frontiers | Frontiers in Bioengineering and Biotechnology

TYPE Original Research
PUBLISHED 05 June 2023
DOI 10.3389/fbioe.2023.1160223

Next-generation cell line selection methodology leveraging data lakes, natural language generation and advanced data analytics

Stephen Goldrick[1]*, Haneen Alosert[1], Clare Lovelady[2], Nicholas J. Bond[3], Tarik Senussi[2], Diane Hatton[2], John Klein[4], Matthew Cheeks[2], Richard Turner[5], James Savery[4] and Suzanne S. Farid[1]

[1]Department of Biochemical Engineering, University College London, London, United Kingdom, [2]Cell Culture and Fermentation Science, Biopharmaceuticals Development, R&D, AstraZeneca, Cambridge, United Kingdom, [3]Analytical Sciences, Biopharmaceuticals Development, R&D, AstraZeneca, Cambridge, United Kingdom, [4]Data Science and Modelling, Biopharmaceuticals Development, R&D, AstraZeneca, Cambridge, United Kingdom, [5]Purification Process Sciences, Biopharmaceuticals Development, R&D, AstraZeneca, Cambridge, United Kingdom

Cell line development is an essential stage in biopharmaceutical development that often lies on the critical path. Failure to fully characterise the lead clone during initial screening can lead to lengthy project delays during scale-up, which can potentially compromise commercial manufacturing success. In this study, we propose a novel cell line development methodology, referenced as CLD_4, which involves four steps enabling autonomous data-driven selection of the lead clone. The first step involves the digitalisation of the process and storage of all available information within a structured data lake. The second step calculates a new metric referenced as the cell line manufacturability index (MI_{CL}) quantifying the performance of each clone by considering the selection criteria relevant to productivity, growth and product quality. The third step implements machine learning (ML) to identify any potential risks associated with process operation and relevant critical quality attributes (CQAs). The final step of CLD_4 takes into account the available metadata and summaries all relevant statistics generated in steps 1–3 in an automated report utilising a natural language generation (NLG) algorithm. The CLD_4 methodology was implemented to select the lead clone of a recombinant Chinese hamster ovary (CHO) cell line producing high levels of an antibody-peptide fusion with a known product quality issue related to end-point trisulfide bond (TSB) concentration. CLD_4 identified sub-optimal process conditions leading to increased levels of trisulfide bond that would not be identified through conventional cell line development methodologies. CLD_4 embodies the core principles of Industry 4.0 and demonstrates the benefits of increased digitalisation, data lake integration, predictive analytics and autonomous report generation to enable more informed decision making.

KEYWORDS

cell line development, machine learning, data analytics, natural language generation, Industry 4.0

Check for updates

OPEN ACCESS

EDITED BY
Meiyappan Lakshmanan,
Indian Institute of Technology Madras,
India

REVIEWED BY
Jong Kwang Hong,
Yonsei University, Republic of Korea
Seo-Young Park,
Sungkyunkwan University, Republic of Korea
Tran Bich Ngan Nguyen,
AbbVie Operations Singapore Pte Ltd,
Singapore

*CORRESPONDENCE
Stephen Goldrick,
✉ s.goldrick@ucl.ac.uk

RECEIVED 06 February 2023
ACCEPTED 22 May 2023
PUBLISHED 05 June 2023

CITATION
Goldrick S, Alosert H, Lovelady C, Bond NJ, Senussi T, Hatton D, Klein J, Cheeks M, Turner R, Savery J and Farid SS (2023), Next-generation cell line selection methodology leveraging data lakes, natural language generation and advanced data analytics.
Front. Bioeng. Biotechnol. 11:1160223.
doi: 10.3389/fbioe.2023.1160223

1 Introduction

Cell line development (CLD) is a critical task within biopharmaceutical manufacturing and selects the lead clone for the master cell bank (MCB) which provides the starting material for the entire life span of a therapeutic drug candidate. Traditional methods for CLD are time-consuming and require a large number of experiments to identify the best candidate, which can lead to delays in the development timeline and higher costs. This selection process typically aims to achieve two objectives. The first is to reduce a large heterogeneous pool of between a 1,000–10,000 cell lines to a single clone expressing high levels of a therapeutic protein that meet all relevant product quality specifications (Munro et al., 2017). The second is to ensure the selected lead clone will scale-up appropriately and consistently deliver the required product demand whilst achieving the desired product quality specifications. Both objectives are equally important; however, the majority of research focuses on the first objective of selecting a high producing clone, with scale-up considerations a secondary objective. This oversight is normally due to strict timelines and availability of data resulting in a decision on the lead clone selection before an in-depth evaluation of all scale-up and process considerations can be properly accessed. This paper proposes a more holistic methodology for CLD that better leverages the data generated during CLD to select the lead clone to satisfy productivity, quality and scalability objectives.

A summary of some CLD related challenges is depicted in Figure 1. The first challenge relates to the early stages of CLD where the focus is on high-throughput evaluation of thousands of clone candidates cultured using a low volume (<1 mL) and operated

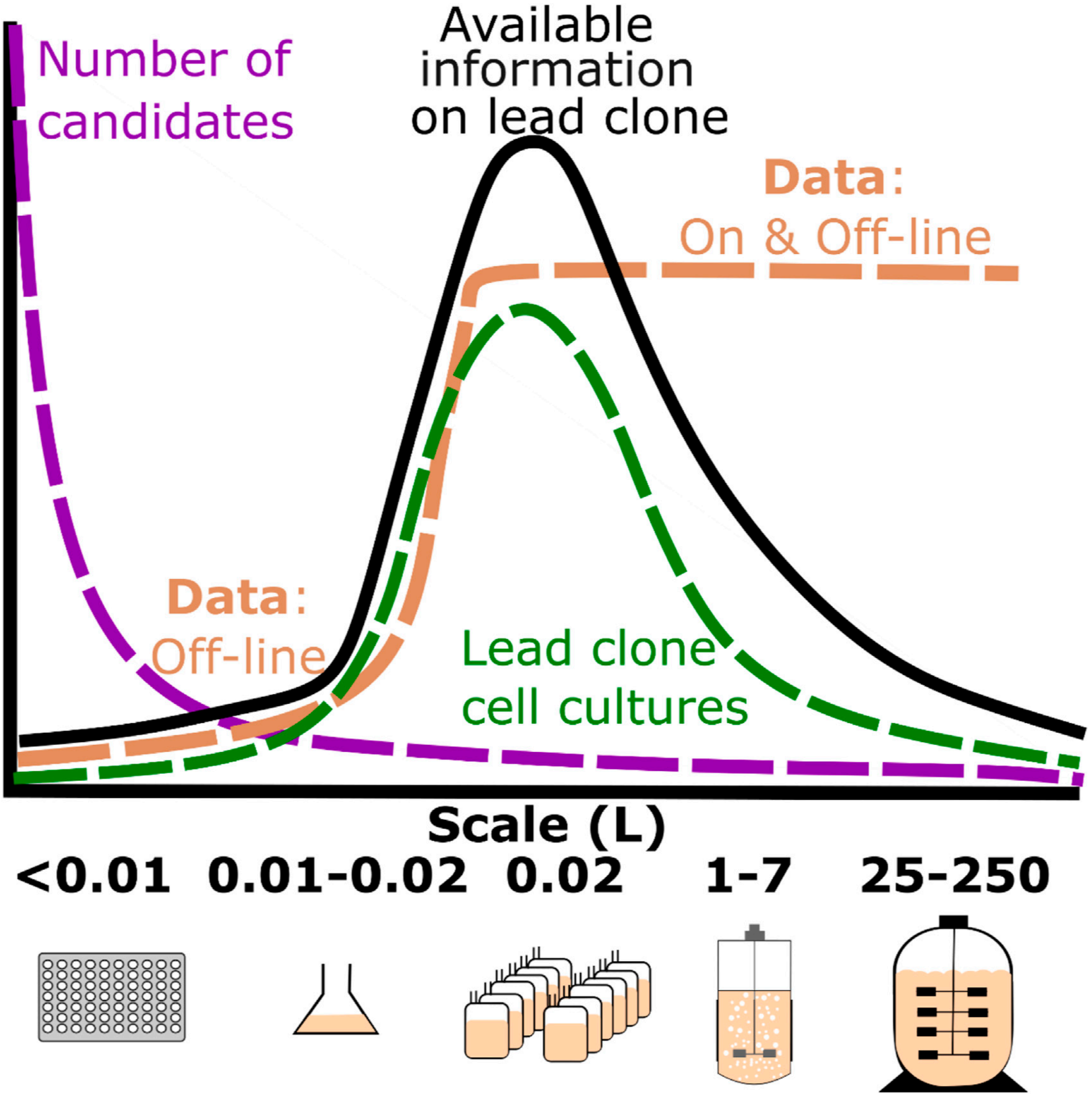


FIGURE 1
Schematic diagram summarizing the number of candidate clones, lead clone cell cultures, data recorded and available information on the lead clone for a typical therapeutic protein as it transitions from micro-scale (<0.1 L) to pilot scale manufacturing (250 L). The number of clone candidates is highlighted by the purple line, the number of lead clone culture runs is shown by the green line and data available at each scale by the brown line which is primarily defined by either off-line analytics or both off-line and online process data recorded at each scale. The available information trend line shown in black is a function of the number of lead clone cultures multiplied by the data available at the various scales.

