Discovering Changes in Cell Stability Using Process Mining: A Case Study

Johnson Zhou*, Abel Armas-Cervantes*, Zahra Dasht Bozorgi*, Ellen Otte†, Artem Polyvyanyy*
*The University of Melbourne, Australia
{johnson.zhou1,zahra.dashtbozorgi,abel.armas,artem.polyvyanyy}@unimelb.edu.au
†CSL Limited, Australia
ellen.otte@csl.com.au

***Abstract*—A bioprocess is a series of biological, chemical, and physical operations used to produce a product using living cells or their components. Bioprocesses are often used for the production of monoclonal antibodies (mAbs). The first step of the mAb production bioprocess is to take a vial containing a small amount of the selected cell line and grow those cells until they are of sufficient quantity. This step is known as the seed train in bioprocess development. During the seed train phase, it is essential to monitor the stability of the cells and their growth due to challenges such as variations in cell behaviour, batch-to-batch differences, and potential changes in cultivation conditions. In this paper, we present a case study where process mining is used to analyse the stability of cell lines during the seed train phase at a large pharmaceutical company in Australia. In order to do so, first it was necessary to transform the collected seed train data into an event log. Next, process models were discovered for high- and low-growth seed trains. We then derived insights into the performance of the seed train growth rate whereby characteristics of cell cultures in early stages can be associated with growth rate performance in later stages. Finally, we showed how the discovered models can be used to predict the growth performance of new seed trains.**

***Index Terms*—bioprocess development, seed train performance insights, process mining**

I. Introduction

The biopharmaceutical industry relies on production processes that utilise living cells to produce therapeutic products. These processes are known as *bioprocesses* and are often used for the production of monoclonal antibodies (mAbs), a type of protein used to treat various diseases. The first step of the mAb production bioprocess is to take a vial containing a small amount of *cell line* culture – group of cells sharing common genetic heritage – from a frozen state and grow those cells until they are of sufficient quantity. This step is known as the *seed train* in bioprocess development [1].

Cell stability is the ability of the cells to maintain ideal growth rates over time, which has been associated with stable productivity and consistent yields during manufacturing. It is common to observe batch-to-batch variation in the stability of cell cultures despite consistent manufacturing methods and identical genetic make-up of the cells. The reasons for these variations are not fully understood [2]. Manufacturing methods manage variation through robust standard operating procedures (SOPs). The SOPs allow laboratory technicians to detect and *react* to changes in cell stability to mitigate negative outcomes. While effective, delayed reacting to poorly stable cell cultures increases costs and time burden. There are no data-driven methods yet available to define a decision surface distinguishing low from high cell stability at the early stages of seed train. Finding such a decision surface can contribute to quicker, more effective selection of stable cell lines, which can reduce production costs and time.

Typically, data about cell growth and age is collected every few days during the seed train step, and the decision on how to react is left to the laboratory technician. This approach is not standardised and can be subjective, leading to difficulties in the replication of the results in future batches. However, the collected data contains valuable insights about trends that can be used to make standardised decisions for optimal growth. Two main challenges exist in applying data-driven methodologies to seed trains. The first is that the amount of existing data is small due to the high cost of running wet-lab experiments. The second is that the developed methodology must be interpretable to pass strict regulatory requirements. These challenges serve as motivation to use process mining to monitor cell stability in seed trains.

Process mining methods are shown to require fewer samples than other data-driven techniques (e.g., complex machine learning or time series techniques) [3]. Also, using process discovery techniques, visual representations of the seed trains can be generated to aid with interpretability. Consequently, in this work, we address the following two research questions:

- **Research question 1 (RQ1).** Can process mining techniques be used for predicting cell stability in seed trains?
- **Research question 2 (RQ2).** How can a decision surface be created to detect poorly stable seed train experiments?

First, we conceptualise the seed train as a process, as studied in process mining [4]. We refer to this conceptualisation as the *bioprocess paradigm*, which enables the definition of *latent cell actions* grounded in cell action attributes. We hypothesise that latent cell actions are representative of the hidden intents of cells that enable derivation of the insights on the variations in cell stability. Second, we propose a method using the bioprocess paradigm to model and create a decision surface to help identify poorly performing cell cultures at an early stage. To the best of our knowledge, this is the first time process mining is used to analyse seed train data.

Postprint, August 2024

The remainder of the paper is organised as follows. The next section presents the background of this work. Section III presents the bioprocess paradigm. Section IV describes the details of data acquisition and preparation. Results of the experiments are presented in Section V. Finally, conclusions and future work are discussed in Section VI.

II. Background

Process mining is a family of tools and techniques to analyse business processes based on event logs. These event logs are collected by information systems supporting the execution of such business processes [4], [5]. An event log contains process executions captured as sequences of events. These sequences are referred to as traces, and all events in the same trace are part of the same case (same process execution). In a trace, events represent action instances ordered by their times of occurrence.

Given a set of actions A, an *event* is a tuple $e = \langle a, c, t, M \rangle$, where $a \in A$ is an *action* that triggered the event, c is a *case identifier* that refers to the process execution that triggered the event, t is a *timestamp* at which the event was observed, and $M \in \mathcal{M}$ is a collection of $m \geq 0$ number of *attributes* (d_i) and their *values* (v_i), that is, $M = (d_1, v_1), \ldots, (d_m, v_m)$, where $i \in [1 \,..\, m]$ and $\mathcal{M}$ is the universe of all attribute-value collections. A *trace* is then a non-empty sequence of all events that refer to the same case identifier ordered by their timestamps, and an *event log* is a collection of traces.

