A Digital Twin Framework for Bioprocess Development Using IoT Sensor Data

Zahra Dasht Bozorgi[1*], Artem Polyvyanyy[1†], Marcello La Rosa[1†], Ellen Otte[2], Abel Armas-Cervantes[1]

[1*]Computing and Information Systems, The University of Melbourne, Parkville, Melbourne, 3010, Victoria, Australia.
[2]Biopharmaceutical Product Development, CSL Limited, Melbourne, 3000, Victoria, Australia.

*Corresponding author(s). E-mail(s): zahra.dashtbozorgi@unimelb.edu.au;
Contributing authors: artem.polyvyanyy@unimelb.edu.au;
marcello.larosa@unimelb.edu.au; ellen.otte@csl.com.au;
abel.armas@unimelb.edu.au;
†These authors contributed equally to this work.

Abstract

The advent of Industry 4.0 has revolutionized the manufacturing landscape, introducing advanced technologies such as the Internet of Things (IoT) to optimise production processes. This book chapter proposes a comprehensive digital twin framework that harnesses the power of IoT technologies, specifically sensor data, to enhance efficiency and performance in process industries. Our proposed framework is designed to leverage real-time sensor data from bioprocesses and combine it with manually collected data to create a virtual representation of the real process. This envisioned digital twin not only mirrors the current state of the system but also enables predictive analysis and proactive decision-making. We discuss how using the notion of process, as defined in the business process management area, enables the integration of various components of bioprocesses and how IoT-enabled technologies can create a real-time connection between the physical and virtual processes. We also discuss the challenges of realizing the proposed digital twin and offer potential solutions to those challenges. We demonstrate the applicability of the proposed framework via a case study with a large pharmaceutical company. In particular, in the context of predictive process monitoring, we show the current baseline can be outperformed if our framework is adopted.

Postprint, August 2025

1 Introduction

A bioprocess refers to a series of steps or operations designed to harness living organisms (such as bacteria, yeast, or mammalian cells) or their cellular components to produce desired products or carry out specific biochemical transformations. These processes are commonly employed in various industries, including pharmaceuticals, biotechnology, food and beverage, agriculture, and environmental remediation. In bioprocessing, living organisms are cultivated under controlled conditions, typically in bioreactors or fermentation tanks, where they are provided with nutrients and other necessary conditions for growth and metabolism. The organisms then produce the desired products through metabolic pathways or biochemical reactions. Bioprocesses can involve a wide range of activities, including cell culture, fermentation, purification, and downstream processing.

Typically, bioprocesses are composed of several *unit operations* which are executed sequentially. The process starts by thawing vials of frozen microorganisms. The cells are then grown in a cell culture where they are fed nutrients to help growth and target protein production. The cell culture unit operations are known as upstream processing. The goal of upstream is to produce as much high-quality product as possible. After the production steps, several purification steps are carried out to remove contaminants and purify the products. These steps are collectively known as downstream processing. The last operation after the downstream steps is known as formulation. In the formulation step, the product of interest is prepared as a substance that can safely be used by customers. Figure 1 shows the upstream, downstream, and formulation unit operations.

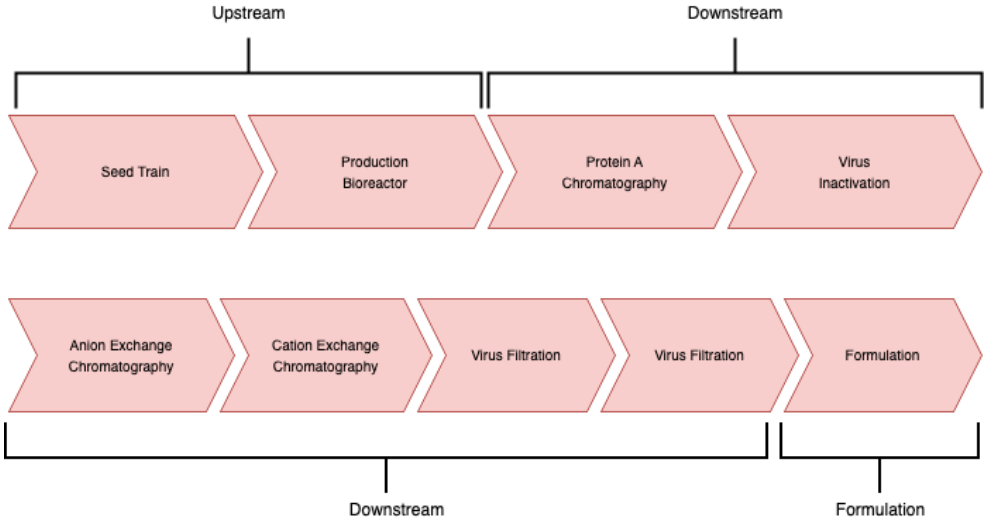


Fig. 1: Unit Operation of a typical bioprocess

In recent years, bioprocess industries have experienced a paradigm shift towards the adoption of digital technologies to streamline and optimize bioprocess development. With the increasing complexity and diversity of bioproducts, there is a pressing need for advanced methodologies to effectively capture, analyze, and optimize the intricate

processes involved in bioproduction. Digital twin technology [1] has emerged as a promising approach to address these challenges by creating virtual replicas of physical systems, enabling real-time monitoring, analysis, and optimization. However, the application of digital twins in the realm of bioprocess development remains relatively unexplored and is often confined to a single step in the bioprocess. This results in multiple disconnected digital twins that do not capture the relationship between the various unit operations. This makes optimization and long-term decisions challenging.

This book chapter presents a novel digital twin framework encompassing all unit operations in a typical end-to-end bioprocess. The novelty of our framework lies in the integration of concepts and techniques from business process management and process mining, which are widely used for analysis and improvement of business processes in a variety of industries. Using these concepts, it is possible to represent bioprocesses as a series of human and biological steps, as well as leverage data collected during the execution of the bioprocesses to identify improvement opportunities.

In the proposed digital twin framework, two types of improvements are considered: short-term and long-term. Short-term improvements are done on the fly during process execution. For example, using a controller that learns its strategy from the digital twin's process control component, a short-term strategy is to dynamically adjust the parameters to keep the process in stable condition. Long-term tactical improvements are those that can be derived from the digital twin and are implemented in future instances of the process. For example, using the digital twin, one might determine an optimized feeding strategy which is then used in future bioprocess executions. Our envisioned digital twin should be equipped to address both types of improvement as will be discussed later in the chapter.

The rest of this chapter is organized as follows. We begin by reviewing the current literature on digital twins in both process mining and bioprocessing domains in Section 2. We then present the envisioned framework, its various components, and use cases in Section 3. In Section 4, we describe the challenges of achieving the envisioned framework. We present an example use case and partial implementation of the proposed digital twin as a proof of concept in Section 5, and finally, conclude the chapter in Section 6.

2 Related Work

The related work reviewed in this section is divided into two subsections. Subsection 2.1 reviews existing works on digital twins of organizational processes. Existing works on digital twins of bioprocesses are reviewed in Subsection 2.2.

2.1 Organisational Digital Twins

Since their conception, digital twins [1] have been adopted in various settings. One such setting is the use of Digital Twin technology to represent organizational business processes. While a digital twin is a virtual representation of real-life phenomenon that is indistinguishable from its real-life counterpart, van der Aalst et al. [2] describes digital model and digital shadow as other virtual representations that vary in their level of integration with the real-life phenomena they represent. Figure 3 shows the

difference between 1. digital model, 2. digital shadow, and 3. digital twin. A digital model is created using manual and offline collected data from a real process. There is no real-time connection between reality and the digital model. The dotted lines between reality and the digital model represent the loose coupling between them. A digital shadow goes one step further in connection to reality. It is synchronized with reality by having real-time data fed into it (solid arrow from reality to the digital shadow), however, any insights and decisions resulting from the model are fed back to reality offline and manually. The digital twin goes further than digital shadows by having an online feedback loop connecting the digital twin to reality. Most existing process mining techniques are either digital models or digital shadows, while digital twins currently remain aspirational.

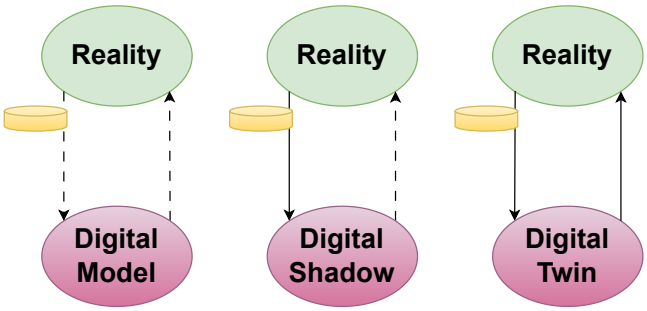


Fig. 2: Differences between a digital model, a digital shadow, and a digital twin (Figure adapted from [2])

The history of the emergence of the digital twin concept is explored in [3]. The authors explore the role of digital shadows as building blocks for a broader digital twin. Additionally, it builds a connection between digital shadows and mathematical (or physical) models by showing examples of how such models are currently used in building digital shadows. However, such work is mainly concerned with the conceptual link between data management and algorithmic services and is not concerned with the challenges of implementing digital shadows and organizational twins.