in batch mode. The initial screening typically involves a high number of candidates assessed using high-throughput 96 well plates or microtitre plates (MTPs). Subsequent screening of hundreds of clones can be performed in high throughput fed-batch using multi-well plates in automated culturing systems (Rameez et al., 2014; Silk et al., 2010). The early stages of CLD follow a well-established procedure to appropriately select the cell lines that can move forward (Lin et al., 2019; Li et al., 2010; Hong et al., 2018). The cell lines are transfected with the gene of interest for the targeted product, methods include transfecting plasmids through calcium phosphate, cationic lipid-based lipofection, electroporation and using polymer-based reagents amongst others (Li et al., 2010). These high-throughput platforms screen using cell markers, typically MTX (methotrexate) or MSX (methionine sulfoximine) for glutamine synthetase (GS) mediated cells (Li et al., 2010; Lin et al., 2019). Each selected cell is then cloned into its own colony, as regulators require evidence that the cells are derived from a single progenitor cell to ensure monoclonality (European Medicines Agency, 2006; Chen et al., 2020). Methods such as single cell printing using the ClonePix system and Fluorescence-activated cell sorting (FACS) flow cytometry are methods implemented to ensure separate cell lines for their colonies (Kim, Diao, and Shuler, 2012). In order to identify a stable highly expressing cell line that meets the desired quality requirements a high number of candidates are required to be assessed. This involves various analytical assays such as ELISA to eliminate non-producing cells, followed by subsequent screening for highly productive clones (Kim, Diao, and Shuler, 2012; Frye et al., 2016). These initial cell line screenings experiments typically involve the use of high-throughput 96 well plates or microtitre plates (MTPs). These experiments are time-consuming, do not accurately reflect scale-up conditions and require a significant amount of analytical equipment to evaluate best performing cell lines. Furthermore, the low volume of these batch cultures during initial screening restricts the availability of off-line analytics resulting in the assessment criteria for each clone solely dependent on protein concentration and/or cell growth metrics.

In order to speed up the selection of the most productive cell lines and reduce product development timelines, the industry has developed different high throughput (HT) methods. Systems such as microtitre plates (MTPs) with a volume of 100–400 µL are great during initial screening allowing for HT runs in parallel, however they lack automation and their orbital shaking design can generate variability in comparison to other systems (Li et al., 2010; Huang and Kwiatkowski, 2015). Deep well plates (DWP) with a volume of 0.5–2 mL also have HT capabilities however similar to MTPs can be affected by poor mixing as their shaking is not comparable to mixing performed by impellers (Ben-Tchavtchavadze, Perrier, and Jolicoeur, 2007; Huang and Kwiatkowski, 2015). Additional automated systems have also been developed such as the SimCell™ microfluidic and micro-bioreactor systems; the Pall Micro24 system with a deep well design, each mini reactor has a standard baffled design and a working volume of up to 7 mL (Warr et al., 2013; Huang and Kwiatkowski, 2015). These systems have control units that monitor cell culture process parameters and can successfully predict key product quality attributes; it has also shown potential for scale-up applications through media screening and process optimisation (Ravindran et al., 2019; Huang and Kwiatkowski, 2015; Whitford, 2006). Other popular HT microbioreactor systems include the ambr® 15 cell culture and ambr® 250 systems that contain impellers and are comparable scale-down mimics of bench-scale systems (Huang and Kwiatkowski, 2015).

As shown in Figure 1, the micro-bioreactor stage is one of the most important stages as it is the first stage where both off-line and on-line process variables are recorded. At this stage the number of clones has been reduced to manageable numbers in the region of 10–100 clones. Additionally, this is the first stage of the selection process that evaluates the impact of controlling the key environment variables (e.g., pH and dissolved oxygen) in conjunction with fed-batch operation on the key productivity and growth metrics of each clone. This stage of the CLD process is a critical decision point in the selection of the lead clone as all the necessary information is available to make an informed decision on the optimum choice of lead clone as both process and off-line data analytics are available. However, this decision is heavily focused on off-line productivity metrics such as final titre with little consideration of the on-line variables such as pH or dissolved oxygen (DO_2). Yet, these environmental variables can have a major influence on the metabolism of mammalian cell cultures; for example, Konakovsky et al. (2016) demonstrated the impact of adjusting the pH set-point on the production and consumption rates of lactic acid in addition to influencing the glucose consumption rates. Therefore, it is important to better understand the impact of these environment variables on the available productivity and growth metrics at this stage as they may play a significant role as the lead clone is scaled to commercial manufacturing. The on-line data is therefore a very valuable asset as it can help assess problems during scale-up such as evaluating the oxygen consumption rate of the cells to ensure the final commercial-scale bioreactor can meet the oxygen requirements for those high yielding cells (Garcia-Ochoa and Gomez, 2009). Furthermore, having access and analysing all of this data might highlight some latent operation issues that require subsequent evaluation and could pose a material impact on the tight timelines.

This paper outlines the development of a data-driven methodology to enhance lead clone selection that not only considers the available titre concentration and product quality that is the primary source of information during conventional CLD but also leverages the significant untapped data resource containing all the available off-line, on-line and metadata. To help automate this selection process this paper outlines the use of a simple natural language generation (NLG) algorithm to evaluate all available information which summarises and contextualises the large volume of information into a human readable report. This automatically generated report outlines the key metrics and other useful correlations to assist the operator for their lead clone selection. This methodology aims to remove a significant burden of time on all scientists and engineers that repeatedly spend large amounts of time on writing up and presenting results on this critical business decision. This autonomous cell line selection protocol demonstrates a novel application of NLG within the biopharmaceutical space that leverages all data within the data lake. This enables more informed data-driven decisions that pave the way towards Industry 4.0 implementation within the bioprocessing manufacturing sector.

2 Materials and methods

2.1 Cell line, culture propagation, and bioreactor systems

All cell lines used a Chinese hamster ovary (CHO) host that expressed high levels of an antibody-peptide fusion protein and were cultivated in chemically defined CHO media. The cells were maintained at 37°C under 5% carbon dioxide, shaken at a constant rpm and passaged 2–3 times per week for propagation and scale-up for inoculation. The cell line screening was carried out on an advanced micro-bioreactor ambr® 15 system (Sartorius., 2020) with 48 single vessels split into four separate culture stations where each vessel was operated with a 11–15 mL working volume. The temperature and pH of each culture station was controlled to 33°C and 7.0, respectively and the agitation rate was ramped up to ensure the dissolved oxygen concentration of 50% could be maintained. The feeding strategy involved five equally spaced additions of the feed after the initial feed day indicated. The culture pH was controlled to 7.0 through the addition of sodium carbonate and sparging with CO_2 gas with its control strategy implementing a pH dead-band equal to 0.1. Antifoam was added as required. Daily at-line samples were analysed for viable cell density (VCD) and viability (Viab) using the Vi-Cell Automated Cell Viability Analyzer (Beckman Coulter, Brea, CA, United States), and glucose (Gluc) and lactate (Lact) were analysed using the 2900D Biochemistry Analyzer (YSI, Yellow Springs, Ohio, United States). The rpm set-points, initial seeding density, and supplemented feeds contain proprietary information, which prevents us from providing details about them in this section.

2.2 Titre analysis and purity

Volumetric antibody-peptide fusion titres (Titre) in cell culture supernatants were quantified by protein A affinity chromatography using a protein A ImmunoDetection sensor cartridge (Applied Biosystems, Warrington, United Kingdom) coupled to an Agilent 1,200 series HPLC (Agilent, Berkshire, United Kingdom). Peak areas relative to a reference standard calibration curve were used to calculate titres. These samples were measured on days 8, 10, 12, and 14 for the ambr® 15 system. The trisulfide bond (TSB) was quantified through a TQS triple quadrupole mass spectrometer (Waters, Milford, MA, United States). The monomer purity including fragment and aggregate concentration was monitored with high-performance size exclusion chromatography (HP-SEC) using a TSK-GEL G3000SWXL column (7.8 mm × 30 cm) from Tosoh Bioscience (King of Prussia, PA, United States) with an Agilent 1200 HPLC system (Agilent Technologies, Santa Clara, CA, United States).

