must consider factors such as process scale, cell culture format, and the available cellular material's diversity and limitations, emphasizing the need for non-destructive sampling techniques. One critical process parameter is the viable cell concentration (VCC), which is pivotal for determining process progression, harvest readiness, and informing control strategies. Traditionally measured using offline instruments, there is a push towards PAT tools that offer continuous, operator-free monitoring of VCC to mitigate the variability and costs associated with manual sampling.

Table 1
Impact of expression system on process variables relative to each other. Values for growth rate and cell densities are a function of various parameters such as the specific strain, process operating mode and media.

Process parameter	Bacteria	Yeast	Mammalian
Growth rate	Fast	Fast	Slower
Cell densities	High	High	Lower
Sensitivity to environmental changes	Less Sensitive	Less Sensitive	Highly Sensitive
Post translational modifications	Different types of modifications	Different types of modifications Limited N-linked & O-linked glycosylation	Performs N-linked & O-linked glycosylation
Optimal growth conditions	Prefers ~37 °C; pH: 6.0–7.5	Prefers 20–30 °C; pH: 4.0–6.0	Prefers 37 °C; pH: 7.0–7.4

Spectroscopic sensors, which have been employed for small molecule and biologics processing, are potential tools for in-line monitoring in cell therapy. Sensors such as Dielectric (or impedance), Raman, and fluorescence spectroscopy provide non-invasive, non-destructive means for online monitoring. However, the raw data from these sensors necessitate interpretation through statistical tools, such as principal component analysis (PCA), to correlate with the critical quality attributes (CQAs) they intend to measure. Depending on their specific application, these sensors can indirectly quantify cell concentration through chemometric analysis and measure amino acids and metabolites directly via spectra quantification (Aifuwa, 2022; Corredor et al., 2022).

In the realm of cell therapy expansion processing, rocking motion (RM) bioreactors are prevalent, and dielectric spectroscopy is a primary technique compatible with this culture expansion method. This technique exposes cells to an electric field, causing ion movement within cells and the medium, turning cells into tiny capacitors due to the polarization of their membranes. It allows for the indirect quantification of viable cell concentration (VCC). However, the challenge lies in the fact that permittivity is influenced by cell concentration, volume, and type. Given the dynamic shifts in cell morphology and phenotype, combined with the heterogeneity of patient material, obtaining consistent measurements for cell concentration using dielectric spectrometry becomes challenging. Overcoming these challenges would necessitate the development of statistical tools to isolate these factors, ensuring the technique's successful integration into cell therapy processing (Aifuwa, 2022; Corredor et al., 2022).

Critical parameters such as pH, dissolved oxygen (DO), metabolites, and amino acids are fundamental to monitoring cell therapy processing. Although PAT sensors exist for biologics bioprocessing, there is a plethora of technologies designed specifically for cell therapy processing, particularly in disposable formats suitable for RM expansion systems. While continuous pH and DO monitoring tools are commercially available, the industry sees ongoing developments in miniaturised sensor formats. Monitoring amino acids, essential for cell growth and metabolic pathways, continuously can optimize culture performance. Incorporating rapid amino acid analysers, such as 908 Devices Rebel, might offer in-depth, in-process amino acid characterisation. Additionally, tools such as Sartorius BioPAT Trace or SBI's single-use optical sensors can be vital for monitoring key culture metabolites such as glucose and lactate. Though not yet commercially tailored for autologous cell therapy processing, Raman spectroscopy holds potential as a future tool in this sector (Aifuwa, 2022; Corredor et al., 2022).

3. Navigating the digital twin revolution: Current landscape & future perspectives

Digital twin implementation within the bioprocess industry can offer a plethora of benefits, such as predicting drug behaviour, simulating drug production processes, and enhancing drug discovery. However, despite these advantages, there has been limited implementation observed within the industry. A detailed contemporary study by Perno et al., 2022 analysed the several barriers and enablers to the widespread implementation of DTs as presented in Fig. 5, thereby providing an outlook to future areas of improvement.