Process models are graphical representations of processes. In process mining, process discovery algorithms generate models from event logs. Some algorithms use directed graphs – *a.k.a.* Directly-Follows Graphs (DFG) [4] – in which nodes represent process actions and edges represent the direct precedence between action instances observed next to each other in, at least, one trace. Other algorithms use Petri nets (PNs), a graphical and mathematical modelling tool widely used for process analysis [6]. PNs have places, transitions and arcs connecting places to transitions and vice-versa. In a PN, these components are graphically represented as circles, boxes and arrows, respectively.

Conformance checking is a popular process mining operation to evaluate whether a particular event log can be explained by a given process model. An *alignment* is a conformance checking concept that describes the "best match" between a trace and possible executions of a process model. One can use *alignment fitness* [7] grounded in alignments to quantify how well a process model describes an event log with a value between 0 and 1, where larger values denote the quality of the model to describe more of the log traces.

III. Bioprocess paradigm

In contrast to most data-driven industries, bioprocesses remain largely understudied [2]. This is characterised by a general shortage of models and techniques representing how bioprocesses are executed and an incomplete understanding of the factors that contribute to process control and production yield. Of most relevance are recent works related to the modelling and optimisation of both the seed train and bioreactor processes using statistical methods [1], [8]–[10].

Conventional process paradigm (Figure 1A) typically considers an *actor* interacting with one or more *processes* in the form of *actions*, which produce observable outcomes. Outcomes are usually based on an arbitrary set of performance metrics defined in the context of an overall objective or goal. The results can act as feedback mechanisms (*action-reaction* loops), which allow the actor to make suitable adjustments for the process to progress [11]. It is implied that actors are predominantly human, either as individuals (being customers, sellers) or organisations (being suppliers, contractors) [4], [5]. *Human actors* are internal to the system and can directly interact with the processes to achieve results. Processes are predominantly deterministic and have directly measurable parameters, either quantitatively or qualitatively.

The difficulty for conventional process paradigms to capture the dynamics of bioprocesses is that the primary vehicle for production are living cells, which are complex and perpetual biological systems of their own accord. Cells are latent biological actors (*cell actors*) that harbour their own *latent cellular actions*. These cellular actors and processes are latent because their actions and changes are challenging to measure parametrically and deterministically [12].

This creates several challenges to modelling bioprocesses. First, human actors are external to the bioprocess system. Second, the results observed through bioprocess interactions are a culmination of both human and cell actions. The action-reaction cycle is, therefore, no longer under the direct influence of human actions, as shown in Figure 1B.

We define the latent cellular processes by considering their role within the extended bioprocess action-reaction cycle. The intuition is that although it may not be possible to categorically define exactly what a cell's actions are at any given time, it is possible to observe the effects of their actions in light of known observations. Formally, we define the set of latent cellular actions $A_C \subseteq A$ as a quantized representation of observations in the form of attributes relevant to cellular function $\mathcal{M}_C \subseteq \mathcal{M}$, referred to as *cell action attributes* (Figure 1C), through a quantization function Q, such that:

$$Q : \mathcal{M}_C \rightarrow A_C \,. \tag{1}$$

Concretely, we implement function Q using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [13]. Density Density-based techniques are advantageous when domain knowledge is limited as the number of clusters does not need to be pre-defined [14]. Additionally, the technique is better suited to handle data relationships that are not linearly separable. To overcome the limitation of multiple densities, a multi-density approach was used similar to the Multi-Level DBSCAN (ML-DBSCAN) algorithm proposed by Liu et al. [15]. In this approach, data is iteratively and incrementally clustered, starting with the highest density. Details of the implementation, as well as