Achieving a digital twin of organizations is highly desirable since it provides a virtual environment where different actions and decisions can be tested without sacrificing quality or wasting resources in the real world. Many works using the notion of a process to build a digital replica of a workflow have been proposed in the literature. For example, in the work by van der Aalst et al. [4], the authors discuss the limitations of modelling paradigms such as Turing machines and Markov chains. Such models are limited in their expressive power – e.g., they cannot capture concurrency between activities, a key aspect of organizational business processes – and therefore, are limited in their capability to create a digital twin of an organization. Therefore, the authors propose the use of more sophisticated process modelling languages, such as object-centric Petri-nets [5], to represent the workflow and interaction of objects within organizations. The benefits of object-centric Petri nets in digital twinning are further explored in [6].

In another work, Park and van der Aalst [7] present a digital twin of an organization as the digital replica of a production process or the entire organization. Their proposed digital twin includes an interface model representing the business process and its

supporting information systems. This representation is in the form of a process model (e.g., a Petri net). The interface model presents the process analyst with the current state of the process and highlights bottlenecks and configurations. The analyst then defines process constraints and suitable actions. The constraints and actions are fed into an action engine that continuously monitors the processes and triggers actions based on the monitoring results. Another work by Park et al. [8] uses digital twins of organizations to assess the effect of information system updates in a process-aware manner. In this work, they define digital twins interface models as object-centric Petri nets. Updates in the information system are modelled as different configurations of the Petri net model.

The connection between process prediction and digital twins is explored in an article by Brockhoff et al. [9]. In this work, the digital twin is defined as a software system that actively represents, controls, and optimizes a cyber-physical system. This digital twin incorporates process mining services such as process discovery and conformance checking components to derive digital shadows and actionable insights. While this work mentions extraction of underlying processes using continuous measurements, they do not discuss the challenges of such extractions and possible solutions to these challenges, which is the focus of this book chapter.

Bano et al. [10] present a digital twin cockpit that is process-aware. They define this cockpit as the user interaction part of the digital twin that can handle processes related to the physical object. Similar to our proposed approach, they incorporate sensor data. They use the sensors to create events and extract event logs from sensors installed in a healthcare setting. However, their work also deals with discrete events (e.g., a nurse entering the patient's room) whereas in our work, we aim to create digital twins of objects involving processes with no discrete events.

Biological processes and their modelling paradigms are mentioned in the work by Becker and Pentland [11]. In this work, they propose to use inspirations from biological models such as regulatory network modelling and incorporate those inspirations in creating digital twins of organizations. However, they do not explore modelling of bioprocesses themselves as we do in this work.

The works mentioned in this section offer valuable insights into how process mining concepts and methodologies can be used in creating digital twins of processes and organizations. However, these works mainly focus on the complexities of various inter-related processes that exist within organizations., while in this book chapter we seek to use the notion of process and harness the power of process mining techniques to deal with the challenges of twinning biological processes and their interaction to organizational workflows.

2.2 Bioprocess Digital Twins

In recent years, digital twins have gained traction in the field of bioprocess development. As mentioned in Section 1, bioprocesses are comprised of several unit operations each serving a different aim in the development of the bioprocess. Each of these unit operations has its unique associated challenges, such as heterogeneous data types, varying degrees of uncertainty, and complexity of the steps involved. Therefore, the majority of existing works in the bioprocessing field focus on a single unit operation

rather than the end-to-end process. In this section, we review existing work on digital twinning in the bioprocess development field grouping them by the unit operation targeted.

2.2.1 Seed Train

After thawing up a small vial of cells, the first step is to increase their number in a unit operation known as the seed train. There are several challenges such as complex cell metabolism, batch-to-batch variation, the uncertainty of cell behaviour, and the effects of cultivation conditions. Rodriguez and Frahm [12], outline the necessities of digitizing the seed train step. However, their envisioned digital twin is a mechanistic model fit to observable parameters that predicts key performance indicators of the seed train, such as viable cell density. Additionally, their digital twin does not automate decision making. Decisions such as passaging strategy are still made manually using the prediction model. They also explore the role of data-driven strategies and uncertainty estimation.

2.2.2 Production Bioreactor

This step comes after cell expansion and is one of the more complex unit operations due to its high uncertainty [13]. Consequently, the majority of digital twin literature in the bioprocessing field focuses on this step.

Park et al. [14] propose a digital twin framework that leverages process analytical technologies (PAT). In particular, they explore new data collection technologies such as soft sensors, and their impact on achieving realtime synchronization between the real process and the digital twin. Their envisioned framework considers the digital twin to be a mechanistic or data-driven model or a hybrid of the two. However, this view is limiting because such models are developed for single unit operations. To the best of our knowledge, no mechanistic or data-driven model presents the process end-to-end. Therefore, using this framework, we would need to have multiple unconnected digital twins for each unit operation, making optimization of the entire process challenging. Nevertheless, this framework is widely adopted in the bioprocess literature with many works proposing to create digital twins of the bioreactor step [15–18].

2.2.3 Downstream Purification

The downstream steps occur after the bioreactor step to separate the product of interest from other byproducts and remove impurities. Khuat et al. [19] identify typical problems addressed in the downstream process step that benefits from digitization and twinning:

- Monitoring and prediction problems in the capture chromatography step
- Monitoring and prediction problems in the polishing chromatography step
- Control of a chromatography process
- Scaleup and prediction problems in Filtration
- Optimisation of purification sequences

A rich body of literature proposes to tackle the problems stated above. For instance, Tiwari et al. [20] propose digital twins as a means to design a control strategy for the

chromatography step. Similar works propose data-driven approaches for prediction problems in chromatography [21], scale-up problems in filtration [22], and various optimization problems in the downstream [23, 24]. In the majority of these works, building a real-time connection between the models and the real process is not clear. Moreover, the models are developed for various purposes using heterogeneous data sources. Therefore there is no clear link between the various steps of the downstream.

2.2.4 Generic Digital Twin Frameworks

It is clear that while having unit operation digital twins is very useful for solving various tasks, it is not sufficient for optimizing the entire process. Another line of work centers on a more generic idea of digital twinning. For instance, Taylor et al. [25] suggest holistic process modelling as the main engine for digital twins. They suggest improvements such as combining small and large-scale data for model building, improving uncertainty intervals, and establishing extrapolation procedures for non-controllable parameters. However, they do not specify a modelling paradigm for multi-unit operation digital twins.

In [26, 27], the authors investigate digital twins for design of experiments. Their goal is to reduce the number of experiments by identifying critical process parameters via in-silico modelling. Another work by Sokolov et al. [28] argues for the use of hybrid modelling (models consisting of both data-driven and mechanistic components) as the enabler of digital twins. This opinion is supported in other works as well [29, 30]. None of the works mentioned above, however, propose any methodology for connecting the various model components and representing the end-to-end process as a unified entity. We aim to address this gap via the framework proposed in the next section.

3 Framework

In the realm of process industries that utilize biologics, the digital twin concept has emerged as a transformative methodology, aiming to revolutionize processes to enhance efficiency, quality assurance, and adaptability. In this section, we clarify the context and scope of the digital twin and propose a conceptual framework and its various use cases. The digital twin concept involves the creation of a dynamic and interactive virtual model that emulates the behaviour of its physical counterpart in real-time. By combining data from sensors and manual inputs, the digital twin becomes an accurate and responsive representation of the bioprocess. This virtual replica serves as a powerful tool for various tasks. We describe a comprehensive end-to-end digital twin that contains upstream, downstream, and formulation unit operations to provide an overall picture of the process to end users. There are many tasks that can utilize a digital twin. We group all these tasks into four categories, namely process characterization, process optimization, process monitoring, and process control. Therefore, the envisioned digital twin consists of four key components each dedicated to one category.

We define the digital twin as a process model that acts as the digital replica of the physical process. This digital replica is close enough to reality that it can be used for process characterization, control, monitoring, and optimization. The envisioned digital twin is connected to the physical process in real-time in a bidirectional way. It receives

data from the physical process in an online manner and provides feedback either directly or indirectly to the physical process via automated or manual actions. This information transfer occurs via the data and action units as described in Section 3.2. Knowledge about specific tasks and unit operations can be incorporated via model plug-ins as described in Section 3.3