Implementation of DTs within the bioprocess industry can face issues relating to system integration. This refers to the challenge of transitioning from outdated systems to modern technologies. It encompasses integrating new systems into the existing ones and harmonising various components of current systems (Perno et al., 2022). As these systems age, the integration challenge amplifies. Additionally, particularly in light of recent high-profile data breaches (Millar, 2021), security is another area of concern with respect to DT implementation. This barrier covers risks related to data handling - from collection and exchange to storage and processing, as well as the protection of intellectual property. Data security is universally crucial, with specific emphasis on its

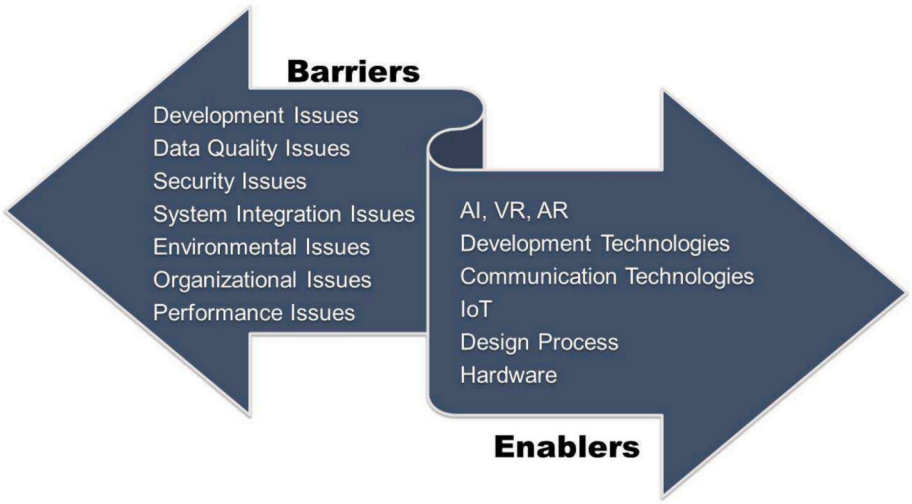

Fig. 5. Summary of enablers and barriers to DT implementation within the biopharmaceutical sector.

significance and the inherent challenges. To address this, blockchain has been recommended to boost security and ensure transparency (Yaqoob et al., 2020; Perno et al., 2022). Additionally, DTs are subjected to constraints in the hardware and software, which affect the data flow between the physical and virtual systems, impacting the performance of DTs, particularly when considering that their defining feature is their capacity to depict and monitor their physical counterparts instantaneously. Facilitating this real-time connection is complex, especially when it involves managing extensive data in real-time (Perno et al., 2022). Furthermore, a company's internal decisions play a pivotal role in DT development. Poor data management – where data is dispersed, isolated, and seldom updated – compromises data quality, security, and integration. Therefore, for a successful digital transformation with DTs, companies must establish a robust data management system (Perno et al., 2022).

Despite these challenges, there are a number of enablers to DT implementation which are being developed by various sectors. Recent advancements in Artificial Intelligence (AI), Virtual Reality/ Augmented Reality (VR/AR), and other development technologies (as discussed earlier in this paper) are addressing DT implementation challenges. AI, combined with big data analytics, aids in creating tools and models that can detect patterns and correlations and make predictions based on real asset data. VR and AR have emerged as key technologies that enhance the usability of DTs and assist in creating training systems for new operators. Advancements in simulation software have facilitated the creation of detailed models that mimic physical assets accurately. For enhanced simulation performance, techniques such as reduced-order models (ROMs) are utilised. With these technological advancements, the time and resources needed for DT development are minimised, and model accuracy is increased. Furthermore, improvements in software and communication technologies are addressing data quality issues. Ensuring the data's availability, accuracy, and quality for DTs is crucial. To mitigate potential impacts on DT performance due to these challenges, there is a need for comprehensive groundwork to enable efficient data transfer between the physical entity and its DT (Perno et al., 2022). Utilising cloud and edge computing alongside reliable communication protocols ensures data quality and validity. Centralised databases further enhance data quality, especially when sourcing from diverse origins. Security is bolstered through communication technologies such as OPC-UA, MTConnect, and MQTT, which offer features such as encryption and secure socket layers. Although safeguarding sensitive data transfer between DTs and physical assets demands significant effort, these protocols have proven effective in industrial settings. System integration challenges are addressed by innovations in IoT/IIoT, with more vendors offering unified platforms that align with existing systems. Overcoming integration hurdles between legacy and new IoT/IIoT systems, vendor-neutral protocols like OPC-UA are being employed (Perno et al., 2022; M. Rathore et al., 2021; González-Figueredo et al., 2019). With respect to the human factor, in order to achieve successful adoption, knowledge enhancement is critical. The adoption of Industry 4.0 standards equips companies with a solid data framework for DT development. Additionally, training the workforce enhances the environment as it fosters the creation of new DT methodologies, tools, and standards. Organisational challenges are addressed through technologies that enable knowledge and refined design strategies. The retraining of staff and the adoption of Industry 4.0 standards help businesses stay updated with current tech trends. This ensures an adequate supply of DT expertise, facilitating informed decisions. Clearly defining DT requirements also aids in formulating well-informed decisions regarding data management and pertinent technologies prior to the onset of DT development. On the performance side, the rising accessibility of robust hardware, combined with decreasing costs and growing computational strength, allows for detailed digital representations of physical assets using sophisticated simulation and ML models. This guarantees smooth real-time interactions between digital twins and their physical counterparts (Perno et al., 2022).