2.3 Cell line manufacturability index (MI_{CL})

The cell line manufacturability index (MI_{CL}) adapts the standard weighted sum model (Fishburn, 1967) that has been used in the biotech sector to weigh up conflicting criteria for process decisions (Pollock et al., 2017), capacity sourcing decisions (George et al., 2007) and manufacturability indices for formulations (Yang et al., 2017). MI_{CL} aggregates all available productivity, product quality and growth parameters into a single metric enabling easier evaluation and comparison of m cell lines considering n criteria:

$$MI_{CL,i} = \sum_{j=1}^{j=n} w_j \times r_{ij} \quad \text{for } i = 1, 2, 3 \ldots m \tag{1}$$

where w_i is the normalised weight of each criteria j and r_{ij} is the dimensionless rating of cell line i and criteria j, calculated as:

$$r_{ij} = \left| \frac{x_{ij} - x_{j,worst}}{x_{j,best} - x_{j,worst}} \right| \tag{2}$$

where x_{ij} is the individual ranking of cell line i for criteria j, $x_{j,best}$ is the best overall ranking for criteria j and $x_{j,worst}$ is the worst cell line ranking for criteria j. $x_{j,best}$ and $x_{j,worst}$ are subjectively defined as either a maximum or minimum for each individual criteria based on expert knowledge. For example, for the parameter titre, $x_{j,best}$ would be the highest titre based on analysing all available titres in the cell line selection and $x_{j,worst}$ would be the lowest titre concentration. However, for final lactate concentration, $x_{j,best}$ would be the lowest lactate concentration with $x_{j,worst}$ equal to the highest concentration. The w_j is the normalised weight of each criteria and can be adjusted for an individual project. The weight can be dependent on several factors including the mode of action of the molecule, market demand, sub-class of molecule (e.g., different weightings and criteria can be selected for monoclonal antibodies (mAbs), fragment antibodies (Fabs), bispecifics, or fusion proteins). The normalised weights ensure that the best possible manufacturability index of an individual cell line MI_{CL} would be less than or equal to 1, i.e., if a cell line i has the best value for each criteria j its MI_{CL} would be equal to 1.

2.4 Rule-based natural language generation

The automatically generated report developed in this work implemented a rule-based natural language generation (NLG) algorithm based around the architecture designed for Data-to-Text systems outlined by Becker et al. (2007). The Data-to-Text systems are NLG algorithms that are specifically designed to generate texts from sensor data or other relevant non-linguistic data types. Four stages are required to implement this approach to generate a text from data:

- Signal Analysis: Analysis of the numerical data to identify correlations and useful trends.
- Data Interpretation: Identifying the key messages or patterns within the data.
- Document Planning: Deciding on a document structure that will communicate the key messages from the data analysis and interpretation.

- Microplanning and Realisation: Creating the actual text using the NLG.

Within this work the "Signal Analysis" and "Data Interpretation" stages were completed in steps 2 and 3 of the CLD_4 methodology. The "Document Planning" and "Microplanning and Realisation" were completed by Step 4 of the CLD_4 and used a simple template to communicate the key messages from the data analysis relevant to the decision making within CLD. The metrics utilised within the report were calculated using Eqs 3, 4. These included the required capacity to meet the production target based on an estimated number of patients and dose size of the drug within the CLD pipeline.

$$D_{product} = N_{patients} \times d_{patient\_year} \tag{3}$$

$$V_{batch} = \frac{D_{product}}{T_{est} \times N_{batches} \times Y_{DSP} \times P_s} \tag{4}$$

where $D_{product}$ is the annual product demand (g), $N_{patients}$ is the estimated number of target patients, $d_{patient\_year}$ is the annual dose required per patient (g), V_{batch} is the bioreactor working volume required per batch (L), T_{est} is the estimated titre from the lead clone (g L^{-1}), $N_{batches}$ is number of batches per year, Y_{DSP} is the expected overall process yield (%) and P_s is the batch success rate.

2.5 Data analysis and visualisation

All online, offline, and meta data were imported and analyzed using algorithms developed in Python 3.9.12 (Python programming language) and Matlab 2021b (The MathWorks, Inc., Natick, MA). R (R Foundation for Statistical Computing, Version 4.1.2, Vienna, Austria) was used to generate the correlation matrix. Matlab 2021b was used to visualise all the graphs.

3 Results and discussion

Data is now a major asset for biopharmaceutical companies that continues to grow in both size and complexity, and yet to date this data resource has not yet been fully exploited (Narayanan et al., 2020). This paper helps address this issue within cell line development, where only a small fraction of the available data is utilised in lead clone selection. This work proposes a novel data-driven workflow to improve selection of the lead clone through the analysis of multiple interconnected sources of information using advanced machine learning algorithms. This workflow represents the next-generation of lead clone selection by leveraging all available information for better decision making. The workflow is referenced as CLD_4 as it involves four steps and embodies the core principles of Industry 4.0 enabling autonomous data-driven decisions. The first step of CLD_4 involves pulling the data in its raw format into a data lake and subsequently classifying the data into different categories, storing the data appropriately within a data warehouse in addition to preprocessing and calculating key features within the data. The second step assesses this structured data within the data warehouse to calculate a new metric referenced as the Cell Line Manufacturability index (MI_{CL}) which ranks all individual cell lines based on selected criteria. The third step involves the application of data analytics using the preprocessed data from step 1 to gain an understanding of how the process operation influences the critical quality attributes and to highlight any potential issues that may occur during scale-up operations and/or commercial manufacturing. Step 4 autonomously generates a report utilising the essential bits of information calculated from Step 1 and the MI_{CL} from Step 2 to select the lead clone in addition to compiling the key insights generated from the data analytics from Step 3. In addition to selecting the lead clone, this final fourth step provides recommendations supporting long-term business decisions related to capacity planning and future production targets based on additional metadata recorded by the scientists.

3.1 CLD$_4$ step 1: data classification and storage

The first step of CLD_4 was to import all the unstructured raw data related to the cell line development activities into a data lake. A summary of the raw data types is outlined in Table 1. Part of this data transformation involves classifying all available biopharmaceutical data related to CLD activities into five different categories: 1. Process Parameters, 2. Growth, 3. Productivity, 4. Product Quality and 5. Meta-information. A detailed description of the 5 categories used in this work is shown in Table 1 which outlines the frequency, storage system and format of the data.

It is evident from Table 1 that one of the major challenges with the consolidation and digitalisation of these raw data sources within the data lake is the multiple formats and storage locations of each data type. This challenge was echoed by Steinwandter et al, (2019) who discussed the multiple data formats and outputs recorded by the different analytical devices and bioreactor systems to be one of the major challenges faced by the biopharmaceutical industry. To help alleviate these issues, there are a number of initiatives such as FAIR data principles (Wilkinson, 2016) that aim to improve infrastructure for better data management and ensure the data is Findable, Accessible, Interoperable and Reusable. Additional standards include the Allotrope Data Format (Millecam et al., 2021) that has the goal of standardising these data formats within industry by ensuring all analytical providers and vendors output their data in a standardised platform-independent data format (i.e., HDF5), which could greatly simplify the consolidation and curation of the data. For this work, an in-house algorithm was developed to gather and store all the data resources shown in Table 1 into a data lake. The raw data then follows an ETL (Extract Transform Load) procedure where the data is extracted from the raw sources and mapped to a large queryable table within a data warehouse. The ETL algorithm imported all the data recorded by the ambr 15 bioreactor that resulted in the importation of a total of 592 csv files (12 csv files per cell culture run and 4 csv files per cell culture station) in addition to importing all other data recorded from each of the analytical devices outlined in Table 2. All this information was converted into a queryable format within the data warehouse based on bioreactor ID, timestamp and source of each data type, this allows all of the online, offline and meta information of each bioreactor run to be easily extracted and used for subsequent analysis. The data was stored as JavaScript Object

TABLE 1 The five data classifications related to CLD_4 are summarized by their recording frequency, source and storage format.

Category	Description	Frequency	Source	Format	Examples
Process parameters	Bioreactor process conditions	Every 1–90 s	Single bioreactor system (ambr® 15) includes all data recorded by multiple sensors	Multiple csv files with bioreactor ID and timestamps (proprietary format)	Process data and set-points of dissolved oxygen (DO_2), temperature (T), pH (pH), stirrer speed (RPM) and gas flow rates (F_{O2}, F_{CO2})
Growth	All variables related to cellular growth and nutrient consumption	Every 24 h	Recorded across multiple analytical devices (e.g., YSI 2900, Vi-Cell XR cell counter)	All data saved with timestamp and unique bioreactor reference on internal server and exported as csv/excel format	Glucose ($Gluc$), lactate ($Lact$), viable cell density (VCD), growth rate (Spec growth rate)
Productivity	Off-line variables recorded	Every 24/48 h	Recorded across multiple analytical devices (e.g., Agilent 1,200 series HPLC)	Specific to analytical instrument, data exported using Excel based template containing bioreactor ID and timestamp	Titre concentration (mAb) and specific productivity ($q_{antibody}$)
Product quality	Information specific to protein structure	Recorded at harvest of cell culture run	Recorded across multiple analytical devices (SEC-UPLC, Xevo TQS triple quadrupole mass spectrometer)	Specific to analytical instrument, data exported using Excel based template containing bioreactor ID and timestamp	Monomer ($Mono$), aggregation (Agg) and trisulfide bond (TSB)
Metadata	Non-numerical data recorded, includes cell culture process observations and predicted target market of molecule	Infrequent/not often recorded	No standard recording format, electronic lab notebooks, PowerPoint presentations and management meeting	No standard format, manually recorded by scientist with additional information requested by upper management	Foaming observations, contamination issues, molecule reference, estimated patient dose and estimated market demand

Notation (JSON) files which is a lightweight format for storing and transporting data, is easy to read and can handle different data types and formats. By structuring this data within a data warehouse, algorithms can be universally applied to the data regardless of previous data format, structure or stage in the screening process which greatly simplifies the application of data analytics. Within this paper the primary focus is on the data recorded during the microbioreactor stage as this stage contains the most amount of information as outlined in Figure 1. Combining data from early and/or late screening could further improve the decision on lead clone provided the cell lines are labelled consistently across the scales and the offline analytics are available for analysis.