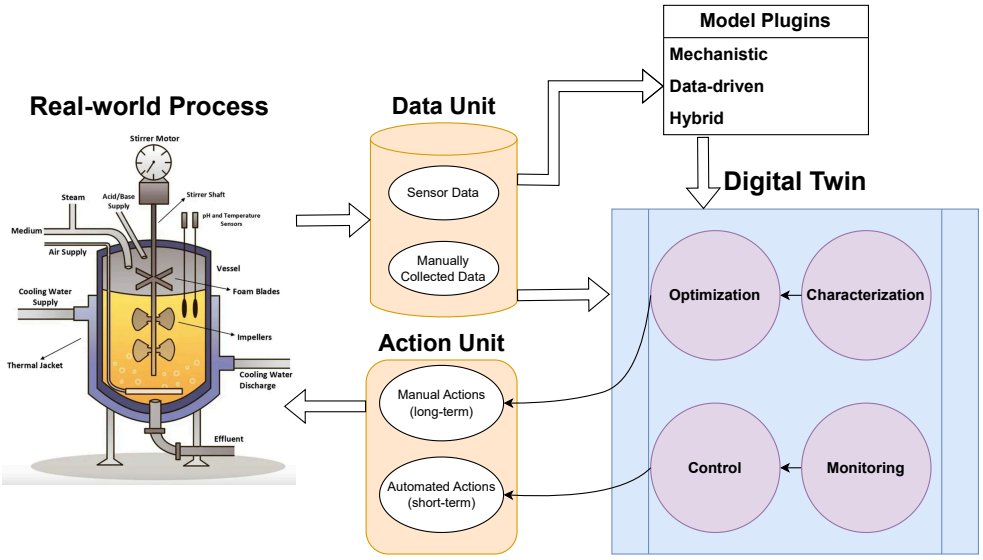


Fig. 3: Proposed Framework

3.1 The Digital Twin

3.1.1 Process Characterization Component

Understanding the inherent characteristics of each unit operation is the primary focus of the Process Characterization Component. It involves understanding and quantifying the behaviour of a bioprocess under various conditions to ensure consistent and reproducible production of bioproducts. This is usually done by defining critical process parameters (CPPs) and critical quality attributes (CQAs). CQAs are the attributes of the final product that determine its quality, safety, and efficacy. CPPs are process parameters that have a significant impact on CQAs. The effect of CPPs on CQAs is commonly done via a procedure called the design of experiments (DoE). This involves systematically running experiments each with varying parameters. This process can be expensive, so, the digital twin facilitates DoE by providing a detailed, real-time simulation that allows for in-depth analysis. Engineers and operators can explore the digital twin to gain insights into the underlying dynamics of the process, identifying the most likely CPPs, and specific ranges for each CPP. This involves utilizing models to characterize raw material behaviour, product formation kinetics, and scalability. For this component, it is beneficial to have data coming from various process conditions to capture the underlying dynamics of the physical system.

3.1.2 Process Optimisation Component

When the dynamics of the process are sufficiently understood in the characterization component, this knowledge can be used to improve the process. The Process Optimisation Component identifies opportunities for process improvement and efficiency enhancement. This involves leveraging the digital twin for scenario analysis, what-if simulations, and optimization strategies to improve overall process performance. Below are some example scenarios for process optimization.

- Scenario 1: Raw Material Variability. Simulation of variations in raw material quality provides insights into critical control points, allowing for the development of strategies to mitigate the effects of variability.
- Scenario 2: Equipment Downtime. Simulating scenarios with equipment failures allows the assessment of the impact on overall production timelines, aiding in the design of robust manufacturing schedules and maintenance plans.
- Scenario 3: Process Parameter Optimization. The digital twin supports the optimization of process parameters to maximize product yield and quality. Scenarios involving changes in feeding strategy, temperature, pressure, or flow rates can be simulated to identify optimal operating conditions.

Using process mining simulation engines, it is possible to simulate the above scenarios and confirm hypotheses in-silico before running costly wet lab experiments.

3.1.3 Process Monitoring Component

Continuously monitoring and assessing the ongoing manufacturing process is the role of the Process Monitoring Component. This includes employing sensor data and real-time analytics to detect deviations, anomalies, or potential issues during the execution of the physical process. In this component, the digital twin serves as a live dashboard, offering a comprehensive overview of the ongoing bioprocess. By continuously comparing the digital twin's behaviour with real-time sensor data, operators can quickly detect anomalies, deviations, or inefficiencies, allowing for proactive intervention before issues escalate. A wide range of process mining techniques can be applied in this setting. For example, conformance checking can be used to determine if the ongoing bioprocesses are progressing according to guidelines. Predictive process monitoring can also be used to predict the conditions of the bioprocess in the near future. Having access to prediction allows for proactive intervention in the process which is done in the control component described next.

3.1.4 Process Control Component

Process control in the context of bioprocessing refers to the systematic management and regulation of various parameters, conditions, and variables within a production process to ensure that it operates within desired specifications. It involves monitoring taking the results of the monitoring component and making adjustments in real-time to maintain optimal performance. The real-time synchronization between the digital twin and the physical process enables operators to make informed decisions promptly. With the ability to predict potential issues and simulate different control strategies, the

digital twin empowers operators to optimize process parameters, improve efficiency, and minimize downtime. Below the main aspects of the control components are described:

- **Real-Time Monitoring:** The process control component closely collaborates with the process monitoring component. The control component receives the results of the monitoring component and uses prescriptive process monitoring methodologies to determine actions based on the monitoring results.
- **Feedback Control Mechanisms:** The digital twin facilitates the implementation of feedback control mechanisms. By comparing real-time data with the expected or optimal conditions, the digital twin can automatically trigger adjustments to operating parameters. The digital twin also serves as a testing platform for various control strategies, before they are applied to the physical bioprocess.
- **Optimization Algorithms:** The optimization component discovers optimal parameters and conditions offline. This knowledge can be used by sophisticated algorithms embedded in the control component to optimize process parameters based on real-time data. This adaptive optimization ensures that the manufacturing process operates efficiently under varying conditions.
- **Integration with Control Systems:** The control component needs to seamlessly integrate with existing process control systems. This integration enhances the capabilities of traditional control systems by providing a more detailed and adaptive understanding of the manufacturing environment.

By continuously learning from data and adapting to changing conditions, the digital twin contributes to a culture of continuous improvement. Insights gained from the digital twin can inform adjustments to standard operating procedures, leading to enhanced efficiency and product quality.

3.2 Data and Action Units

The data and action units are the links between the physical and virtual systems. The digital twin is built using data collected from the real process and in turn, it provides input to the real process. All communication from the physical process to the digital twin goes through the data unit and from the digital twin to the physical system through the action unit.

3.2.1 The Data Unit

Traditionally data from bioprocesses has been collected manually by scientists. Typically a daily sample is taken and analyzed offline. However, in recent years, more efficient and frequent techniques for sampling have been introduced via sensors. These sensors monitor media conditions such as pH, temperature, and dissolved oxygen, as well as cellular processes and feed media addition. The main benefit of using sensors is that they collect process information in or near real-time, making it possible to sync the digital twin with the physical process constantly.

Another category of sensors, known as soft sensors, may be used to collect information from the bioprocess. One popular soft sensing technique that is successfully used in bioprocess industries is Raman spectroscopy [19, 31]. Raman can be used to

measure a variety of process parameters and variables, such as cell profiles, nutrients, and metabolites.

The data unit collates information from sensors and offline measurements and converts them to event logs. The event log can then be used to discover the process that serves as the digital twin. As more information arrives from the sensors, the event log is expanded and the digital twin is rebuilt to keep the it synced with the physical process.

3.2.2 The Action Unit

The action unit handles the information flow from the digital twin to the physical process. This information can be in the form of new experimental designs resulting from the characterization component of the digital twin, or the control strategy derived from the process control component. Some information is fed directly to the physical process, such as the parameter values coming from the control component. But other information such as the experimental design influences the physical process indirectly via human actors. The information communicated to the physical process directly must be transferred in real-time to ensure interventions are executed in a timely manner while also adhering to strict guidelines defined in the design of experiments stage.

3.3 Model Plug-ins

Here, we describe the relationship between prior knowledge about specific unit operations and the proposed digital twin. The digital twin should utilize various modelling techniques to represent the behaviour of the microorganisms and their interaction with the cell culture environment and materials. Three main paradigms of modelling are typically used in the literature: 1. Data-driven modelling, 2. Mechanistic Modelling, and 3. Hybrid Modelling.

In the upstream steps, machine learning algorithms analyze historical data on raw materials, fermentation, and cell culture. These data-driven models predict variables such as cell growth, nutrient consumption, and metabolite production based on historical patterns. Mechanistic models delve into the biological intricacies of fermentation and cell culture, capturing the underlying kinetics and dynamics of microbial growth, substrate utilization, and product formation. The hybrid modelling approach blends data-driven and mechanistic components. For instance, a machine learning model can be used to empirically estimate the parameter values of mechanistic equations based on data.

In the downstream phase, data-driven models come into play to predict product yield, purity, and quality attributes. These models integrate historical data on equipment performance, process conditions, and product characteristics. Complementing this, mechanistic models simulate the physical and chemical processes involved in downstream operations, considering mass transfer, chromatography column dynamics, and filtration kinetics.

These models, regardless of which paradigm is used, typically describe one unit operation and are only concerned about optimizing one step. However, such models do not provide information about how each unit operation affects the subsequent ones, and global optimization across all steps is still under-explored. The end-to-end digital

twin serves as an overall platform where the entire process can be studied with various modelling techniques. For instance, such models can be used to generate data under various conditions. This generated data can be the input to the digital twin to see the impact of each model on the process as a whole.

4 Challenges and Guidelines

Each unit operation in the process gives rise to specific challenges. There are also challenges associated with the end-to-end process. In this section, we provide an overview of some of these challenges and propose directions for future research to tackle them.

4.1 Building Process Models from Real-valued Data

Traditionally, BPM and process mining techniques have been developed to manage organizational workflows. Workflow models typically deal with discrete steps carried out by actors and involving one or many objects. As a result, most process mining techniques are equipped with ways of dealing with categorical sequences, and not continuous numeric values such as multidimensional time series. Bioprocess data, whether collected manually or via sensors, has a unique structure that workflow models alone are not sufficient to handle. Similar to a typical organization's workflow, a bioprocess involves activities carried out by actors and involves objects. However, these actions are fixed to comply with strict regulatory and safety requirements. Therefore, in an event log from a bioprocess manufacturing facility, there are few process variants and any suggested change to the workflow by the digital twin needs to go through a comprehensive and lengthy regulatory approval. In such a setting, conformance checking – one of the main operations in process mining – may be used to detect any deviations from the regulatory requirements.