4. Conclusion

Digital Twins (DTs) are tools used to predict system behaviour based on historical data. The accuracy of these predictions relies on the mathematical models the DT uses, which must be accurate reflections of the system. The challenge lies in creating models tailored to specific purposes and scenarios. Various mathematical models have been created for bioprocess operations; these models are either knowledge-centric, rooted in foundational principles, or data-centric. The latter approach is popular in industry due to its systematic nature and alignment with the Quality by Design (QbD) principle. However, the data-centric method has notable disadvantages. It often necessitates an increased number of experiments, particularly with an increasing number of process parameters, leading to a "curse of dimensionality". Additionally, insights from this method are generally limited to the specific data set analysed, which restricts the broader application of the knowledge gained. The knowledge-centric approach offers a deeper comprehension of systems and is built on equations refined over time. This method allows for knowledge transfer across projects as physical laws remain consistent. However, its application can be challenging due to the extensive understanding and expertise needed, especially as the

number of process parameters and their potential interactions increase (Sokolov et al., 2021). As a solution, hybrid modelling has been introduced, blending the knowledge-centric approach with data-driven methods. Hybrid models leverage the depth of the knowledge-centric method and the simplicity of the data-centric approach. They provide a cost-effective and modular way to understand key system behaviours, enhancing the effectiveness of Digital Twins (DTs).

The advent of automation in bioprocessing is pivotal for robust process control, with digitalisation being the contemporary focus. Developments in sensors, both inline and at-line, wireless technology, and advancements in compact sensor calibration are revolutionising the space. However, seamless integration of automated workstations into continuous workflows remains a challenge due to inconsistencies in software versions, proprietary interfaces, and data formats. The overarching need for successful integration is the utilisation of standardised communication protocols and user-friendly graphical interfaces (A. S. Rathore et al., 2021). Despite the integration of enterprise control systems with various business and shop-floor systems posing a significant challenge, there is demonstrated progress within this field. Platforms such as SCADA illustrate end-to-end process integration, and the digital twin concept has been demonstrated for small-scale production (A. S. Rathore et al., 2021).

Navigating the complexities of bioprocesses whilst capturing their overall characteristics and variability through modelling is challenging, particularly due to the limited presence of reliable online sensors. Most process variables are assessed offline or online, leading to increased costs and inefficient information delivery. Furthermore, the durability of online sensors in harsh environments and manual operations of at-line sensors further complicate data acquisition. The solution has been the rise of cybernetical-physical systems controlled by soft sensors and algorithms. Soft sensors have emerged as pivotal tools for real-time evaluation, showcasing potential in assessing biomass, product concentrations, and other crucial quality attributes (CQAs). However, for real-time implementation, advancements are needed in non-invasive analytical techniques, sensor adaptability across production systems, regulatory compliance, and user-friendly interfaces for seamless digitalisation and data integration.

Author contributions

KM Conceptualization, investigation, and writing of the research; FVSdS Conceptualization and revision of the paper, MSG Conceptualization and revision of the paper; JM contributed to technologies and paper revision from an industrial perspective, and all authors revised and approved final paper.

AI disclaimer

During the preparation of this work the author(s) used ChatGPT at concept stage of the review to aid in the formation of the structure. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare no competing interests.