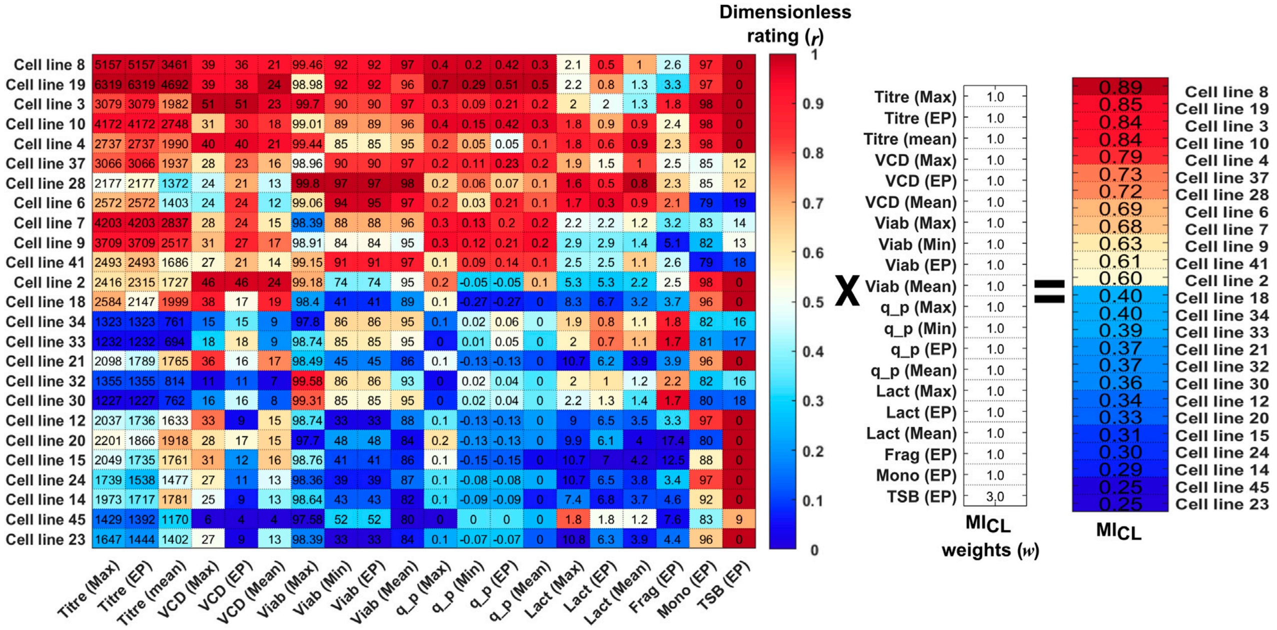
To simplify the analysis and reduce the complexity of the available time-series data stored in the data warehouse, a selection of features were extracted from this data. The features calculated in this work included the maximum, minimum, end point, cumulative, average, standard deviation, and time above/below set-point as summarised in Table 2. The feature selection method used in this paper is based on industrial experience and the features selected were found to be suitable for this case study. However, Table 2 is not a comprehensive list of features and additional features could be added or removed during on the process. The calculation of these engineered features plays a crucial role in steps 2, 3, and 4 of the CLD_4 methodology and has two major advantages. The first is the significant reduction of the data size, this is particularly important for the on-line variables. For example, the pH recorded for an individual ambr® 15 bioreactor cell culture typically contains ~15,000 data points, based on a 90 s interval for a 16 days period. The majority of the important trends within this pH trend can be summarised by only 6 new features (max, min, end-point, average, std, time above set-point (S.P) and time below S.P) resulting in a drastic reduction of the data in addition to greatly simplifying the subsequent analysis and visualisation of the data. Using the suggested features shown in Table 2, this can be reduced to 6 data points resulting in ~2,500 fold reduction in data. This data reduction also significantly improves the ease of data analysis in comparison to analysing the complete time-series. Goldrick et al. (2017) highlighted the challenge of using the complete time-series data for end-point predictions of a CQA which involved complex batch-wise unfolding operations, data pre-processing and interpolation methods before a robust PLS and MLR model could be generated.

3.2 CLD_4 step 2: cell line manufacturability index (MI_{CL}) calculation

The next step formulates the selection of a lead clone as a multi-criteria decision-making problem enabling scientists to evaluate the performance of each individual clone. The new metric is referenced as the cell line manufacturability index (MI_{CL}). The MI_{CL} was calculated for the 48 clones using the 20 selection criteria shown in Figure 2. For illustration purposes; Figure 2 shows the top and bottom 12 clones ranked from 1–12 and 36–48 based on their MI_{CL} value. The MI_{CL} is a weighted sum metric defined by Eq.1 that considers 20 different selection criteria related to key features extracted in step 1 that summaries the productivity, growth, product quality of each cell line. The rating of each selection criteria was calculated using Eq. 2, where $x_{i,worst}$ and $x_{i,best}$, is the worst and best value for each cell line within the run, respectively. The maximum possible value of MI_{CL} is equal to 1 taking into account the weight of each criteria as defined in Eq. 1. The raw data

TABLE 2 A summary of variables and their engineered features within each of the five data categories by *CLD$_4$* implementation. Control strategy outlines the objective within the bioreactor to either maximum (Max.), minimise (Min.) or control the variable at its set-point (S.P.).

Variable (Units)	Control strategy	Feature engineering							
		Max	Min	End point	Cumul	Avg	Std	Time above S. P.(hrs)	Time below S.P. (hrs)
Productivity									
Titre (mg L^{-1})	Max	✓	✗	✓	✗	✓	✗	✗	✗
$q_{antibody}$ (pg $cell^{-1}$ day^{-1})	Max	✓	✓	✓	✗	✓	✗	✗	✗
Growth									
VCD (cells × 10^6 mL^{-1})	Max	✓	✗	✓	✗	✓	✗	✗	✗
Viability (%)	Max	✓	✓	✓	✗	✓	✗	✗	✗
Lactate (g L^{-1})	Min	✓	✗	✓	✗	✓	✗	✗	✗
Glucose (g L^{-1})	S.P.	✓	✓	✓	✗	✓	✗	✓	✓
Product Quality									
Aggregates (%)	Min	Endpoint only							
Fragments (%)	Min	Endpoint only							
Monomer (%)	Max	Endpoint only							
Trisulfide bond (%)	Min	Endpoint only							
Process Parameters									
Temperature (°C)	S.P.	✓	✓	✓	✗	✓	✓	✓	✓
pH (–)	S.P.	✓	✓	✓	✗	✓	✓	✓	✓
DO_2 (%)	S.P.	✓	✓	✓	✗	✓	✓	✓	✓
Base addition (mL)	S.P.	✗	✗	✗	✓	✗	✗	✗	✗
$Flow_{O2}$ (mL min^{-1})	S.P.	✓	✗	✓	✓	✗	✗	✗	✗
$Flow_{CO2}$ (mL min^{-1})	S.P.	✓	✗	✓	✓	✗	✗	✗	✗
Metadata									
Project ID reference	Scientist	Inoculation date	Bioreactor position	Cell line reference	Projected Titre	Molecule reference		Observations	Estimated patient dose