One major challenge in bioprocess development is that such processes involve living microorganisms. These organisms' behaviour can be unpredictable and difficult to control. Also, determining the best conditions for their performance is non-trivial. The behaviour of these living cells is captured by data collected manually or from sensors installed in the equipment. Typical sensor measurements consist of process parameters which are fixed throughout a specific unit operation and process variables which are subject to change. These data are in the form of multidimensional time series taken at fixed time intervals depending on the sensor type. One major challenge when representing the bioprocess digitally using a BPM notion of process is that the sensor data needs to be converted into event log format since process models are formed using discrete activities, actors, or objects. This is a significant challenge because if not done correctly, important information may be lost in the process of conversion to event logs. One possible solution is to use unsupervised clustering algorithms to determine latent activities in the measurements, as done in [32].

4.2 Generating Control Signals

As mentioned in Section 3.1, the process control component of the digital twin is used to test control strategies before it is applied to the physical process. One concern is that the control strategy heavily relies on the quality of the model on which the digital twin is built. Low-quality data and models can lead to sub-optimal control decisions. So, a major challenge is to ensure that the model is of sufficient quality to be used in control decisions. To address this challenge, a rigorous quality assurance procedure must be implemented that ensures that the output of the digital twin meets the standards and requirements.

The other challenge of generating control signals relates to the previous challenge regarding the digital twin being discrete. Typically control signals specify fine-grained numeric values for controlled process parameters. However, a discretized digital twin can only recommend a range of values. One way to overcome this challenge is to combine it with a numerical model that can determine the exact value of the controlled parameter after the process model has determined the range.

4.3 Shifting from Batch Processing to Continuous Processing

Currently, the majority of bioprocesses are batch processes. This means a batch of material goes through the unit operations serially. Each batch has its unique identifier. In the context of process mining, each batch can correspond to a case. However, there has been a gradual shift towards continuous manufacturing due to its benefits of improving product quality and processing times [33]. For such continuous processes, one challenge is to identify the notion of case identifier as the material continually enters the processing reactor and its waste and unwanted material are continuously removed. This problem can be addressed via segmentation techniques in the robotic process mining literature.

4.4 Determining Simulation Fidelity

Another main challenge in building a digital twin of bioprocesses is to determine the fidelity of the models built into the digital twin. Simply, we need to determine how close the digital twin should be to the physical process. This depends on what the digital twin is being used for and on the unit operation being modelled. Some tasks require higher fidelity while for other tasks low-fidelity models might suffice.

To address this challenge, the first step is to review industry standards for each unit operation and for the overall end-to-end process. Also, error bounds must be established for each task and the digital twin's performance should be assessed against these bounds.

4.5 Incorporating Models of Specific Unit Operations

We might wish to use the knowledge gained from other modelling techniques that describe the bioprocess via the model plugins component of the framework. There is extensive literature on using data-driven, mechanistic, and hybrid techniques to model specific aspects of the process. While we envision the end-to-end bioprocess

as a process model, some specific tasks will benefit from using the models that have been developed in the literature. So a key challenge is to identify opportunities and frameworks for blending such existing models with the end-to-end process model.

4.6 Degree of Hybridization

High quality data availability is a major challenge in modelling bioprocesses. Due to the tight regulatory controls placed in bioprocess industries, the data coming from such facilities do not contain many variations. This makes creating what-if scenarios a challenging task. Users might wish to use the digital twin to analyze what will happen if something goes wrong. For example, how would the developed control strategy perform if there is a power outage? Or how should the actors proceed if one of the critical process parameters goes out of the established bounds? If the digital twin solely relies on data, it might not perform optimally unless a similar situation has occurred in the past. In such situations, it might be beneficial to incorporate process knowledge from mechanistic models. One way to achieve this is to generate data from such mechanistic models and discover the digital twin from the generated data. This is also not optimal since existing mechanistic models are too simple to model all aspects of the process. Therefore, the most obvious solution is to use a combination of real data and the generated data from mechanistic models. Nevertheless, the challenge remains, to what degree should these data be combined? Which process behaviours should be coming from real data and which from the generated data?

4.7 Scale-up

Much of the existing models and knowledge about bioprocesses come from lab experiments. Before a product is manufactured at a large scale, smaller lab and pilot-scale experiments are conducted and much of the existing data comes from such experiments. It is reasonable to assume that the envisioned digital twin will also be developed using small-scale data since manufacturing scale data is scarce. So another challenge that needs to be addressed is the scalability of the digital twin. Is a digital twin developed by data coming from a five-litre bioreactor going to accurately model a five thousand-litre reactor? Such questions should be answered using a pilot scale experiment for the digital twin. Once the safety of the digital twin on all other aspects is determined, it should be trialled in a large manufacturing facility before it can be deployed for regular use. Also, the mechanics of the larger environment should be incorporated into the process model to account for variations that might occur due to the larger equipment and material used in manufacturing.

4.8 Unit Operation-Specific Challenges

Each unit operation comes with its own specific challenges. Depending on the level of detail included in the digital twin, we might need to model individual unit operations. In the upstream phase, the microorganisms used in the bioprocess are grown in a culture and are encouraged to make a target product by tightly controlling the culture environment. The upstream is the step where the process heavily relies on the behaviour of the living microbes to achieve the desired outcome. So this step has a high level of

uncertainty and a complicated underlying behaviour which is difficult to capture by a model. On the other hand, the downstream steps rely less on living organisms and more on efficient use of equipment and understanding of fluid dynamics to successfully carry out the filtrations required. When designing the digital twin for the upstream steps, it is important to include uncertainty quantification techniques and investigate how the level of uncertainty can impact the downstream steps. For the downstream steps, the digital twin benefits from an object-centric approach where all the equipment and materials are central to the design of the process model and their impact is well established.

5 Case Study with a Pharmaceutical Company

In this section, we demonstrate the application of our envisioned digital twin framework via a case study with a large pharmaceutical company in Australia. This case study focuses on the monitoring component of the digital twin. In the first instance as a proof of concept, we focus on a single unit operation rather than the entire end-to-end process. However, the unit operation we have chosen is the production bioreactor, a very complex step in the bioprocess. In this case study, we show how monitoring of bioprocess parameters and quality attributes is possible via existing process mining solutions.

5.1 Soft Sensors and Bioprocess Monitoring

During the bioreactor phase, monitoring various process parameters and quality attributes is essential. This is because monitoring these parameters provides insights into cell activity and how the bioprocess unfolds. Typically, nutrients, metabolites, protein concentrations, and information about cell growth are monitored in real-time during bioreactor runs. Monitoring these parameters allows scientists to make crucial decisions on how to intervene in the process according to indications of cell behaviour. In particular, predictive monitoring is especially beneficial in this scenario because it allows for preemptive measures to be taken and for proactive interventions to be possible. For example, consider the metabolite lactate. It is a natural byproduct of cell activity when it is producing a target protein. Small amounts of lactate in the cell culture are normal. However, excessive levels of lactate acidify the cell culture medium leading to low overall performance [34]. So, lactate levels are constantly monitored during bioreactor runs, and if they start to go high, actions are taken to reduce them. Predictive monitoring of lactate concentrations allows such actions to be taken before lactate reaches undesired levels.

Traditionally, monitoring has been carried out by manually collecting samples from the bioreactor and analyzing the samples offline. However, process analytical technologies (PAT) tools are becoming increasingly common to measure monitored parameters during various stages of the bioprocess, including the bioreactor. One subset of such tools is spectroscopy. Various spectroscopy techniques have emerged in the bioprocessing domain, allowing for precise and real-time monitoring of process parameters. A comprehensive review of spectroscopy use cases in the bioprocess industries is presented in [19].

Raman spectroscopy is a PAT technique used in bioprocess development for various purposes. During the bioreactor phase, it is used for real-time monitoring of critical process parameters such as glucose, lactate, glutamine, glutamate, ammonia, and viable cell density (VCD) [35]. To achieve this, Raman probes are inserted into the bioreactor. These probes collect spectral measurements at specific time intervals (every fifteen minutes in our experiments). Spectral measurements need to be correlated to analytical measurements via statistical or machine learning models. These models are calibrated on manually collected measurements or data coming from other sensors. The calibration step is typically done offline and then the model is deployed for real-time monitoring. Once the calibrated model is applied to spectral measurements, the concentrations of the aforementioned parameters can be estimated in real-time. These estimates form a multivariate time series as they are concentration values collected at a fixed time interval.

5.2 Predictive Process Monitoring

The calibrated model provides current concentration values from spectral data. But it does not provide predictive capabilities. So, for predictive monitoring of critical process parameters, predictive techniques can be used to estimate future concentration values. Since the concentration values are time series, one natural way of predicting future values is time series forecasting. We can find various time series forecasting techniques applied in the bioprocessing domain, particularly for forecasting critical process parameters during the bioreactor phase [36, 37].