References

Aifuwa, I., 2022. PAT strategies and applications for cell therapy processing. In: Current Opinion in Biomedical Engineering, vol. 24. Elsevier B.V. https://doi.org/10.1016/j.cobme.2022.100405

Alhuthali, S., Kontoravdi, C., 2022. Population balance modelling captures host cell protein dynamics in CHO cell cultures. PLoS One 17 (3 March). https://doi.org/10.1371/journal.pone.0265886.

Alsadik, B., 2019. Kalman Filter. In Adjustment Models in 3D Geomatics and Computational Geophysics (pp. 299–326). Elsevier. https://doi.org/10.1016/b978-0-12-817588-0.00010-6.

Alt, N., Zhang, T.Y., Motchnik, P., Taticek, R., Quarmby, V., Schlothauer, T., Beck, H., Emrich, T., Harris, R.J., 2016. Determination of critical quality attributes for monoclonal antibodies using quality by design principles. Biologicals 44 (5), 291–305. https://doi.org/10.1016/j.biologicals.2016.06.005.

Benyahia, B., Brumano, L.P., Pessoa, A., Santos da Silva, F.V., 2020. Biopharmaceutical development, production, and quality. In: New and Future Developments in Microbial Biotechnology and Bioengineering: Microbial Biomolecules: Properties, Relevance, and their Translational Applications. Elsevier, pp. 69–89. https://doi.org/10.1016/B978-0-444-64301-8.00003-2.

Blasch, E., Pham, T., Chong, C.Y., Koch, W., Leung, H., Braines, D., Abdelzaher, T., 2021. Machine learning/artificial intelligence for sensor data fusion-opportunities and challenges. IEEE Aerosp. Electron. Syst. Mag. 36 (7), 80–93. https://doi.org/10.1109/MAES.2020.3049030.

Bleisinger, O., Malek, C., Holbach, S., 2022. Machine learning based simulation for design space exploration. Proceedings of the Design Society 2, 1521–1530. https://doi.org/10.1017/pds.2022.154.

Brunner, V., Siegl, M., Geier, D., Becker, T., 2021. Challenges in the development of soft sensors for bioprocesses: A critical review. In: Frontiers in Bioengineering and Biotechnology, vol. 9. Frontiers Media S.A. https://doi.org/10.3389/fbioe.2021.722202.

Chen, Y., Yang, O., Sampat, C., Bhalode, P., Ramachandran, R., Ierapetritou, M., 2020. Digital twins in pharmaceutical and biopharmaceutical manufacturing: a literature review. In: Processes (Vol. 8, Issue 9). MDPI AG. https://doi.org/10.3390/pr8091088.

Chopda, V., Gyorgypal, A., Yang, O., Singh, R., Ramachandran, R., Zhang, H., Tsilomelekis, G., Chundawat, S.P.S., Ierapetritou, M.G., 2022. Recent advances in integrated process analytical techniques, modeling, and control strategies to enable continuous biomanufacturing of monoclonal antibodies. J. Chem. Technol. Biotechnol. Vol. 97 (Issue 9), 2317–2335. John Wiley and Sons Ltd. https://doi.org/10.1002/jctb.6765.

Corredor, C., Ketcham, S., McHugh, K., Lane, G., Vale, N., Rehmann, M., Both, D., 2022. Process analytical technologies (PAT) applications for upstream biologics. Am. Pharm. Rev. (preprint). In: https://www.americanpharmaceuticalreview.com/Featured-Articles/585192-Process-Analytical-Technologies-PAT-Applications-for-Upstream-Biologics/.

Ding, B., 2018. Pharma industry 4.0: literature review and research opportunities in sustainable pharmaceutical supply chains. Process. Saf. Environ. Prot. 119, 115–130. https://doi.org/10.1016/j.psep.2018.06.031.

Gargalo, C.L., Udugama, I., Pontius, K., Lopez, P.C., Nielsen, R.F., Hasanzadeh, A., Mansouri, S.S., Bayer, C., Junicke, H., Gernaey, K.V., 2020. Towards smart biomanufacturing: a perspective on recent developments in industrial measurement and monitoring technologies for bio-based production processes. J. Ind. Microbiol. Biotechnol. 47 (11), 947–964. https://doi.org/10.1007/s10295-020-02308-1.