	Titre (Max)	Titre (EP)	Titre (mean)	VCD (Max)	VCD (EP)	VCD (Mean)	Viab (Max)	Viab (Min)	Viab (EP)	Viab (Mean)	q_p (Max)	q_p (Min)	q_p (EP)	q_p (Mean)	Lact (Max)	Lact (EP)	Lact (Mean)	Frag (EP)	Mono (EP)	TSB (EP)
Cell line 8	5157	5157	3461	39	36	21	99.46	92	92	97	0.4	0.2	0.42	0.3	2.1	0.5	1	2.6	97	0
Cell line 19	6319	6319	4692	39	38	24	98.98	92	92	96	0.7	0.29	0.51	0.5	2.2	0.8	1.3	3.3	97	0
Cell line 3	3079	3079	1982	51	51	23	99.7	90	90	97	0.3	0.09	0.21	0.2	2	2	1.3	1.8	98	0
Cell line 10	4172	4172	2748	31	30	18	99.01	89	89	96	0.4	0.15	0.42	0.3	1.8	0.9	0.9	2.4	98	0
Cell line 4	2737	2737	1990	40	40	21	99.44	85	85	95	0.2	0.05	0.05	0.1	1.8	0.6	0.9	2.3	98	0
Cell line 37	3066	3066	1937	28	23	16	98.96	90	90	97	0.2	0.11	0.23	0.2	1.9	1.5	1	2.5	85	12
Cell line 28	2177	2177	1372	24	21	13	99.8	97	97	98	0.2	0.06	0.07	0.1	1.6	0.5	0.8	2.3	85	12
Cell line 6	2572	2572	1403	24	24	12	99.06	94	95	97	0.2	0.03	0.21	0.1	1.7	0.3	0.9	2.1	79	19
Cell line 7	4203	4203	2837	28	24	15	98.39	88	88	96	0.3	0.13	0.2	0.2	2.2	2.2	1.2	3.2	83	14
Cell line 9	3709	3709	2517	31	27	17	98.91	84	84	95	0.3	0.12	0.21	0.2	2.9	2.9	1.4	5.1	82	13
Cell line 41	2493	2493	1686	27	21	14	99.15	91	91	97	0.1	0.09	0.14	0.1	2.5	2.5	1.1	2.6	79	18
Cell line 2	2416	2315	1727	46	46	24	99.18	74	74	95	0.2	-0.05	-0.05	0.1	5.3	5.3	2.2	2.5	98	0
Cell line 18	2584	2147	1999	38	17	19	98.4	41	41	89	0.1	-0.27	-0.27	0	8.3	6.7	3.2	3.7	96	0
Cell line 34	1323	1323	761	15	15	9	97.8	86	86	95	0.1	0.02	0.06	0	1.9	0.8	1.1	1.8	82	16
Cell line 33	1232	1232	694	18	18	9	98.74	85	85	95	0	0.01	0.05	0	2	0.7	1.1	1.7	81	17
Cell line 21	2098	1789	1765	36	16	17	98.49	45	45	86	0.1	-0.13	-0.13	0	10.7	6.2	3.9	3.9	96	0
Cell line 32	1355	1355	814	11	11	7	99.58	86	86	93	0	0.02	0.04	0	2	1	1.2	2.2	82	16
Cell line 30	1227	1227	762	16	16	8	99.31	85	85	95	0	0.02	0.04	0	2.2	1.3	1.4	1.7	80	18
Cell line 12	2037	1736	1633	33	9	15	98.74	33	33	88	0.1	-0.13	-0.13	0	9	6.5	3.5	3.3	97	0
Cell line 20	2201	1866	1918	28	17	15	97.7	48	48	84	0.2	-0.13	-0.13	0	9.9	6.1	4	17.4	80	0
Cell line 15	2049	1735	1761	31	12	16	98.76	41	41	86	0.1	-0.15	-0.15	0	10.7	7	4.2	12.5	88	0
Cell line 24	1739	1538	1477	27	11	13	98.36	39	39	87	0.1	-0.08	-0.08	0	10.7	6.5	3.8	3.4	97	0
Cell line 14	1973	1717	1781	25	9	13	98.64	43	43	82	0.1	-0.09	-0.09	0	7.4	6.8	3.7	4.6	92	0
Cell line 45	1429	1392	1170	6	4	4	97.58	52	52	80	0	0	0	0	1.8	1.8	1.2	7.6	83	9
Cell line 23	1647	1444	1402	27	9	13	98.39	33	33	84	0.1	-0.07	-0.07	0	10.8	6.3	3.9	4.4	96	0

Titre (Max)	1.0
Titre (EP)	1.0
Titre (mean)	1.0
VCD (Max)	1.0
VCD (EP)	1.0
VCD (Mean)	1.0
Viab (Max)	1.0
Viab (Min)	1.0
Viab (EP)	1.0
Viab (Mean)	1.0
q_p (Max)	1.0
q_p (Min)	1.0
q_p (EP)	1.0
q_p (Mean)	1.0
Lact (Max)	1.0
Lact (EP)	1.0
Lact (Mean)	1.0
Frag (EP)	1.0
Mono (EP)	1.0
TSB (EP)	3,0

0.89	Cell line 8
0.85	Cell line 19
0.84	Cell line 3
0.84	Cell line 10
0.79	Cell line 4
0.73	Cell line 37
0.72	Cell line 28
0.69	Cell line 6
0.68	Cell line 7
0.63	Cell line 9
0.61	Cell line 41
0.60	Cell line 2
0.40	Cell line 18
0.40	Cell line 34
0.39	Cell line 33
0.37	Cell line 21
0.37	Cell line 32
0.36	Cell line 30
0.34	Cell line 12
0.33	Cell line 20
0.31	Cell line 15
0.30	Cell line 24
0.29	Cell line 14
0.25	Cell line 45
0.25	Cell line 23

FIGURE 2
Calculation of the MI_{CL} for the top 12 and bottom 12 clones based on the dimensionless rating (r) of the data from the 20 variables multiplied by the selected weights (w) of each variable. The raw data for each clone is shown in the data matrix and each column is coloured coded from high (dark red) to low (dark blue) based on the variables' dimensionless rating.

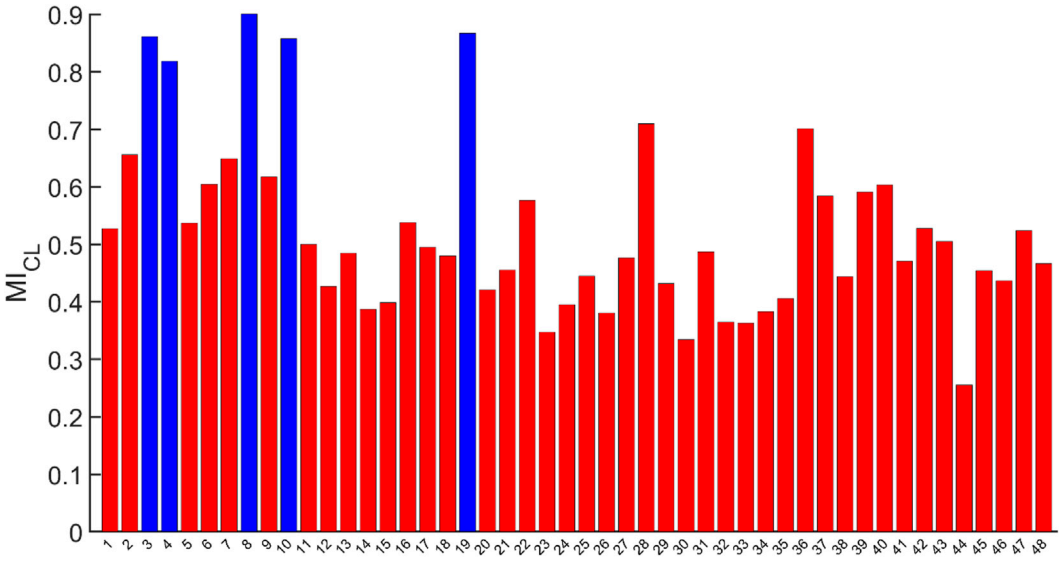


FIGURE 3
MI_{CL} calculated for each of the 48 clones of the ambr® 15 that considered all 20 variables with their predefined weighting where the top 5 clones are shown in blue and identified in order as 8, 19, 3 10, and 4.

stored in the data warehouse utilised for the selection of the top clones in this work is shown in Figure 2. Considering the selection criteria for the maximum titre (*Titre (Max)*) of the 48 clones displayed in Figure 2, the lowest ("*worst*") titre was clone 30 with a titre equal to 1.2 g L^{-1} which equated to a rating value (r_{ij}) of 0, indicated by the dark blue pattern. The highest ("*best*") titre was cell line 19 with a value equal to 6.3 g L^{-1}, resulting in a r_{ij} of 1, indicated by the dark red pattern. Within the data set evaluated in this study there were three variables which had a reverse rating where the higher the value the lower the rating, i.e., lactate (*Lact*)