There are a few limitations in applying time series forecasting techniques for predictive monitoriong. First, these techniques do not incorporate the notion of a case. The input type assumed by time series forecasting methods does not account for the fact that measurements might come from independent experiments. For example, if we have glucose concentrations $< a_1, a_2, \ldots, a_t >$ for bioreactor A, and $< b_1, b_2, \ldots, b_t >$ for bioreactor B, and we wish to predict a_{t+1} and b_{t+1}, we cannot pool both sequences together as input to the forecast model. We need to train separate models for each bioreactor. This means that each model needs to be trained using data from the initial stages of the bioreactor and then applied to get predictions for the remainder of the bioreactor run. This is not optimal because we cannot use patterns and knowledge gained from previous bioreactor runs in the prediction task.

Second, the accuracy of the forecasts typically decreases as the prediction horizon increases. For instance, suppose that we train the model on glucose concentration values of the first two days. We can get highly accurate forecasts for day three. But the forecasts may be highly unreliable for day thirteen. This means that to get accurate results, models need to be updated at least every few days for each bioreactor run. This can be costly and inefficient.

In this case study, we propose using the notion of a process as it is defined in the Business Process Management and Process Mining communities for the predictive monitoring task. Given data collected using Raman spectroscopy, we aim to predict glucose and lactate concentrations in the next time step. To achieve this, we follow the pipeline outlined in Figure 4.

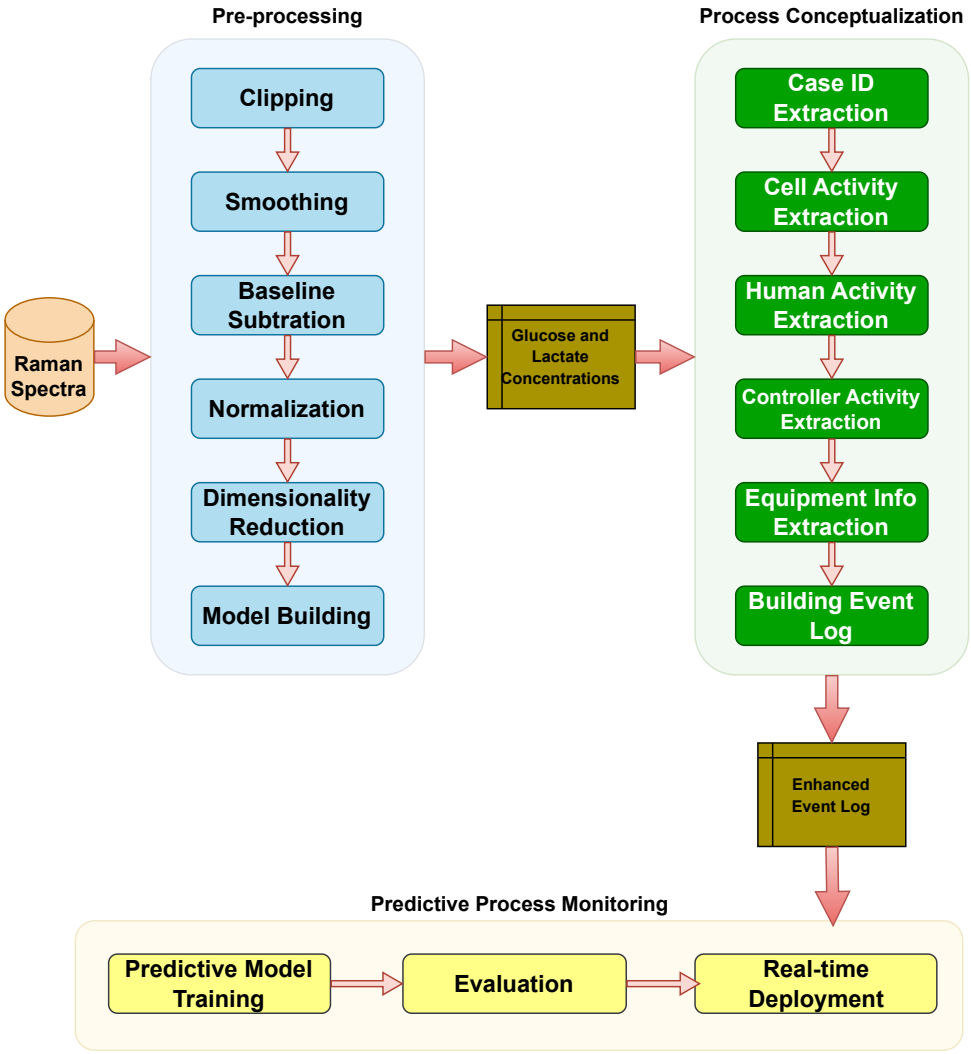


Fig. 4: Predictive Bioprocess Monitoring Pipeline: The activities shown in blue are the pre-processing steps described in Section 5.2.1, the activities shown in green are the process conceptualization steps described in Section 5.2.2, and the activities in yellow are the process monitoring steps of Section 5.2.3

5.2.1 Data Pre-processing

The raw Raman spectra do not include the substance concentrations. Several pre-processing steps need to be carried out to obtain concentration values. The first step is clipping. Clipping involves removing data points that are outliers or contain artifacts or other noise. This step helps to clean the spectrum in improve subsequent analyses. The second step is smoothing. Smoothing techniques such as moving average, Savitzky-Golay, or Gaussian filtering, are applied to reduce high-frequency noise while

preserving the spectral features. Smoothing improves signal-to-noise ratio (SNR) and enhances the visibility of peaks corresponding to molecular vibrations (which helps to distinguish concentrations of different substances in the cell culture). The third step, namely baseline subtraction, is performed to remove the background signal caused by fluorescence, Rayleigh scattering, or other non-resonant processes. There are various of methods for doing this such as polynomial fitting, asymmetric least squares, or adaptive iteratively reweighted penalized least squares (airPLS). Baseline subtraction reveals the true spectral features and facilitates accurate peak identification and quantification. Next, in the fourth step normalization is applied to adjust for intensity variations between spectra caused by factors such as sample concentration, instrument settings, or laser power fluctuations. Common normalization techniques include area normalization, maximum normalization, vector normalization, or probabilistic quotient normalization. But for Raman spectra in particular vector normalization is preferred. Normalization ensures that spectral comparisons are not biased by differences in overall intensity and improves the consistency of analyses across samples. It should be noted that normalization should be done after baseline correction [38]. The last and fifth step is dimensionality reduction. In this step, techniques such as principal component analysis (PCA) or partial least squares (PLS), are employed to reduce the number of variables while retaining most of the spectral information. PCA identifies principal components that capture the majority of variance in the data, enabling visualization and interpretation of spectral differences. PLS regression combines dimensionality reduction with regression analysis to model relationships between spectral data and sample properties, facilitating quantitative analysis and predictive modelling. PLS is often used to calibrate a model that maps spectral features to concentration values [35].

5.2.2 Process Conceptualization

Once concentration values are estimated from the Raman spectra, the next step is to conceptualize these readings as a process. Conventionally in the buainess process management and process mining communities, a process is a sequence of activities involving actions and objects to achieve an outcome. This definition cannot be directly applied to the bioprocess. Indeed, bioprocesses do involve actions performed by actors (e.g., scientists or automated controllers), but another important aspect of the bioprocess, especially during the bioreactor phase, is the *reaction* of the living microorganisms (cells) to the actors' actions. These cellular reactions are central to the bioprocess and determine its success or failure. However, they are considered latent processes as their behaviours and transitions are currently not fully understood. Consequently, these latent processes are characterized by measurements of various substances, such as nutrients, metabolites, amino acids, and cell data. For example, glucose is an essential nutrient for cell metabolism. By monitoring glucose levels, the rate of cell metabolism can be estimated. If glucose levels are going down rapidly, it indicates that the cells are metabolizing fast and using a lot of glucose.

To conceptualize the latent bioprocess, the various measurements collected manually or via sensors (e.g., Raman sensors) need to be incorporated into the conventional definition of a process. The conventional business process has activity labels, whereas we have time series measurements. Therefore, to create activity labels for the bioprocess,

we use clustering. The benefit of clustering is that it is unsupervised, so no prior knowledge is required to construct the labels. Also, clustering captures the natural shape of the data, so the discovered activities correspond directly to the collected measurements, and by extension, to the behaviour of the cells. Since timestamps already exist as part of the measurements, their translation into the conventional process is trivial.

Another important aspect of business processes is the case identifier. Each activity is paired with an identifier unique to the process instance it belongs to. In this case study, each bioreactor run is an independent instance of the bioprocess. Each run is characterized by the vessel used, the experiment it belongs to and the project that contains that experiment. Therefore in this setting, the case ID is a tuple of the following form: (project_ID, experiment_ID, vessel_ID).

Using the above-mentioned methods, we can construct event logs consisting of case IDs, activity labels, timestamps, and optional case attributes. Additional information can be added to enhance the event log. For example, information about the initial conditions of the bioreactor, or instruments settings can be added to the event log as case attributes. In this case study, we use Raman probe settings as case attributes.

5.2.3 Predictive Model Training

Once the bioprocess event log is created, the next step is training a predictive model to predict the next activity (cluster) given a prefix. Within the area of process mining, predictive process monitoring is experiencing a surge in activity, with numerous new techniques being proposed every year. These techniques use a variety of machine learning and deep learning methods for predicting various components of the process, with next activity prediction being one of the main focuses.