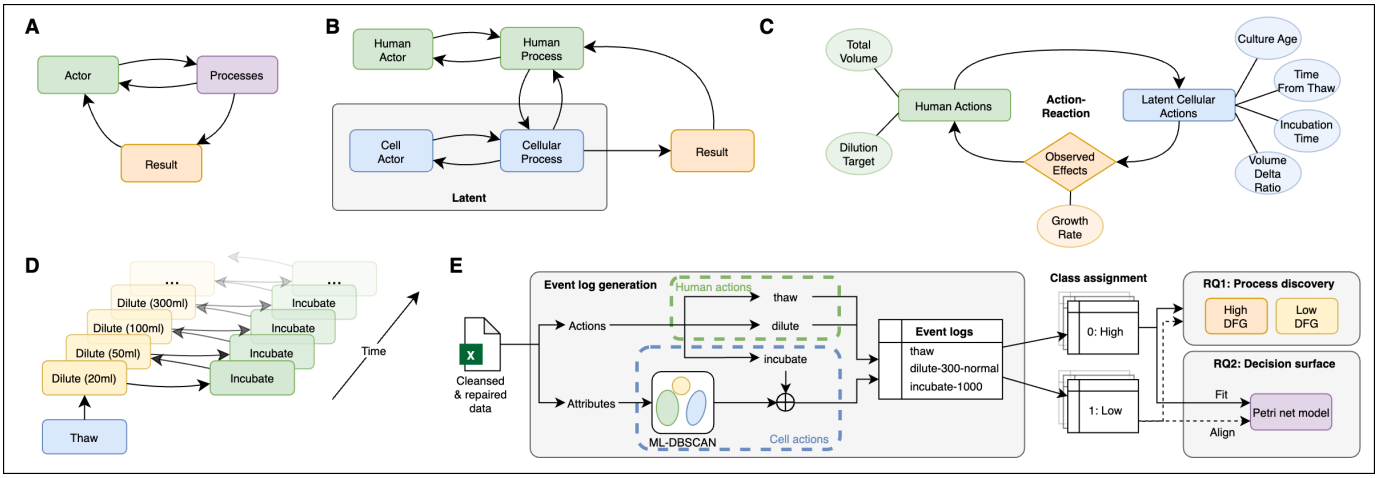

Figure 1: Schematic visualisations showing relationships and direction of influence between components for **A:** conventional process paradigm, **B:** bioprocess paradigm, **C:** contributors of cellular actions and associated attribute fields, **D:** seed train bioprocess, and **E:** data processing pipeline used for RQ1 and RQ2 experiments.

the choice for cell action attributes using real-world data, are discussed in the next section.

IV. Data Preparation

A. Dataset

We used a dataset comprising historical data from real-world seed trains. It contains timestamped actions taken by laboratory technicians when conducting seed train experiments. Some experiments were branched into several divergent experiments, each resulting in a separate trace in the dataset. A de-identified version of the dataset that comprises 94 traces over 2,154 events we used in our study is publicly available [16].

Cell line refers to the type of cell culture used in the seed train experiments with specific cellular properties. Two cell lines, namely *A* and *B*, represent the majority of traces within the available data. The actual names of the cell lines are obfuscated in consideration of commercial confidentiality. Cell lines may be genetically different, so findings in one cell line may not be fully transferable to another.

Seed train processes comprise three actions. *Thaw* is the initiation of a trace whereby a cell culture is thawed from a frozen state. *Dilute* refers to an action whereby a small inoculate of culture medium is diluted to a larger volume (known as total volume) by adding fresh medium. There is no material difference between culture and fresh medium. Finally, *incubate* refers to incubating diluted culture medium in controlled conditions to proliferate the growth of the cells.

Dilution typically results in a reduction of cell density within the total volume. This density measure is known as the *viable cell density* (VCD). The SOPs define the level of volume expansion by asserting a *target volume* and *dilution target*, which is the VCD after dilution. For example, a typical dilution can have a target volume of 300mL and a dilution target of 0.3×10^6 cells mL^{-1}. Incidentally, we refer to this specific dilution as the *steady state*. Dilutions are the avenue by which laboratory technicians can react to variations in cell stability by reducing the target volume, dilution target, or both. Cell growth is measured by *growth rate*, which is the rate of change in VCD over the incubation period, typically 72 hours.

Seed train experiments typically begin with the thaw action for the first 24 hours, followed by the dilute and incubate actions performed in an oscillation pattern (Figure 1D). The reactionary nature of dilutions means that there are no directly measurable negative outcomes. To address this, we use *time to steady state* as a proxy measure of cell stability. Specifically, we apply a label of "0-high stability" to experiments that reach steady state dilution by day 4 (referred to as "high") and "1-low stability" otherwise (referred to as "low"). We derive the number of days from SOPs with the help of a domain expert.

For each timestamp, the data includes a total of 11 attributes related to time, quantity, volume, and performance, see Table I. No cell biochemistry data was available. We performed basic data repair with the help of a domain expert and normalised the attribute values by removing the mean and scaling according to the 25th and 75th percentile range.

B. Event log creation

Human actions. Dilutions are human actions, and they are differentiated by both the target volume and dilution target. We represent a dilution action as a tuple $\langle \text{dilute}, \text{total volume}, \text{dilution target category} \rangle$. A typical dilution target may be 0.3×10^6 cells mL^{-1}. To manage granularity, we place dilution targets into categories of normal (0.3×10^6 cells mL^{-1}), low ($< 0.3 \times 10^6$ cells mL^{-1}) and high ($> 0.3 \times 10^6$ cells mL^{-1}). This is guided by the requirements as stipulated by the SOPs. For example, a dilute action with a target of 0.3×10^6 cells mL^{-1} at a total volume of 300mL would be written as "*dilution-300-normal*" on the event log.

Cell actions. Incubation is a cell action, as it requires no specific input from the technician due to standardised and tightly controlled incubation conditions. We refer to the actions of cells as *latent cell actions* as we do not make any

Table I: Event attributes; **(*)** denotes an attribute associated with cellular process identified as a cell action attribute.