Gargalo, C.L., de Las Heras, S.C., Jones, M.N., Udugama, I., Mansouri, S.S., Krühne, U., Gernaey, K.V., 2021. Towards the development of digital twins for the bio-manufacturing industry. Adv. Biochem. Eng. Biotechnol. 176, 1–34. https://doi.org/10.1007/10_2020_142.

Geremia, M., Bezzo, F., Ierapetritou, M.G., 2024. Design space determination of pharmaceutical processes: effects of control strategies and uncertainty. Eur. J. Pharm. Biopharm. 194, 159–169. https://doi.org/10.1016/j.ejpb.2023.12.008.

Gerzon, G., Sheng, Y., Kirkitadze, M., 2022. Process analytical technologies – advances in bioprocess integration and future perspectives. J. Pharm. Biomed. Anal. 207 https://doi.org/10.1016/j.jpba.2021.114379.

González-Figueredo, C., Alejandro Flores-Estrella, R., Rojas-Rejón, A., O., 2019. Fermentation: metabolism, kinetic models, and bioprocessing. In: Current Topics in Biochemical Engineering. IntechOpen. https://doi.org/10.5772/intechopen.82195.

Goos, G., Bertino, E., Gao, W., Steffen, B., Woeginger, G., Aachen, R., Aachen, G., Moti, Y., 2021. Computer safety, reliability, and security. http://www.springer.com/series/7408.

Grangeia, H.B., Silva, C., Simões, S.P., Reis, M.S., 2020. Quality by design in pharmaceutical manufacturing: a systematic review of current status, challenges and future perspectives. In: European Journal of Pharmaceutics and Biopharmaceutics (Vol. 147, pp. 19–37). Elsevier B.V. https://doi.org/10.1016/j.ejpb.2019.12.007

Gupta, S., Kar, A.K., Baabdullah, A., Al-khowaiter, W.A.A., 2018. Big data with cognitive computing : a review for the future. Int. J. Inf. Manag. 42 (June), 78–89. https://doi.org/10.1016/j.ijinfomgt.2018.06.005.

Helgers, H., Hengelbrock, A., Schmidt, A., Strube, J., 2021. Digital twins for continuous mrna production. Processes 9 (11). https://doi.org/10.3390/pr9111967.

Horst, J.P., Turimella, S.L., Metsers, F., Zwiers, A., 2021. Implementation of quality by design (QbD) principles in regulatory dossiers of medicinal products in the European Union (EU) between 2014 and 2019. Ther. Innov. Regul. Sci. 55 (3), 583–590. https://doi.org/10.1007/s43441-020-00254-9.

ICH, 2009. ICH guideline Q8 (R2) on pharmaceutical development. https://www.ema.europa.eu/en/documents/scientific-guideline/international-conference-harmonisation-technical-requirements-registration-pharmaceuticals-human-use-considerations-ich-guideline-q8-r2-pharmaceutical-development-step-5_en.pdf.

Jagschies, G., 2018. Brief review of the biopharmaceutical and vaccine industry. In: Biopharmaceutical Processing: Development, Design, and Implementation of Manufacturing Processes. Elsevier, pp. 33–58. https://doi.org/10.1016/B978-0-08-100623-8.00002-5.

Jelsch, M., Roggo, Y., Kleinebudde, P., Krumme, M., 2021. Model predictive control in pharmaceutical continuous manufacturing: a review from a user's perspective. Eur. J. Pharm. Biopharm. 159, 137–142. https://doi.org/10.1016/j.ejpb.2021.01.003.

Jiang, Y., Yin, S., Dong, J., Kaynak, O., 2021. A review on soft sensors for monitoring, control, and optimization of industrial processes. In: IEEE Sensors Journal, Vol. 21,