Predicting the next activity in this setting is not sufficient. This is because the predictive monitoring method only predicts the next cluster, but not the concentration value. In many cases, external actors of the bioprocess make decisions based on concentration values. Moreover, many control strategies are implemented with concentration values as input. Therefore, the predicted cluster needs to be converted to concentration values. To achieve this, we consider the predicted concentration to be the average value of the given substance in the predicted cluster:

$$Y_{s|c} := \frac{1}{n} \sum_{i=1}^{n} s_i \tag{1}$$

where $Y_{s|c}$ is the predicted concentration of substance s given cluster c, and s_i are concentration values of substance s in cluster c, and n is the number of data points in cluster c.

5.3 Experimental Setup

We apply the above methodology to a real-life dataset from a pharmaceutical company. The dataset consists of 35 bioreactor runs for the production of monoclonal antibodies. The bioreactor runs comprise different projects and different experiments within the same project. For each bioreactor, Raman spectra are available with a sampling

frequency of 15 minutes. Moreover, we use daily manually collected samples to calibrate the model mapping spectra to concentration values.

We performed the pre-processing steps using the following methods. Clipping is carried out to include spectra only within the range of 600-2000. Smoothing is done using a one-dimensional Gaussian filter and baseline correction with the airPLS method. After that, vector normalization is performed. To obtain concentration values, a PLS model is trained using spectral data as input and the daily concentration values of glucose and lactate as output.

Once glucose and lactate values are obtained at 15-minute intervals, we discover latent cellular activities via clustering. We use two common clustering methods to investigate the effect of clustering performance on the overall performance. The methods used are Kmeans and DBSCAN. We select the value of k for Kmeans through various strategies and compare the results. We use the elbow method and the Bayesian information criterion (BIC). In addition, we use the same number of clusters automatically discovered by DBSCAN.

We use a neural network to perform predictive monitoring. We use an embedding layer to represent the traces as a fixed-length vector of size 50. This vector is the input to an LSTM layer consisting of 100 units. It is then followed by a dense layer consisting of n units with n being the number of identified clusters. The softmax activation function is used to get probabilities for each cluster.

We use leave-one-out cross-validation (LOO-CV) to evaluate model performance. This means that for each training iteration, one of the bioreactors is reserved as the validation set, and the rest of the data are used in training. We report the average performance across all bioreactors. The clustering results are evaluated visually, as the input is two-dimensional (glucose and lactate).

We compare our results with time series forecasting since predictive monitoring is done via this approach in the bioprocessing literature [36]. For each bioreactor, a separate forecasting model is trained on the first two days of glucose and lactate concentrations, and forecasting is done on the subsequent days until the end of the bioreactor run.

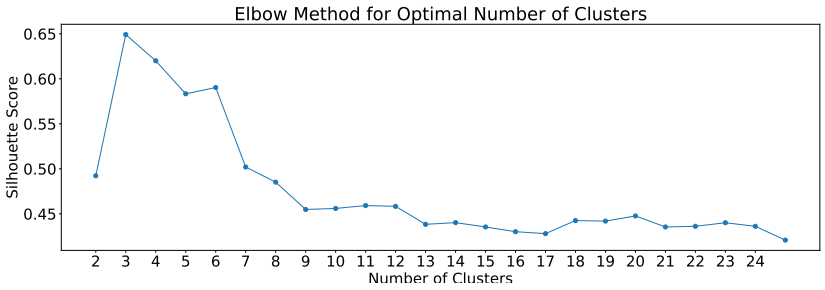
5.4 Results

To evaluate the performance of the PLS model, we calculate the mean squared error (MSE) and r2 score for the validation set and compare the results with a random regressor. Table 1 shows that the PLS model significantly outperforms the random classifier, and hence it can be used to obtain concentration values from the Raman spectra.

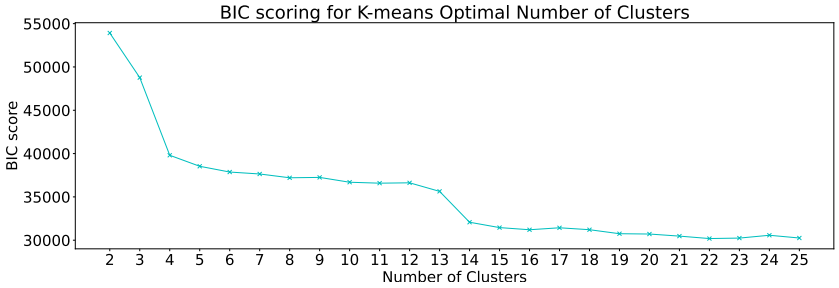
Figure 5 shows the results of the elbow method and the Bayesian information criterion for selecting the number of k for Kmeans. It can be seen that using the elbow method, the optimal number of clusters is 3, whereas the BIC shows that performance improvement diminishes after 14 clusters. 3 and 14 were picked as possible values for the

	MSE	R2 score
PLS	23.53	0.48
Random Regressor	48.1	-0.13

Table 1: Performance of the PLS model in terms of MSE and R2 score



(a)



(b)

Fig. 5: (a) The elbow method, showing the silhouette score for different numbers of clusters (higher is better). (b) The BIC score for different numbers of clusters (lower is better).

number of clusters in Kmeans. Also, using DBSCAN, the number of clusters was automatically selected as 8. For comparison purposes, we used 8 clusters for Kmeans as well. Figure 6 shows the clustering results using DBSCAN and Kmeans using $3, 8$, and 14 clusters.

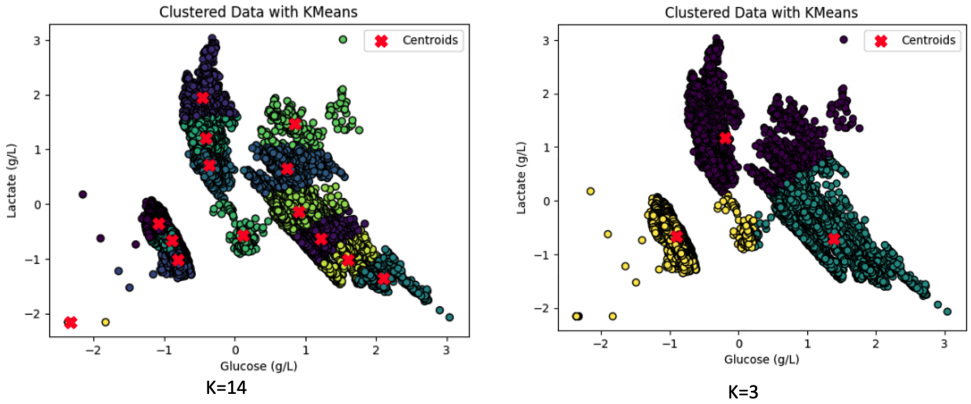
It can be seen visually that DBSCAN best captures the shape of the data and the clusters are well-separated. But in Kmeans the cluster boundaries seem arbitrary. Nevertheless, we proceed to build event logs using both methods and perform predictive monitoring. Table 2 shows the results of both predictive monitoring using our proposed approach (with various clustering configurations) and time series forecasting. The metrics used are mean absolute error (MAE), means squared error (MSE), and root mean squared error (RMSE).

5.5 Discussion

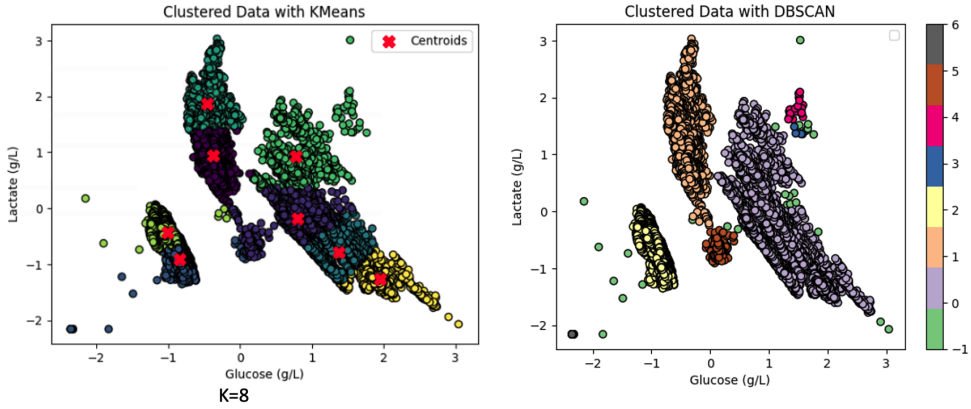
It can be seen from the results in Table 2 that overall the proposed approach outperforms the time series baseline for both glucose and lactate. It empirically confirms the hypothesis that using the notion of a process as part of prediction can improve results. For glucose, the best results in terms of MSE and RMSE are obtained using Kmeans with 3 clusters, even though clustering performance is below DBSCAN. This is because

	Number of Clusters	LSTM Accuracy	Glucose MAE	Glucose MSE	Glucose RMSE	Lactate MAE	Lactate MSE	Lactate RMSE
PPM with DBSCAN	8	0.78	4.33	29.92	5.46	**4.20**	**31.65**	**5.62**
PPM with Kmeans	8	0.86	3.04	16.71	4.08	6.62	45.93	6.77
PPM with Kmeans	3	**0.95**	3.05	**16.15**	**4.02**	5.84	51.03	7.14
PPM with Kmeans	14	0.79	**2.8**	16.38	4.04	5.0	44.49	6.67
Time series baseline	-	-	2.94	31.64	5.63	10.63	379.59	19.48

Table 2: Comparison of results using the proposed approach and time series forecasting.