Attribute	Domain	Description
Working day	Time	Number of days since thaw.
Culture age*	Time	Age of the cell culture.
Time from thaw*	Time	Cum. hours since thaw.
Incubation time*	Time	Time spent in incubation.
Viability	Quantity	The proportion of measured cells are viable.
Total cell density (TCD)	Quantity	Measured cell density.
Viable cell density (VCD)	Quantity	Cell density that is viable.
Dilution target*	Quantity	Target VCD post-dilution.
Growth rate*	Performance	Rate of growth in VCD from the previous record.
Inoculation volume	Volume	Volume of pre-dilution culture used to seed a dilution.
Fresh medium	Volume	Volume of fresh medium used to top up the dilution.
Total volume*	Volume	Total volume achieved by dilution.
Volume delta ratio*	Volume	Derived change in total volume from previous event.

assumptions regarding the internal process of the cells during incubation. Figure 1C and Table I show seven attributes in the available data that characterise incubation. We refer to these attributes as *cell action attributes*. We quantise the cell action attributes by applying ML-DBSCAN and use the cluster identity to differentiate cell actions. For example, if an event contains cell action attributes that have been labelled with cluster "1001", then the action for that event would be written as "incubate-1001". For ML-DBSCAN, we use the values of $[2.4, 6.6, 26]$ for epsilon and four minimum samples as determined empirically from k-distance investigations. Table II shows excerpts from the constructed event logs. We refer to an event log showing human-only actions as a "human" event log and that of human and cell actions as a "cell" event log.

V. Experiments

A. RQ1: Discovering variations in cell stability

We divide traces into two cohorts based on *time to steady state* as per Section IV-A, then conduct automated process discovery by transforming the event logs into DFGs (Figure 1E). We use traces from cell event logs of both cell lines to maximise information discovery.

Process discovery. We construct separate DFGs for each of the high and low stability class traces using Disco[1], a popular process mining tool with capabilities such as automated process discovery and data filtering. Using these DFGs (Figure 2A & B), we see that there are clear differences in the seed train process of high versus low stability cells, where low stability cells are process-wise more complex. These differences are discussed in detail in Section V-C1.

B. RQ2: Decision surface for identifying low stability

We answer RQ2 in three steps. First (RQ2.1), we model cell variants and contrast model conformance between the variant classes. Second (RQ2.2), we use a conformance measure, alignment fitness (herein called alignment), to predict whether a trace belongs to the low- or high-stability class. Third (RQ2.3), we use alignment to determine whether a decision surface could be created at incremental stages of the seed train.

We select traces as train and test sets through 5-fold cross-validation when investigating within similar cell lines and in a zero-shot fashion when investigating between cell lines. Cross-validation was performed to ensure generalisability in the prediction task. Of particular interest is how this approach can be transferred between cell lines due to the different cell genetics. Investigations are conducted in the following manner:

- **A-A (cell):** Train and test sets are drawn from only cell line A using the cell event logs.
- **A-B (cell):** Train set drawn from cell line A and test set drawn from cell line B using the cell event logs.
- **AB-AB (cell):** Train and test sets are drawn from both cell lines A and B using the cell event logs.
- **AB-AB (human):** Train and test sets are drawn from both cell lines A and B using the human event logs (ablative study).

RQ2.1: Contrastive model conformance. We create PN models using inductive miner [17] implemented within pm4py[2]. Noise threshold is a parameter in inductive miner ranging from 0.0 to 1.0, where 0.0 refers to no noise exclusions and 1.0 refers to all noise being excluded. This parameter aids in the exclusion of less significant events from the event log that may be regarded as noise, which can lead to simpler models. We perform a parameter search on the noise threshold in increments of 0.1 by fitting a new PN model at each threshold level from the train set and evaluating using alignments on the test set.

We obtain alignment scores between PN models that are fit on the train set and traces from the test set. The output is an alignment score between 0 and 1. A score of 1 means that the PN model can fully explain the traces in the test set. We hypothesise that we can derive a decision surface by contrasting the alignment scores between models fit on the same class and that of the opposite class. We achieve this by fitting a PN model on each of the low and high classes from a train set and checking alignment on a common test set. Specifically, we use the low class test set due to insights gained from RQ1, which will be discussed in the results section.

RQ2.2: Class prediction. Using results from RQ2.1, we design a prediction pipeline by fitting a PN model on the high class train set and predicting the class of test set traces. Test traces are given a class of 1 (low) if the alignment score is below a prediction threshold of 0.85 and 0 (high) otherwise. A noise threshold of 0.2 is used to fit the PN model. These values were empirically determined from RQ2.1 and are discussed in the results section.

RQ2.3: Incremental decision surface. RQ2.2 makes predictions when taking into account the entire duration of a test trace. This means making a prediction about the stability of the cell culture once the experiment has concluded. In this

[1]https://www.fluxicon.com/

[2]https://pm4py.fit.fraunhofer.de/

Table II: Excerpts from created event logs showing only human actions (left) and human and cell actions (right), with the identity of the "incubate" action being the key difference. Trace identifiers and attributes have been omitted for simplicity. Timestamps have been randomly increased between 0 and 99 years for de-identification purposes.