Issue 11. Institute of Electrical and Electronics Engineers Inc., pp. 12868–12881. https://doi.org/10.1109/JSEN.2020.3033153
Juarez, M.G., Botti, V.J., Giret, A.S., 2021. Digital twins: review and challenges. J. Comput. Inf. Sci. Eng. 21 (3) https://doi.org/10.1115/1.4050244.
Khodarahmi, M., Maihami, V., 2023. A Review on Kalman Filter Models. In: Archives of Computational Methods in Engineering Vol. 30, Issue 1. Springer Science and Business Media B.V, pp. 727–747. https://doi.org/10.1007/s11831-022-09815-7.
Klein, K., Hazell, L., Stolk, P., Shakir, S., 2020. The UK BIO-TRAC study: a cross-sectional study of product and batch traceability for biologics in clinical practice and electronic adverse drug reaction reporting in the UK. Drug Saf. 43 (3), 255–263. https://doi.org/10.1007/s40264-019-00891-6.
Kong, L., Peng, X., Chen, Y., Wang, P., Xu, M., 2020. Multi-sensor measurement and data fusion technology for manufacturing process monitoring: a literature review. In: International Journal of Extreme Manufacturing (Vol. 2, Issue 2). IOP Publishing Ltd. https://doi.org/10.1088/2631-7990/ab7ae6
Kornecki, M., Strube, J., 2019. Accelerating biologics manufacturing by upstream process modelling. Processes 7 (3). https://doi.org/10.3390/PR7030166.
Lindskog, E.K., 2018. The Upstream Process: Principal Modes of Operation. In: Biopharmaceutical Processing: Development, Design, and Implementation of Manufacturing Processes. Elsevier Ltd. https://doi.org/10.1016/B978-0-08-100623-8.00031-1
Liu, S., 2020. Bioprocess Engineering : Kinetics, Sustainability, and Reactor Design. Elsevier.
Luo, Y., Kurian, V., Ogunnaike, B.A., 2021. Bioprocess systems analysis, modeling, estimation, and control. In: Current Opinion in Chemical Engineering, vol. 33. Elsevier Ltd. https://doi.org/10.1016/j.coche.2021.100705
Maddikunta, P.K.R., Pham, Q.V., B, P., Deepa, N., Dev, K., Gadekallu, T.R., Ruby, R., Liyanage, M., 2022. Industry 5.0: A survey on enabling technologies and potential applications. In: Journal of Industrial Information Integration (Vol. 26). Elsevier B.V. https://doi.org/10.1016/j.jii.2021.100257
McAfee, M., Kariminejad, M., Weinert, A., Huq, S., Stigter, J.D., Tormey, D., 2022. State estimators in soft sensing and sensor fusion for sustainable manufacturing. In: Sustainability (Switzerland) (Vol. 14, Issue 6). MDPI. https://doi.org/10.3390/su14063635.
Metta, N., Ghijs, M., Schäfer, E., Kumar, A., Cappuyns, P., Van Assche, I., Singh, R., Ramachandran, R., De Beer, T., Ierapetritou, M., Nopens, I., 2019. Dynamic flowsheet model development and sensitivity analysis of a continuous pharmaceutical tablet manufacturing process using the wet granulation route. Processes 7 (4), 1–35. https://doi.org/10.3390/pr7040234.
Millar, A., 2021. Pharma cyber attacks: Five pharma cybersecurity breaches to know and learn from. Pharm. Technol. [Preprint]. https://www.pharmaceutical-technology.com/features/pharma-cyber-attacks/.
Moser, A., Appl, C., Brüning, S., Hass, V.C., 2021. Mechanistic mathematical models as a basis for digital twins. Adv. Biochem. Eng. Biotechnol. 176, 133–180. https://doi.org/10.1007/10_2020_152.
Narayanan, H., Luna, M., Sokolov, M., Butté, A., Morbidelli, M., 2022. Hybrid models based on machine learning and an increasing degree of process knowledge: application to cell culture processes. Ind. Eng. Chem. Res. 61 (25), 8658–8672. https://doi.org/10.1021/acs.iecr.1c04507.
Narayanan, H., von Stosch, M., Feidl, F., Sokolov, M., Morbidelli, M., Butté, A., 2023. Hybrid modeling for biopharmaceutical processes: Advantages, opportunities, and implementation. In: Frontiers in Chemical Engineering, vol. 5. Frontiers Media S.A. https://doi.org/10.3389/fceng.2023.1157889