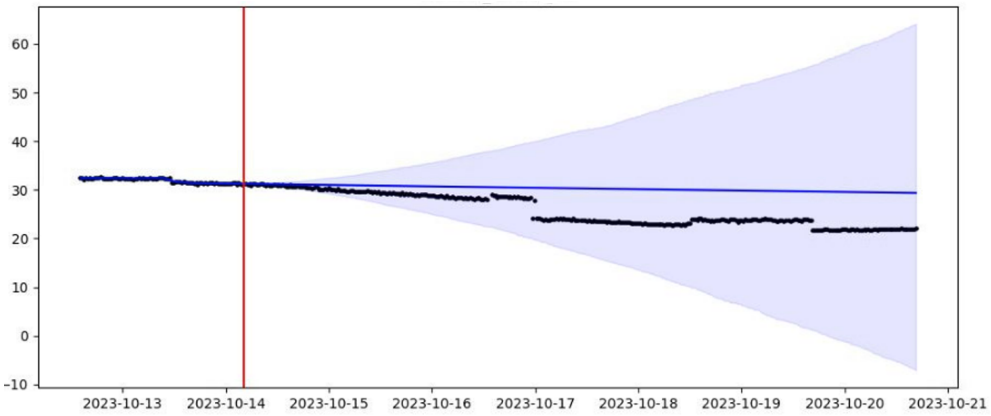
(a)



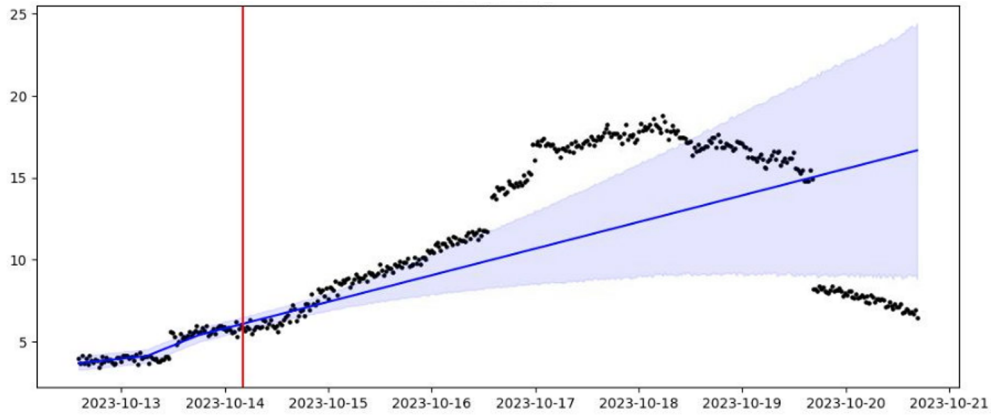
(b)

Fig. 6: (a) Clustering results using Kmeans with 3 and 14 clusters. (b) Clustering results using DBSCAN and Kmeans with 8 clusters.

the LSTM performance is significantly higher when using 3 clusters. As we have limited data available (35 bioreactors), more complex traces are harder to predict. In the case of glucose prediction, even though clustering quality is not optimal, it is sufficient to achieve superior performance to other methods because the correct cluster is identified the majority of the time. Also, it should be noted that glucose levels are relatively stable during the bioreactor run as shown in Figure 7a.



(a)



(b)

Fig. 7: (a) Example of glucose levels over time. (b) Example of lactate levels over time.

The best results for lactate in terms of MAE, MSE, and RMSE are obtained using our approach with DBSCAN. It can be seen that across all clustering configurations, the performance is significantly higher than the baseline. This can be explained by the fact that lactate levels fluctuate during the bioreactor run. So the time series model trained on the first two days of the run does not account for this non-stationarity

and so, the performance deteriorates rapidly and significantly as time progresses. This more complex pattern of lactate levels requires higher clustering quality, meaning that the clusters should more faithfully capture the natural shape of the data. This is why the best performance is observed using DBSCAN.

Overall, it can be seen that for both glucose and lactate, the performance improves over the baseline across all metrics regardless of the choice of clustering method. This demonstrates that incorporating the notion of process in the predictive monitoring task is crucial. Moreover, other aspects of the process could be incorporated into the event log to improve results. For example, we can add human activities to traces to account for sudden fluctuations in concentration levels. The sudden decrease in lactate shown in Figure 7b may be due to an intervention by a scientist or control system. Having this information in the trace is likely to improve the prediction results.

It should be noted that the quality of the predictive monitoring approach relies both on the prediction model and the choice of clustering method for more complex tasks. Since data in this domain is expensive to acquire, event logs can often be small. So one important area for future improvement is to design new methods that can work with small data.

6 Conclusion

In this book chapter we presented a novel digital twin framework for understanding, optimizing, monitoring, and controlling bioprocesses. Digital twins have seen a surge in attention in the business process management and bioprocess development communities. The definitions and functions of the digital twin are varied across disciplines. We discussed how to bridge the gap between these two related communities and present a framework to develop digital twins of bioprocesses using concepts and methods in the business process management and process mining fields. We discussed the challenges of achieving the digital twin and provided guidelines for those challenges. It should be noted that although the proposed framework is geared towards bioprocesses, it applies to other continuous processes that are characterized by regular measurements. Indeed any process with similar data types can benefit from this framework which could include other forms of manufacturing.

Through a case study conducted in collaboration with a prominent pharmaceutical company, we have demonstrated the practical applicability and tangible benefits of our proposed framework. The results of the case study showed that through an approach based on the proposed framework, we can achieve high prediction performance in the process monitoring task compared to a time series forecasting baseline.

Looking ahead, further research into the application of the proposed framework is warranted. In the future, we plan to address the challenges mentioned in the book chapter and offer concrete solutions to them.

Acknowledgments This research was supported under the Australian Research Council's Industrial Transformation Research Program (ITRP) funding scheme (project number IH210100051). The ARC Digital Bioprocess Development Hub is a collaboration between The University of Melbourne, University of Technology Sydney, RMIT

University, CSL Innovation Pty Ltd, Cytiva (Global Life Science Solutions Australia Pty Ltd) and Patheon Biologics Australia Pty Ltd.

References

[1] Grieves, M.: Digital Twin: Manufacturing Excellence through Virtual Factory Replication. A Whitepaper by Dr. Michael Grieves. 2014

[2] van der Aalst, W.M.P., Hinz, O., Weinhardt, C.: Resilient Digital Twins: Organizations Need to Prepare for the Unexpected. Business & Information Systems Engineering **63**(6), 615–619 (2021) https://doi.org/10.1007/s12599-021-00721-z . Accessed 2024-03-26

[3] Liebenberg, M., Jarke, M.: Information systems engineering with Digital Shadows: Concept and use cases in the Internet of Production. Information Systems **114**, 102182 (2023) https://doi.org/10.1016/j.is.2023.102182 . Accessed 2024-03-14

[4] van der Aalst, W.M.P.: Concurrency and Objects Matter! Disentangling the Fabric of Real Operational Processes to Create Digital Twins. In: Cerone, A., Ölveczky, P.C. (eds.) Theoretical Aspects of Computing – ICTAC 2021. Lecture Notes in Computer Science, pp. 3–17. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85315-0_1

[5] Aalst, W.M.P., Berti, A.: Discovering object-centric petri nets. Fundamenta Informaticae **175**(1-4), 1–40 (2020) https://doi.org/10.3233/FI-2020-1946

[6] van der Aalst, W.M.P.: Twin Transitions Powered By Event Data - Using Object-Centric Process Mining To Make Processes Digital and Sustainable. In: Joint Proceedings of the Workshop on Algorithms & Theories for the Analysis of Event Data and the International Workshop on Petri Nets for Twin Transition Co-located with the 44th International Conference on Application and Theory of Petri Nets and Concurrency (Petri Nets 2023). CEUR Workshop Proceedings, vol. 3424. CEUR-WS.org, Caparica, Portugal (2023). https://ceur-ws.org/Vol-3424/invited2.pdf

[7] Park, G., Van Der Aalst, W.M.P.: Realizing A Digital Twin of An Organization Using Action-oriented Process Mining. In: 2021 3rd International Conference on Process Mining (ICPM), pp. 104–111 (2021). https://doi.org/10.1109/ICPM53251.2021.9576846

[8] Park, G., Comuzzi, M., van der Aalst, W.M.P.: Analyzing Process-Aware Information System Updates Using Digital Twins of Organizations. In: Guizzardi, R., Ralyté, J., Franch, X. (eds.) Research Challenges in Information Science. Lecture Notes in Business Information Processing, pp. 159–176. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-05760-1_10

[9] Brockhoff, T., Heithoff, M., Koren, I., Michael, J., Pfeiffer, J., Rumpe, B., Uysal, M.S., Van Der Aalst, W.M.P., Wortmann, A.: Process Prediction with Digital Twins. In: 2021 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C), pp. 182–187 (2021). https://doi.org/10.1109/MODELS-C53483.2021.00032