Human only ("human")			**Human and cell ("cell")**		
Datetime	**Action**	**Working day**	**Datetime**	**Action**	**Working day**
2104-10-09T11:35:00	thaw	0	2104-10-09T11:35:00	thaw	0
2104-10-09T11:35:00	dilution-30-normal	0	2104-10-09T11:35:00	dilution-30-normal	0
2104-10-09T11:40:00	incubate	0	2104-10-09T11:40:00	incubate-1000	0
...	...	...	...	...	...
2104-10-13T12:00:00	incubate	4	2104-10-13T12:00:00	incubate-1002	4
2104-10-16T09:30:00	dilute-300-normal	7	2104-10-16T09:30:00	dilution-300-normal	7
2104-10-16T09:50:00	incubate	7	2104-10-16T09:50:00	incubate-1001	7
...	...	...	...	...	...
2104-11-30T12:45:00	terminate	52	2104-11-30T12:45:00	terminate	52

section, we seek to determine whether our methods can create an incremental decision surface to see how early we can make a decision about the cell culture stability. We achieve this by fitting a PN model on the high class train set as in RQ2.2 (example shown in Figure 2C). For the test set, we truncate the event logs by the working days attribute and evaluate the alignment score. Specifically, we incrementally truncate from working day 1 to 55 in increments of 3 days in line with typical incubation periods of 72 hours. A noise threshold of 0.2 is used, consistent with RQ2.2.

C. Results and discussion

1) RQ1: Discovering variations in cell stability: Consider the DFGs for both the high stability and low stability processes presented in Figures 2A & 2B with several points of interest. First, actions "incubate-1001" and "dilute-300-normal" occur at high frequencies in both cohorts and exhibit a high degree of cyclic behaviour. Cluster "1001" represents cell incubation with a duration of 72 hours and at a volume of 300ml and matches the dilute action for total volume with a normal dilution target (see Figure 2D). This observation is consistent with that of the steady state identified through the SOPs.

Second, there is a clear and non-trivial difference between the seed train processes such that low stability cell cultures have a more complex control-flow as they progress through the seed train. This is especially the case in the period leading up to the steady state ("dilute-300-normal"). In contrast, the high growth cell cultures are not only simpler by way of control-flow but show a distinct lack of cycles prior to reaching the steady state. The observed complexity in the low-growth cohort could be interpreted to be the workings of the reactive procedures by human technicians to stimulate higher cell growth. For example, cell attributes of cluster "1011" (see Figure 2E), which is present in the low growth DFG (Figure 2B), show much higher dilution targets and lower target volume when compared to steady state (Figure 2D).

Consequently, this suggests that we could expect a reduction in alignment fitness when conforming traces from a low stability class to a process model discovered with the high stability class but not in reverse. This is because process models discovered with the low stability class are more complex and therefore more permissive for alignment of less complex (high stability) traces. We use this insight to guide the experimental method of RQ2.

2) RQ2.1: Contrastive model conformance: We contrast alignment fitness scores between "low-low" (Figure 3A) and "high-low" (Figure 3B) across a range of noise thresholds. Here, low-low refers to low stability test traces aligned against a PN model fit on the low stability class, and high-low refers to low stability test traces aligned against a PN model fit on the high stability class. The reported scores are an average from 5-fold cross-validation.

Good alignment between traces of the same class indicates how well our PN model can explain test traces, where higher is better. It is clear that as the noise threshold increases, alignment scores decrease. In contrast, high-low alignment is indicative of the level at which we can reject a test trace as belonging to the high stability class. In this situation, lower is better. The objective is to define a window where the difference in alignment scores between low-low and high-low is the greatest. From the results, we deduce 0.2 for noise threshold and an alignment score of around 0.85, and use these values to study RQ2.2.

We hypothesise from the objective that the larger margins between low-low and high-low would be suggestive of predictive capacity. If we consider cell line investigations, we expect to see that the biggest margin is achieved when both the train and test sets belong to the same cell line (A-A). The margin is reduced when we combine cell lines (AB-AB, cell) or when applied to new cell lines (A-B). In support of the bioprocess paradigm, using event logs from only human actions results in the smallest margin (AB-AB, human).

3) RQ2.2: Class prediction: Prediction results are presented as prediction accuracy and in the form of a confusion matrix in Figure 3E-H for cell lines investigations A-A (cell), A-B (cell), AB-AB (cell) and AB-AB (human), respectively. The accuracy reported is an average from 5-fold cross-validation. The confusion matrix is normalised to the ground truth.

These results are consistent with our hypothesis from RQ2.1 which states that larger margins between low-low and high-low alignments should be indicative of predictive performance. Our results demonstrate that given a sufficiently robust PN model, as indicated by an appropriate choice of noise threshold