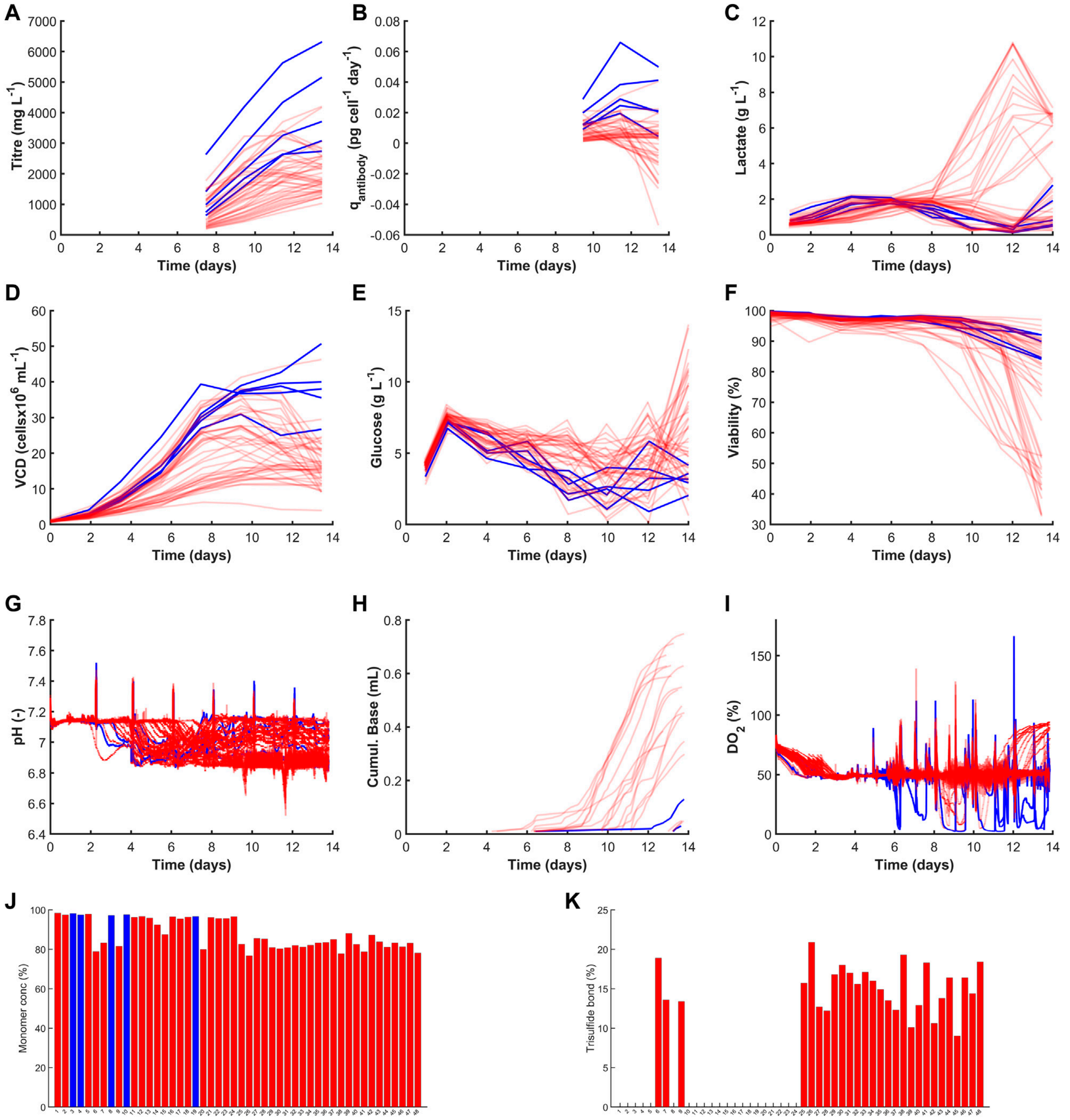


FIGURE 4
Time course profiles and bar graphs of 48 cell lines considered in this work. The cell lines shown in blue highlight the top cell lines selected using the highest MI_{CL} values and those in red were not selected. The figures are separated into their four data categories where the productivity graphs are shown in **(A,B)** for Titre (mg L^{-1}) and $q_{Antibody}$ (pg $cell^{-1}$ day^{-1}). The growth category highlights **(C)** Lactate (g L^{-1}), **(D)** VCD (cellsx10^6 mL^{-1}), **(E)** Glucose (g L^{-1}) and **(F)** Viability (%). The process parameters are shown in **(G)** pH(–), **(H)** Cumul. Base (mL) and **(I)** DO_2 (%). The product quality variables are shown by **(J)** Monomer conc (%) and **(K)** Trisulfide bond (%).

concentration, aggregate (*Agg*) concentration and trisulfide bond (*TSB*) concentration. For these variables, the lowest experimentally recorded value has the highest rating, i.e., the "best" experimental condition and *vice versa* for the lowest value. This is demonstrated by considering the variable lactate where the lowest ("*best*") lactate end point (*Lact (EP)*) was 0.3 g L^{-1} produced by Clone 6 and was given a $r_{i,j}$ rating of 1 and the highest lactate end-point was equal to 7.0 g L^{-1} recorded by clone 15 and given a $r_{i,j}$ equal to 0.

Each of the selection criteria was given the weights displayed in Figure 2, and for this CLD study, the selection criteria for growth and productivity were all given equal weightings of 1. The choice of equal weighting for growth and productivity was based on scientific

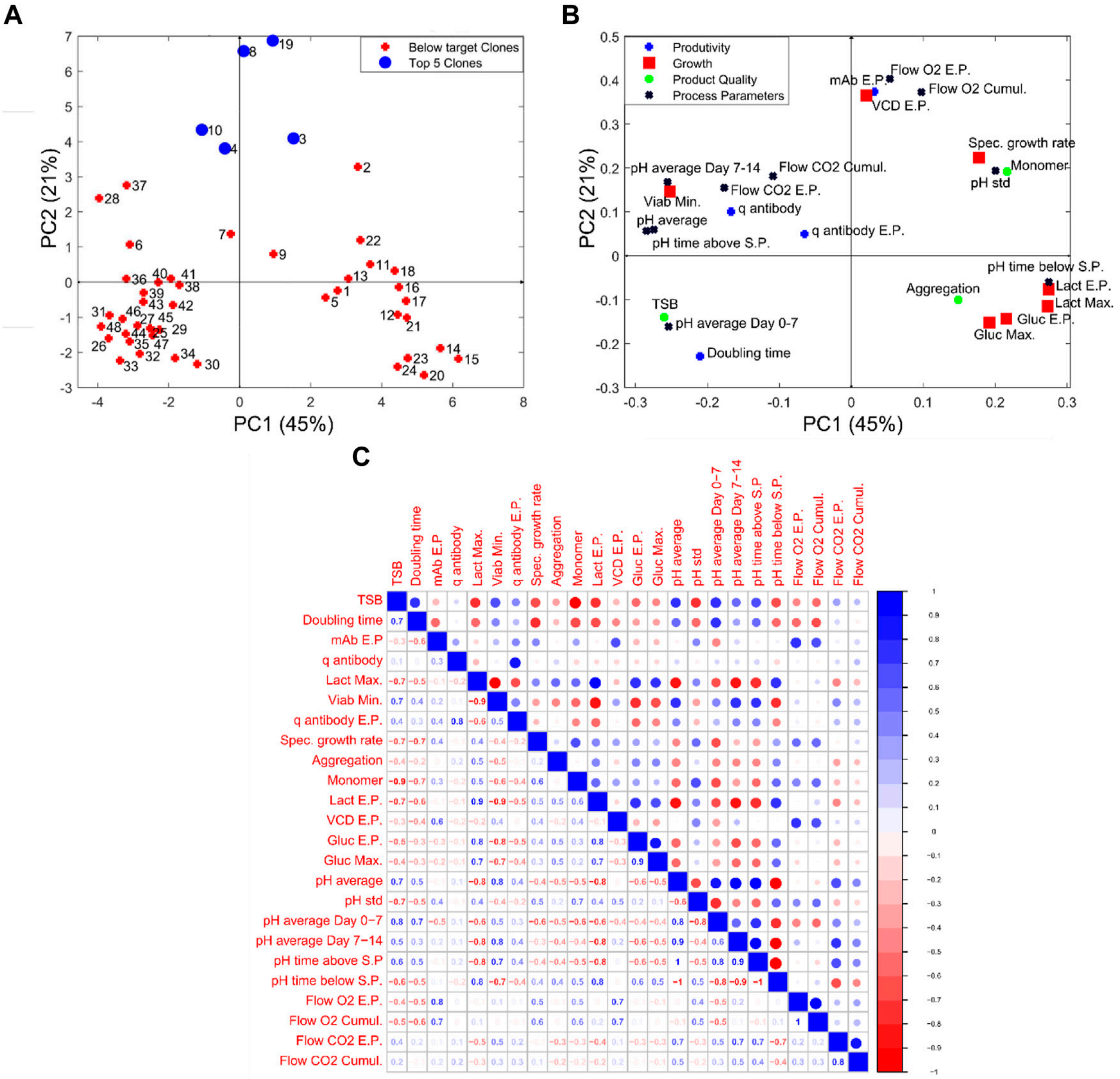


FIGURE 5
Data analysis of the engineered features for the 48 clones with **(A,B)** showing the scores and loadings graph, respectively, of the first and second principal component (PC) from PCA analysis and **(C)** shows the correlation matrix calculated on the same data.

experience and knowledge, for example, it has shown previously that high cell densities and low viabilities can lead to a significant increase in process impurities such as lipids, intracellular proteins and nucleic acids (Roush and Lu, 2008). Additionally, it is has been reported that cell lines that remain in a lactate production state and do not switch to a lactate consumption state towards the end of the culture run can yield low cell growth and productivities (Hartley et al., 2018). Therefore, the growth-related parameters of *VCD* (*Max, Mean* & *End-point*), *Viab* (*Max, Min* & *End-point*) and *Lact* (*Max, Mean* & *End-point*) were all given equal weighting in comparison to the productivity metrics of Titre (*Max*, *Min*, *Mean*, & *End-point*) and q_p (*Max, Min &End-point*). The weightings and choice of variables to be included in this analysis can easily be adjusted to account for different user requirement specifications.