Nayak, A.K., Ahmed, S.A., Beg, S., Tabish, M., Hasnain, M.S., 2019. Application of quality by Design for the Development of biopharmaceuticals. In: Pharmaceutical Quality by Design: Principles and Applications. Elsevier, pp. 399–411. https://doi.org/10.1016/B978-0-12-815799-2.00019-8.
Park, S.Y., Park, C.H., Choi, D.H., Hong, J.K., Lee, D.Y., 2021. Bioprocess digital twins of mammalian cell culture for advanced biomanufacturing. In: Current Opinion in Chemical Engineering, vol. 33. Elsevier Ltd. https://doi.org/10.1016/j.coche.2021.100702
Perno, M., Hvam, L., Haug, A., 2022. Implementation of digital twins in the process industry: A systematic literature review of enablers and barriers. In: Computers in Industry, vol. 134. Elsevier B.V. https://doi.org/10.1016/j.compind.2021.103558
Rathore, A., Winkle, H., 2009. Quality by design for biopharmaceuticals. Nat. Biotechnol. 27 (1).
Rathore, M., Shah, S.A., Shukla, D., Bentafat, E., Bakiras, S., 2021. The role of AI, machine learning, and big data in digital twinning: A systematic literature review, challenges, and opportunities. In: IEEE Access, vol. 9. Institute of Electrical and Electronics Engineers Inc., pp. 32030–32052. https://doi.org/10.1109/ACCESS.2021.3060863
Rathore, A.S., Mishra, S., Nikita, S., Priyanka, P., 2021a. Bioprocess control: current progress and future perspectives. In: Life (Vol. 11, Issue 6). MDPI AG. https://doi.org/10.3390/life11060557.
Rathore, A.S., Mishra, S., Nikita, S., Priyanka, P., 2021b. Bioprocess control: current progress and future perspectives. In: Life (Vol. 11, Issue 6). MDPI AG. https://doi.org/10.3390/life11060557.
Reyes, S.J., Durocher, Y., Pham, P.L., Henry, O., 2022. Modern sensor tools and techniques for monitoring, controlling, and improving cell culture processes. In: Processes (Vol. 10, Issue 2). MDPI. https://doi.org/10.3390/pr10020189.
Rösner, L.S., Walter, F., Ude, C., John, G.T., Beutel, S., 2022. Sensors and Techniques for On-Line Determination of Cell Viability in Bioprocess Monitoring. In: Bioengineering (Vol. 9, issue 12). MDPI. https://doi.org/10.3390/bioengineering9120762.
Sachio, S., Kontoravdi, C., Papathanasiou, M.M., 2023. A model-based approach towards accelerated process development: a case study on chromatography. Chem. Eng. Res. Des. 197, 800–820. https://doi.org/10.1016/j.cherd.2023.08.016.
Sansana, J., Joswiak, M.N., Castillo, I., Wang, Z., Rendall, R., Chiang, L.H., Reis, M.S., 2021. Recent trends on hybrid modeling for industry 4.0. In: Computers and Chemical Engineering, vol. 151. Elsevier Ltd. https://doi.org/10.1016/j.compchemeng.2021.107365
Schlindwein, W., Gibson, M., 2018. Pharmaceutical Quality by Design: A Practical Approach.
Sharma, D., Patel, P., Shah, M., 2023. A comprehensive study on industry 4.0 in the pharmaceutical industry for sustainable development. Environ. Sci. Pollut. Res. https://doi.org/10.1007/s11356-023-26856-y.
Sokolov, M., von Stosch, M., Narayanan, H., Feidl, F., Butté, A., 2021. Hybrid modeling — A key enabler towards realizing digital twins in biopharma?. In: Current Opinion in Chemical Engineering, vol. 34. Elsevier Ltd. https://doi.org/10.1016/j.coche.2021.100715
Tsopanoglou, A., Jiménez del Val, I., 2021. 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, vol. 32. Elsevier Ltd. https://doi.org/10.1016/j.coche.2021.100691
Udugama, I.A., Lopez, P.C., Gargalo, C.L., Li, X., Bayer, C., Gernaey, K.V., 2021. Digital twin in biomanufacturing: challenges and opportunities towards its implementation. Sys.Microbiol. Biomanufact. 1 (3), 257–274. https://doi.org/10.1007/s43393-021-00024-0.
Yaqoob, I., Salah, K., Uddin, M., Jayaraman, R., Omar, M., Imran, M., 2020. Blockchain for digital twins: recent advances and future research challenges. IEEE Netw. 34 (5), 290–298. https://doi.org/10.1109/MNET.001.1900661.