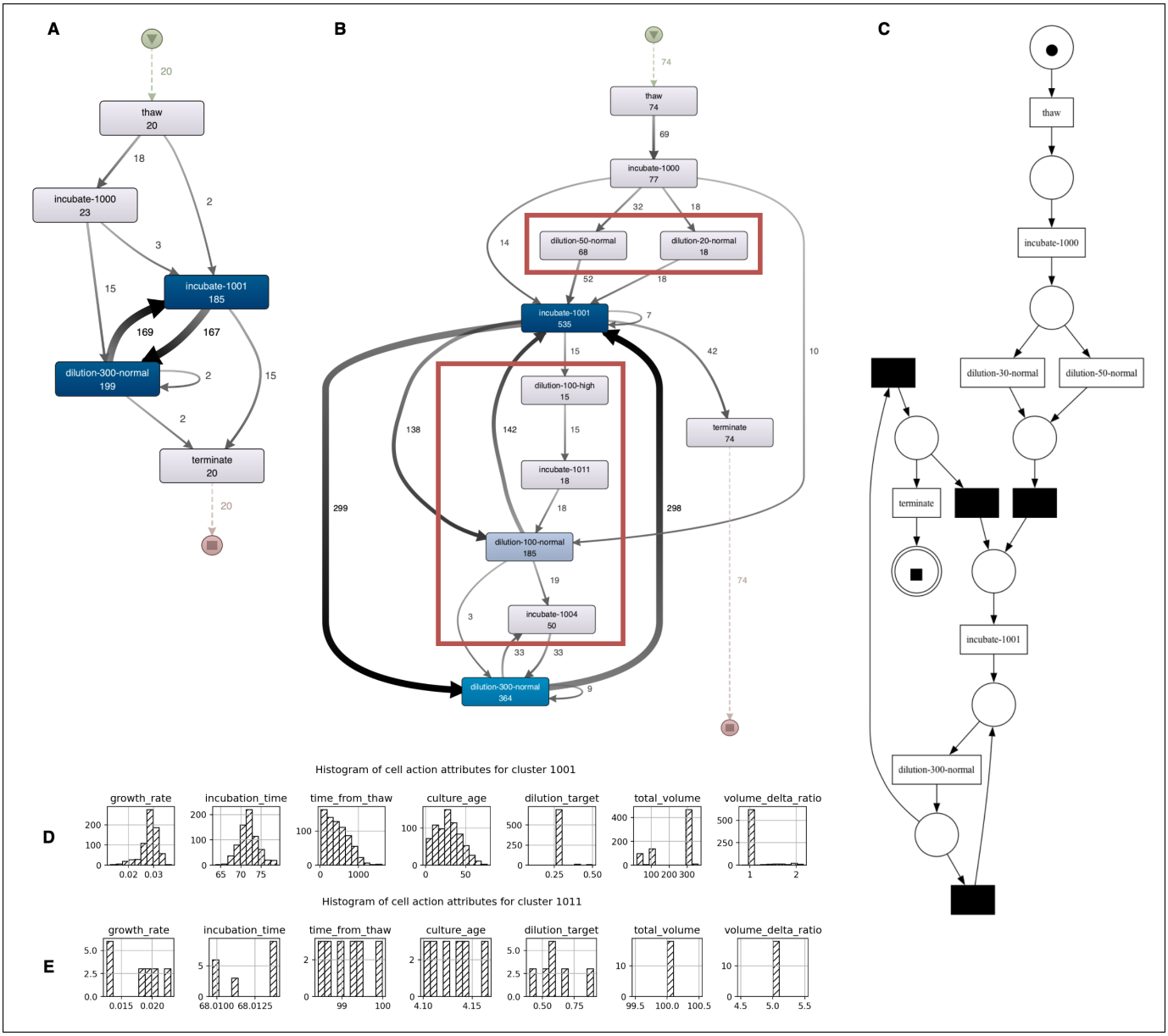

Figure 2: DFGs for high (**A**) and low (**B**) stability cells in the seed train. Red boxes highlight process differences in the low class not present in the high class. These DFGs were created with Disco for event logs for cell lines A and B combined. The level of detail is the same for both DFGs at 10% for both actions and paths. The green and red icons represent the beginning and end of the processes, respectively. **C:** Petri net (PN) fit on the high stability cohort of cell line A with a noise threshold of 0.2 using cell event logs. The filled circle and square within places mark the beginning and end of the process, respectively. Black transitions represent silent transitions, which are used to encode looping or skippable behaviour [18]. **D, E:** Distributions of cell action attributes for clusters "1001" and "1011," respectively. Cluster "1001" is common to both high and low stability cells in the steady state. Cluster "1011" is only present in low stability cells.