However, as a result of the prior knowledge that high TSB concentrations can alter the potency and physical chemical properties of the protein (Goldrick et al., 2017), the TSB concentration at end-point (*TSB EP*) was given a higher weighting equal to 3 to ensure the top clones had minimum TSB levels. Typically, within CLD in addition to the key productivity and growth metrics there may be additional CQAs to be considered to ensure certain quality standards are met. These can include CQAs such as purity, potency, and efficacy which can have an impact on the therapeutic drug's safety and effectiveness in treating the target disease. Other CQAs can be important for downstream processing such as the concentration of fragments or aggregates. The calculation of the MI_{CL} can consider multiple CQAs, the only challenge in adding additional CQAs is deciding on their weight as if these CQAs are more important that some CPPs such as final lactate concentration then the weighting may need to be increased. Hence in this case study the TSB concentration was given a weighting of 3 as it was determined to be sufficient to ensure the

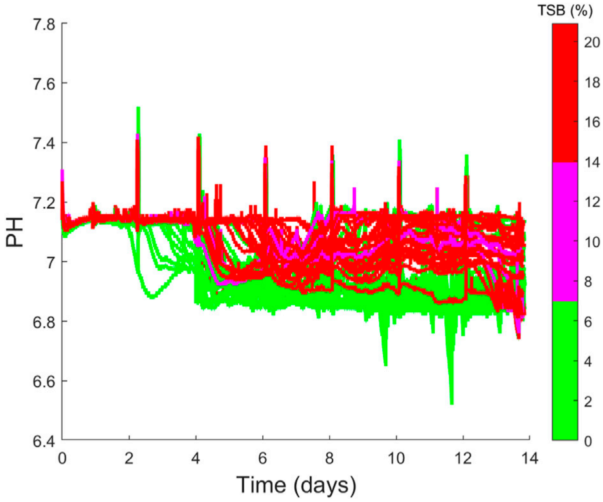


FIGURE 6
Time series profiles of the pH values of the 48 cell cultures considered which are coloured based on their final TSB (%) concentrations, where low TSB is between 0%–7%, medium TSB is between 7%–14% and high TSB is between 14%–21%.

top 5 clones all had 0% TSB concentrations as can be seen in Figure 2.

Using the selection criteria and weights shown the top five clones were 8, 19, 3, 10, and 4 based on their calculated MI_{CL} equal to 0.89, 0.85, 0.84 0.84, and 0.79 as shown in Figure 3. To highlight the performance of these top 5 clones in comparison to the other 48 clones considered in this selection process, some variables are shown in Figure 4. The MI_{CL} provides a simple and comprehensive numerical value summarising the performance of each individual clone considering both the cell lines' ranking within each criteria and the relative importance of each criteria. This simple metric greatly simplifies the scientist's decision in selecting the lead clones that demonstrate favourable growth, productivity and product quality values.

Figure 4 highlights time-course and end-point concentrations of selected variables from each of the four categories: productivity, growth, process parameters and product quality. The observed deviations give an early indication of how these clones will respond to the hydrodynamic shear, shifts in cell culture environmental conditions such as pH, temperature and DO_2 and impact of metabolites concentrations changes such as varying glucose or lactate that may vary at scale-up. The top five clones selected based on the highest MI_{CL} demonstrated high productivity and favourable growth patterns whilst maintaining desired critical quality attributes. The top two clones yielded high titre concentrations equal to 6.3 g L^{-1} (clone 19) and 5.2 g L^{-1} (clone 5) which is a key criteria to minimise operating costs by reducing the capacity requirements at scale-up and confidently meet market demand (Li et al., 2010; Frye et al., 2016; Priola et al., 2016; Krebs et al., 2018). Clone 19 resulted in a four-fold increase in titre compared to clone 30 with the lowest titre equal to 1.2 g L^{-1}. Lin et al. (2019) has investigated the variability in high producing CHO clones and observed a similar 3-4 fold increase from 0.1 g L^{-1} to 0.4 g L^{-1}; they attributed the variability to the selection marker and observed the use of attenuated glutamine synthetase (GS) which removed the need for methionine sulfoxamine (MSX) and was shown to generate more stable clones with high productivities. Subsequent process development activities such as adaptive feeding strategies could further increase productivities of these clones as demonstrated by Gagnon et al. (Gagnon et al., 2011), who reported yields in the region of 9–10 g L^{-1}. However, in addition to high productivities it is paramount to ensure the product quality remains within specification as highlighted in Figure 4A where clones 7 and 9 outperformed three of the top five clones but as these both had high levels of TSB their overall MI_{CL} was reduced and therefore they were not selected. Furthermore, this work could be extended to evaluate the risk of the lead clone being out of specification, which may be a useful metric to ensure a stable and robust clone is selected. This demonstrates the flexibility of the MI_{CL} metric as the weights can easily be changed to select a cell line most appropriate for scale-up and commercialisation.

There were also a significant number of clones that had high lactate metabolism towards the end of a culture run as shown in Figure 4C. The MI_{CL} takes this into consideration by considering the maximum, end-point and mean lactate concentration; if minimising lactate concentration at the end of the culture is a key criteria then the weight of the end-point lactate (w_{LactEP}) could be increased. This is related to the amount of base added which is significantly lower for the top 5 clones in comparison to the other clones as shown in Figure 4H). The high volume of base added is due to the high number of clones where there was no back metabolism of lactate as shown in Figure 4C), thus additional base is required to maintain the pH at its setpoint. The cell lines with the high base addition and high lactate end-point also had a much lower viability which can also be observed in Figure 4F). The selected five clones were also shown to have high viability with an end-point value of between 88.7% and 92%, this was in stark contrast to some of the other clones that had low viability, e.g., clones 12 and 23 both had an end point viability equal to 33% as shown in Figures 2, 4F. The inclusion of the viability in the MI_{CL} is paramount as highlighted by researchers that viability below a certain percentage can be problematic to DSP operations (Papathanasiou et al., 2017). The product quality measurements are displayed as bar graphs as only end-points were available for these measurements. Figures 4J, K show the monomer concentration and TSB, respectively with the top 5 clones shown to have the highest purity levels between 96.6% and 98.5% and all had 0% TSB.

3.3 CLD_4 step 3: data analysis

Steps 1 and 2 of the CLD_4 methodology utilise the available data within the categories related to productivity, growth and product quality to select the top 5 clones. Typically, these three categories are the only data types considered during the selection process and the online process parameters defining the cell culture environment are usually ignored. One of the primary reasons for the exclusion of this data is due to the challenge of extracting useful information from these high dimensional data