[10] Bano, D., Michael, J., Rumpe, B., Varga, S., Weske, M.: Process-aware digital twin cockpit synthesis from event logs. Journal of Computer Languages **70**, 101121 (2022) https://doi.org/10.1016/j.cola.2022.101121 . Accessed 2023-07-31

[11] Becker, M.C., Pentland, B.T.: Digital Twin of an Organization: Are You Serious? In: Marrella, A., Weber, B. (eds.) Business Process Management Workshops. Lecture Notes in Business Information Processing, pp. 243–254. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-94343-1_19

[12] Hernández Rodríguez, T., Frahm, B.: Digital Seed Train Twins and Statistical Methods. In: Herwig, C., Pörtner, R., Möller, J. (eds.) Digital Twins: Tools and Concepts for Smart Biomanufacturing. Advances in Biochemical Engineering/Biotechnology, pp. 97–131. Springer, Cham (2021). https://doi.org/10.1007/10_2020_137 . https://doi.org/10.1007/10_2020_137 Accessed 2023-07-28

[13] Pham, T.D., Manapragada, C., Sun, Y., Bassett, R., Aickelin, U.: A scoping review of supervised learning modelling and data-driven optimisation in monoclonal antibody process development. Digital Chemical Engineering **7**, 100080 (2023) https://doi.org/10.1016/j.dche.2022.100080

[14] 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. Current Opinion in Chemical Engineering **33**, 100702 (2021) https://doi.org/10.1016/j.coche.2021.100702 . Accessed 2023-07-28

[15] Zhao, B., Li, X., Sun, W., Qian, J., Liu, J., Gao, M., Guan, X., Ma, Z., Li, J.: BioDT: An Integrated Digital-Twin-Based Framework for Intelligent Biomanufacturing. Processes **11**(4), 1213 (2023) https://doi.org/10.3390/pr11041213 . Accessed 2023-07-28

[16] Whitford, B.: Bioprocess intensification: aspirations and achievements. BioTechniques **69**(2), 84–87 (2020) https://doi.org/10.2144/btn-2020-0072 . Accessed 2023-07-31

[17] Appl, C., Baganz, F., Hass, V.C.: Development of a Digital Twin for Enzymatic Hydrolysis Processes. Processes **9**(10), 1734 (2021) https://doi.org/10.3390/pr9101734 . Accessed 2023-07-28

[18] Gomis-Fons, J., Schwarz, H., Zhang, L., Andersson, N., Nilsson, B., Castan, A., Solbrand, A., Stevenson, J., Chotteau, V.: Model-based design and control of

a small-scale integrated continuous end-to-end mAb platform. Biotechnology Progress **36**(4), 2995 (2020) https://doi.org/10.1002/btpr.2995 . Accessed 2024-03-14

[19] Khuat, T.T., Bassett, R., Otte, E., Grevis-James, A., Gabrys, B.: Applications of machine learning in antibody discovery, process development, manufacturing and formulation: Current trends, challenges, and opportunities. Computers & Chemical Engineering **182**, 108585 (2024) https://doi.org/10.1016/j.compchemeng.2024.108585 . Accessed 2024-03-14

[20] Tiwari, A., Masampally, V.S., Agarwal, A., Rathore, A.S.: Digital twin of a continuous chromatography process for mAb purification: Design and model-based control. Biotechnology and Bioengineering **120**(3), 748–766 (2023) https://doi.org/10.1002/bit.28307 . Accessed 2024-03-14

[21] Feidl, Garbellini, Luna, Vogg, Souquet, Broly, Morbidelli, Butté: Combining Mechanistic Modeling and Raman Spectroscopy for Monitoring Antibody Chromatographic Purification. Processes **7**(10), 683 (2019) https://doi.org/10.3390/pr7100683 . Accessed 2024-04-08

[22] Agarwal, H., Rathore, A.S., Hadpe, S.R., Alva, S.J.: Artificial neural network (ANN)-based prediction of depth filter loading capacity for filter sizing. Biotechnology Progress **32**(6), 1436–1443 (2016) https://doi.org/10.1002/btpr.2329 . Accessed 2024-04-08

[23] Pirrung, S.M., Van Der Wielen, L.A.M., Van Beckhoven, R.F.W.C., Van De Sandt, E.J.A.X., Eppink, M.H.M., Ottens, M.: Optimization of biopharmaceutical downstream processes supported by mechanistic models and artificial neural networks. Biotechnology Progress **33**(3), 696–707 (2017) https://doi.org/10.1002/btpr.2435 . Accessed 2024-04-08

[24] Liu, S., Papageorgiou, L.G.: Optimal Antibody Purification Strategies Using Data-Driven Models. Engineering **5**(6), 1077–1092 (2019) https://doi.org/10.1016/j.eng.2019.10.011 . Accessed 2024-04-08

[25] Taylor, C., Pretzner, B., Zahel, T., Herwig, C.: Architectural and Technological Improvements to Integrated Bioprocess Models towards Real-Time Applications. Bioengineering **9**(10), 534 (2022) https://doi.org/10.3390/bioengineering9100534 . Accessed 2023-07-28

[26] Bayer, B., Dalmau Diaz, R., Melcher, M., Striedner, G., Duerkop, M.: Digital Twin Application for Model-Based DoE to Rapidly Identify Ideal Process Conditions for Space-Time Yield Optimization. Processes **9**(7), 1109 (2021) https://doi.org/10.3390/pr9071109 . Accessed 2023-07-28

[27] Kuchemüller, K.B., Pörtner, R., Möller, J.: Digital Twins and Their Role in Model-Assisted Design of Experiments. In: Herwig, C., Pörtner, R., Möller, J.

(eds.) Digital Twins: Applications to the Design and Optimization of Bioprocesses. Advances in Biochemical Engineering/Biotechnology, pp. 29–61. Springer, Cham (2021). https://doi.org/10.1007/10_2020_136 . https://doi.org/10.1007/10_2020_136 Accessed 2023-07-28

[28] Sokolov, M., Stosch, M., Narayanan, H., Feidl, F., Butté, A.: Hybrid modeling — a key enabler towards realizing digital twins in biopharma? Current Opinion in Chemical Engineering **34**, 100715 (2021) https://doi.org/10.1016/j.coche.2021.100715 . Accessed 2023-07-28

[29] Mahanty, B.: Hybrid modeling in bioprocess dynamics: Structural variabilities, implementation strategies, and practical challenges. Biotechnology and Bioengineering **120**(8), 2072–2091 (2023) https://doi.org/10.1002/bit.28503 . Accessed 2023-07-31

[30] Narayanan, H., Luna, M., Sokolov, M., Butté, A., Morbidelli, M.: Hybrid Models Based on Machine Learning and an Increasing Degree of Process Knowledge: Application to Cell Culture Processes. Industrial & Engineering Chemistry Research **61**(25), 8658–8672 (2022) https://doi.org/10.1021/acs.iecr.1c04507 . Accessed 2024-03-28

[31] Gerzon, G., Sheng, Y., Kirkitadze, M.: Process Analytical Technologies – Advances in bioprocess integration and future perspectives. Journal of Pharmaceutical and Biomedical Analysis **207**, 114379 (2022) https://doi.org/10.1016/j.jpba.2021.114379 . Accessed 2023-07-28

[32] Su, Z., Yu, T., Lipovetzky, N., Mohammadi, A., Oetomo, D., Polyvyanyy, A., Sardiña, S., Tan, Y., Van Beest, N.: Data-Driven Goal Recognition in Transhumeral Prostheses Using Process Mining Techniques. In: 2023 5th International Conference on Process Mining (ICPM), pp. 25–32. IEEE, Rome, Italy (2023). https://doi.org/10.1109/ICPM60904.2023.10271945 . https://ieeexplore.ieee.org/document/10271945/ Accessed 2024-04-15

[33] Smiatek, J., Jung, A., Bluhmki, E.: Towards a Digital Bioprocess Replica: Computational Approaches in Biopharmaceutical Development and Manufacturing. Trends in Biotechnology **38**(10), 1141–1153 (2020) https://doi.org/10.1016/j.tibtech.2020.05.008 . Accessed 2023-07-28

[34] Bioprocess engineering: Basic concepts. Journal of Controlled Release **22**(3), 293 (1992) https://doi.org/10.1016/0168-3659(92)90106-2 . Accessed 2024-04-16

[35] Tulsyan, A., Khodabandehlou, H., Wang, T., Schorner, G., Coufal, M., Undey, C.: Spectroscopic models for real-time monitoring of cell culture processes using spatiotemporal just-in-time Gaussian processes. AIChE Journal **67**(5), 17210 (2021) https://doi.org/10.1002/aic.17210 . Accessed 2024-03-28

[36] Gan, J., Parulekar, S.J., Cinar, A.: Development of a recursive time series model

for fed-batch mammalian cell culture. Computers & Chemical Engineering **109**, 289–298 (2018)

[37] Alinaghi, M., Surowiec, I., Scholze, S., McCready, C., Zehe, C., Johansson, E., Trygg, J., Cloarec, O.: Hierarchical time-series analysis of dynamic bioprocess systems. Biotechnology Journal **17**(12), 2200237 (2022)

[38] Afseth, N.K., Segtnan, V.H., Wold, J.P.: Raman spectra of biological samples: A study of preprocessing methods. Applied spectroscopy **60**(12), 1358–1367 (2006)