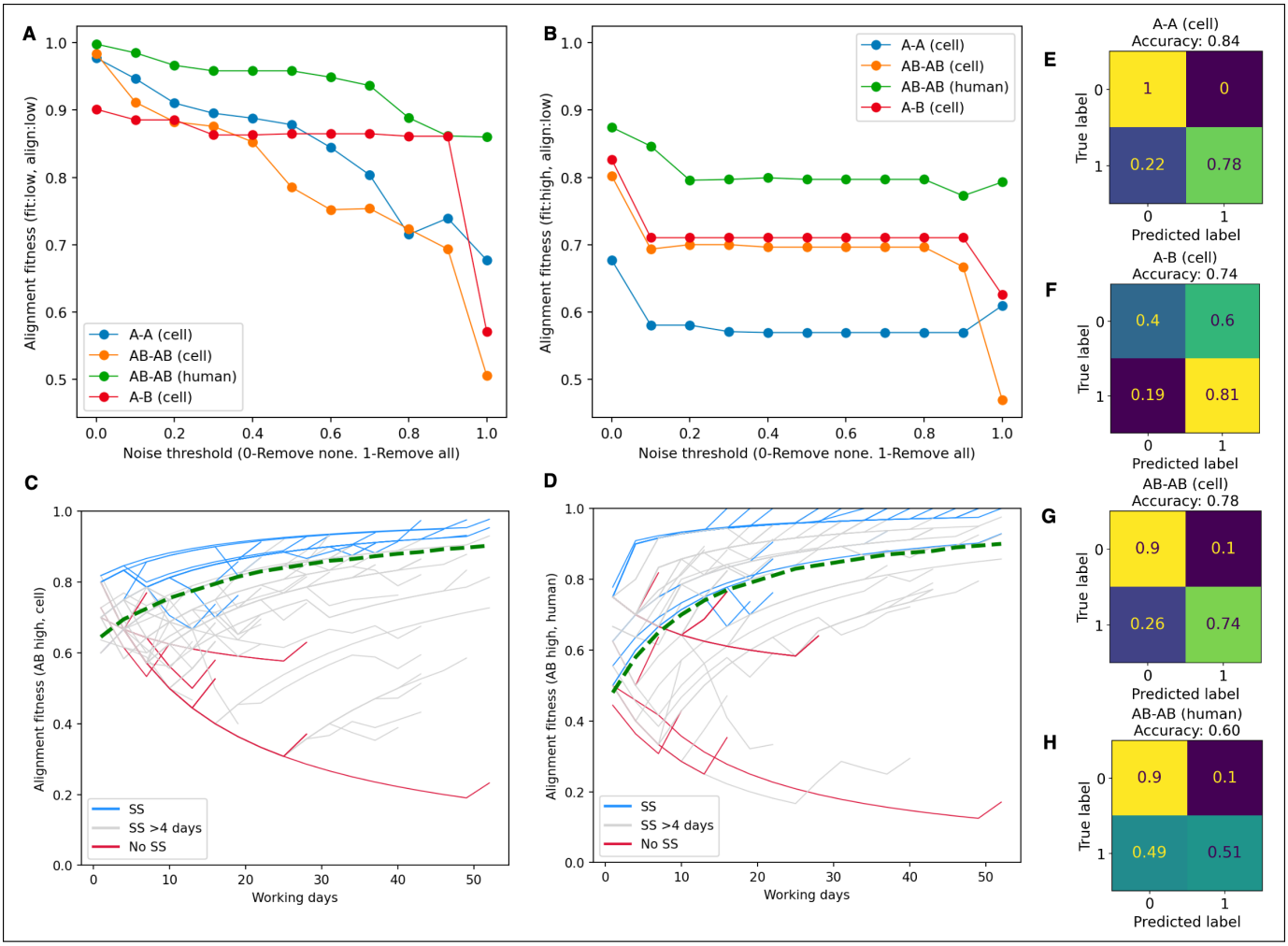

Figure 3: **A:** Average alignment fitness for a PN fit on the low class and tested with the low class across 5-fold cross-validation (low-low). **B:** Average alignment fitness for a PN fit on the high class and tested with the low class across 5-fold cross-validation (high-low). **C:** Alignment fitness for a PN fit on the high class and tested with traces truncated by working days using latent cell actions. Each line is a trace seen during 5-fold cross-validation. "SS" refers to a trace reaching the steady state by day 4 (class 0). "SS >4 days" refers to a trace reaching the steady state after day 4 (class 1) and "No SS" refers to a trace not reaching the steady state by termination (class 1). The green dotted line marks a possible decision surface at varying stages of the trace. **D:** Same as **C** except without using latent cell actions. **E-H:** Average prediction results across 5-fold cross-validation in the form of accuracy and confusion matrix when using a PN fit on the high class to predict the class label of a test set.

(0.2 in our case), predictive performance is only dependent on high-low alignment fitness, where lower is better. Within the same cell line (A-A), our approach can identify up to 80% of low stability cell cultures, referred to as positive predictions, with zero false positives. When applied to new cell lines (A-B), the rate of positive prediction is retained, however, at a higher rate of false positives (high stability predicted as low). This can be expected given the less favourable high-low alignment. The rate of false positives improves when the new cell line is incorporated into the train set (AB-AB, cell). In all cases, leveraging latent cell actions outperforms the same approach using only human actions (Figures 3G-H). These results confirm the practical benefits of complementing event data stemming from performed human actions with additional information on subsequent cell reactions. It is interesting to verify these insights in other domains beyond bioprocesses.

4) RQ2.3: Incremental decision surface: Figures 3C&D show the alignment scores of traces in cell lines A and B through progressive working days for cell event logs and human event logs, respectively. Every trace from cross-validation is represented as a line in the plot. We expand class 1 (low stability) to further differentiate traces that reach the steady state after 4 days (shown in gray) and traces that terminate prior to reaching the steady state (shown in red). This is to provide a further point of distinction with respect to the performance of cell cultures. We represent a possible decision surface by the green dotted line as being the lowest high-low alignment score over time that can separate the majority of

class 0 (high stability) traces.

Note that using our method (Figure 3C), we could establish a decision surface at alignment scores of around 0.75 from day 4 up to around 0.85 by day 50. The decision surface is well defined for the purposes of separating high and low stability traces according to our definition with only 3% of low stability traces above the decision surface. This distinction is not as well defined when using the same approach without consideration of latent cell actions (Figure 3D) with up to 49% low stability traces above the decision surface. The proportion of high stability traces above the decision surface is more than 80% in both approaches.

Of interest is the separable nature of low stability traces that reach the steady state late (grey) and with traces that do not reach the steady state (red). This suggests that there may be several decision surfaces, each suited for a specific application or requirement. For example, shifting the decision surface in a downward fashion may be more accepting of lower-performing cell cultures. The practical implications of this may be explored in a future study.

VI. Conclusion and future work

In this case study, we set out to discover changes in cell stability in the seed train phase of bioprocess manufacturing using process mining. Bioprocesses are different from conventional business processes in that the primary vehicle for production are cells, which are themselves latent actors in an extended action-reaction cycle.

In our first contribution, we apply a bioprocess paradigm that acknowledges the latent cellular actors through their observations. By augmenting the event log with latent cell actions, which are quantized representations of the observed cellular attributes, we discover the process characteristics that define cell cultures that vary in stability. We show that process models discovered using event logs augmented with latent cellular action outperform models derived conventionally based on the actions of human technicians alone. This shows that techniques developed in process mining have the potential to be applied in unexplored fields. However, the correct notion of process must be selected to obtain meaningful results.

Our second contribution presents an avenue for bioprocess manufacturers to identify low-performing cell cultures at an early stage of the seed train to minimise potentially lengthy and costly corrective processes. We show that process models discovered using our method can be transferred to new cell lines at a similar rate of detection with the risk of more false positive cases. This risk can be reduced by incorporating the new cell lines into the model training. This presents opportunities to improve product quality while reducing time to market for high-value biopharmaceuticals.

Considering the small amount of data available, in future work, it is interesting to explore whether process mining can consistently produce better results than other data-driven approaches for the prediction of cell stability.

Reproducibility. The de-identified dataset is available [16].

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] T. Hernández Rodríguez and B. Frahm, "Design, optimization, and adaptive control of cell culture seed trains," in *Animal Cell Biotechnology: Methods and Protocols*, ser. Methods Mol. Biol. Springer, 2020, vol. 2095, pp. 251–267.

[2] M. Sokolov, F. Feidl, M. Morbidelli, and A. Butte, "Big Data in Biopharmaceutical Process Development: Vice or Virtue?" *Chim. Oggi – Chem. Today*, vol. 36, pp. 26–29, 2018.

[3] Z. Su, T. Yu, N. Lipovetzky, A. Mohammadi, D. Oetomo, A. Polyvyanyy, S. Sardiña, Y. Tan, and N. van Beest, "Data-Driven Goal Recognition in Transhumeral Prostheses Using Process Mining Techniques," in *ICPM*, 2023, pp. 25–32.

[4] W. M. P. van der Aalst, *Process Mining: Data Science in Action*. Springer, 2016.

[5] M. Dumas, M. La Rosa, J. Mendling, and H. A. Reijers, *Fundamentals of Business Process Management*. Springer, 2018.

[6] T. Murata, "Petri Nets: Properties, Analysis and Applications," *Proc. IEEE*, vol. 77, no. 4, pp. 541–580, 1989.

[7] A. A. Adriansyah, "Aligning Observed and Modeled Behavior," *Technische Universiteit Eindhoven*, 2014.

[8] S. Kern, O. Platas-Barradas, R. Pörtner, and B. Frahm, "Model-Based Strategy for Cell Culture Seed Train Layout Verified at Lab Scale." *Cytotechnology*, vol. 68, pp. 1019–1032, 2016.

[9] T. Hernández Rodríguez, C. Posch, R. Pörtner, and B. Frahm, "Dynamic Parameter Estimation and Prediction Over Consecutive Scales, Based on Moving Horizon Estimation: Applied to an Industrial Cell Culture Seed Train." *Bioprocess Biosyst. Eng.*, vol. 44, 2021.

[10] T. Hernández Rodríguez, A. Sekulic, M. Lange-Hegermann, and B. Frahm, "Designing Robust Biotechnological Processes Regarding Variabilities Using Multi-Objective Optimization Applied to a Biopharmaceutical Seed Train Design," *Processes*, vol. 10, p. 883, 2022.

[11] J. J. Koorn, X. Lu, H. Leopold, and H. A. Reijers, "From Action to Response to Effect: Mining Statistical Relations in Work Processes," *Inf. Syst.*, vol. 109, p. 102035, 2022.

[12] T. Hernández Rodríguez and B. Frahm, "Digital Seed Train Twins and Statistical Methods," in *Digital Twins: Tools and Concepts for Smart Biomanufacturing*, 2021, pp. 97–131.

[13] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *ACM SIGKDD*, vol. 96, no. 34. AAAI Press, 1996, pp. 226–231.

[14] R. Bhuyan and S. Borah, "A survey of some density based clustering techniques," *CoRR*, vol. abs/2306.09256, 2023.

[15] S. Liu, S. Cao, M. Suarez, E. C. Goonetillek, and X. Huang, "Multi-Level DBSCAN: A Hierarchical Density-Based Clustering Method for Analyzing Molecular Dynamics Simulation Trajectories," *BioRxiv*, 2021.

[16] J. Zhou, A. Polyvyanyy, A. Armas-Cervantes, Z. D. Bozorgi, and E. Otte, "Data for Discovering Changes in Cell Stability Using Process Mining: A Case Study," https://doi.org/10.6084/m9.figshare.26701543.

[17] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, "Discovering Block-Structured Process Models from Event Logs – A Constructive Approach," in *Petri Nets*, ser. Lecture Notes in Computer Science, vol. 7927. Springer, 2013, pp. 311–329.

[18] L. Wen, J. Wang, W. M. P. van der Aalst, B. Huang, and J. Sun, "Mining Process Models With Prime Invisible Tasks," *Data Knowl. Eng.*, vol. 69, pp. 999–1021, 2010.