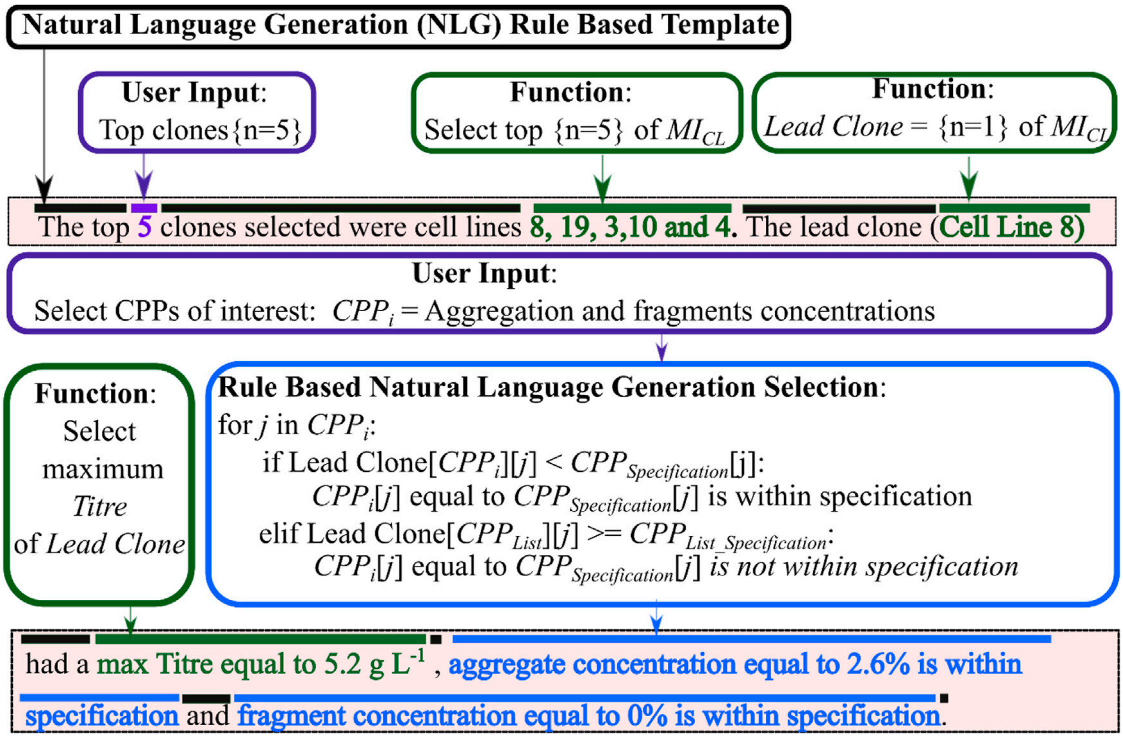
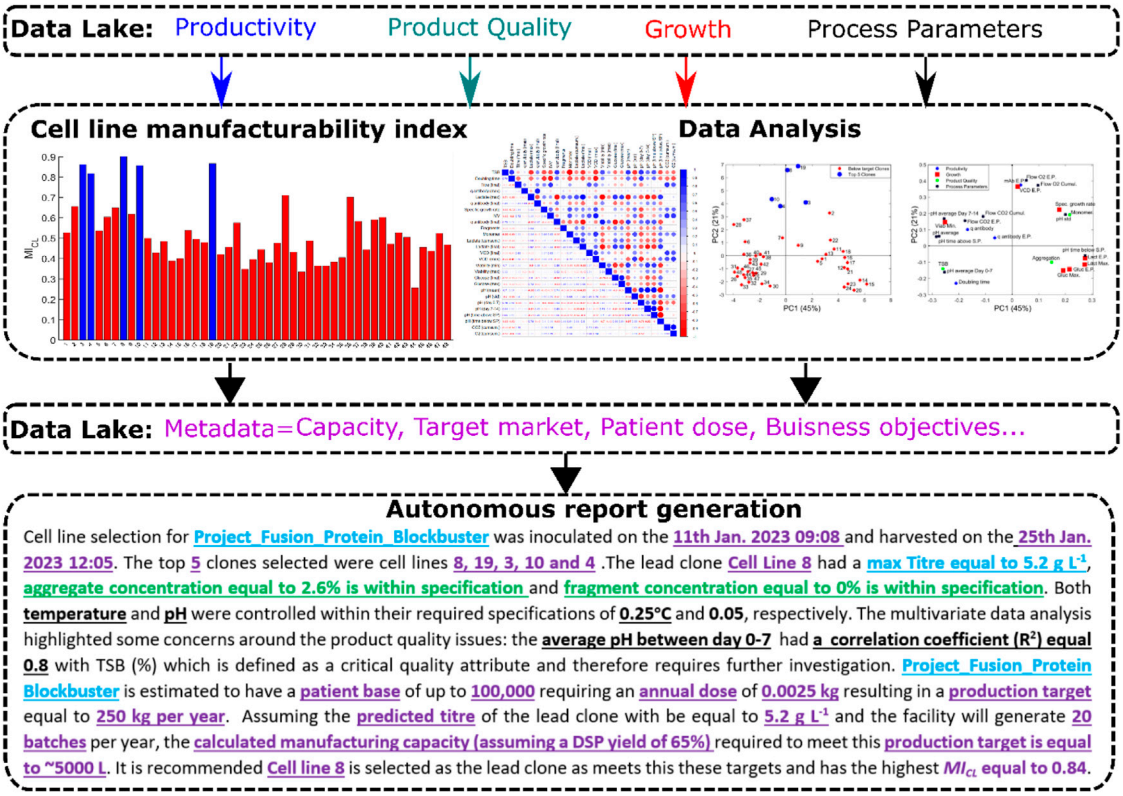

Natural Language Generation (NLG) Rule Based Template
User Input: Top clones{n=5}
Function: Select top {n=5} of MI_{CL}
Function: Lead Clone = {n=1} of MI_{CL}
The top 5 clones selected were cell lines 8, 19, 3,10 and 4. The lead clone (Cell Line 8)
User Input: Select CPPs of interest: CPP_i = Aggregation and fragments concentrations
Function: Select maximum Titre of Lead Clone
Rule Based Natural Language Generation Selection:
for j in CPP_i:
if Lead Clone[CPP_i][j] < $CPP_{Specification}$[j]:
CPP_i[j] equal to $CPP_{Specification}$[j] is within specification
elif Lead Clone[CPP_{List}][j] >= $CPP_{List\_Specification}$:
CPP_i[j] equal to $CPP_{Specification}$[j] is not within specification
had a max Titre equal to 5.2 g L^{-1} , aggregate concentration equal to 2.6% is within specification and fragment concentration equal to 0% is within specification.

FIGURE 7
A sample of the natural language generation rule-based template used in this report. The document template text is highlighted in black, user input data highlighted in purple and text generated using a function shown in green.

sets. Considering the time-series data shown for the 48 clones shown in Figure 4, it is almost impossible to conclude what is the optimal pH profile (displayed in Figure 4G) or investigate if the dissolved oxygen deviations shown in Figure 4I are influencing the growth parameters. Step 3 of this CLD_4 methodology aims to solve these issues by combining feature engineering with the use of ML. The features generated for all the cell culture data was previously discussed in Step 1 of the CLD_4. These newly created features were then analysed using principal component analysis (PCA). To negate the impact of different units used by PCA, all of the analysis was done by first normalizing the data by subtracting the mean and dividing by the standard deviation. This normalization process ensures that variables of different units can be evaluated together. The scores and loadings plots of the data analysed are shown in Figures 5A, B, respectively and were generated as described in the materials and methods. The PCA considered a total of 24 features extracted from each of the 48 clones with the first and second principal components accounting for 45% and 21% of the total variance, respectively. The selected parameters used here were based on user experience and could be extended to include more variables. The scores plot shown in Figure 5A classifies the “Top 5 Clones” (Clones 8, 19, 3, 10, and 4) based on achieving the highest MI_{CL} values (calculated in Step 2) and the remaining clones were classified as “below target clones”. The classification within the loadings plot in Figure 5B is based on the four previously described categories. Through the analysis of the scores and loadings plots, the top five clones cluster together based on similar end-point concentrations of titre (mAb E.P) and VCD (VCD Max) in addition to similar end-point and cumulative flow rates of O_2 (Flow O2 E.P. and Flow O2 Cumul.). The titres of these clones ranged between 4–6 g L^{-1} and had approximately 40–50 × 10^6 cells mL^{-1} at harvest as previously shown in Figures 2, 4. The high cell densities associated with these top clones corresponded to the expected high consumption rates of oxygen (Flow O2 EP and Flow O2 Cumul.) as the oxygen uptake rate (OUR) consumption is typically a function of cell density (Deshpande and Heinzle, 2004).

Additional insights can be leveraged by examining the two distinct clusters observed in the PCA figure between the “below target clones” and the “Top 5 clones” with both clusters exhibiting similar process characteristics. The cluster shown in the right quadrants of the scores plot are correlated with high lactate and glucose concentrations (Lact E.P., Lact Max, Gluc E.P., and Gluc Max) and distinguished by their higher pH values (pH time below S.P, pH time above S.P, pH Std). Although each of these clones are controlled using a fixed pH set-point of 7, the wide deadband of the pH controller observed in Figure 4G allows the pH to drift between the upper and lower bands of this pH set-point. These cell culture runs all had high-end point lactate as observed in Figure 4C, where the end-point lactate concentration was as high as 8 g L^{-1}. As expected the Pearson’s correlation coefficient (R^2) between the maximum lactate concentrations and minimum viability (Viab min) is equal to 0.9; this reduction in viability was shown to significantly decrease the productivity of these cells and corresponded to negative correlation with productivity values



Data Lake: Productivity Product Quality Growth Process Parameters

Cell line manufacturability index

Data Analysis

Data Lake: Metadata=Capacity, Target market, Patient dose, Buisness objectives...

Autonomous report generation

Cell line selection for Project_Fusion_Protein_Blockbuster was inoculated on the 11th Jan. 2023 09:08 and harvested on the 25th Jan. 2023 12:05. The top 5 clones selected were cell lines 8, 19, 3, 10 and 4 .The lead clone Cell Line 8 had a max Titre equal to 5.2 g L^{-1}, aggregate concentration equal to 2.6% is within specification and fragment concentration equal to 0% is within specification. Both temperature and pH were controlled within their required specifications of 0.25°C and 0.05, respectively. The multivariate data analysis highlighted some concerns around the product quality issues: the average pH between day 0-7 had a correlation coefficient (R^2) equal 0.8 with TSB (%) which is defined as a critical quality attribute and therefore requires further investigation. Project_Fusion_Protein Blockbuster is estimated to have a patient base of up to 100,000 requiring an annual dose of 0.0025 kg resulting in a production target equal to 250 kg per year. Assuming the predicted titre of the lead clone with be equal to 5.2 g L^{-1} and the facility will generate 20 batches per year, the calculated manufacturing capacity (assuming a DSP yield of 65%) required to meet this production target is equal to ~5000 L. It is recommended Cell line 8 is selected as the lead clone as meets this these targets and has the highest MI_{CL} equal to 0.84.

FIGURE 8
A summary of the CLD_4 methodology that includes the Manufacturability Index (MI_{CL}), data analysis and NLG report ensuring better lead clones can be selected during CLD that scale-up appropriately and meet the necessary production targets of the drug.

equal to −0.5 as shown in Figure 5C. Similar statistics were reported by Le et al. (2012) through their analysis of 234 cell culture runs and found a R^2 value of −0.87 between final lactate concentration and product yield. This high negative correlation between lactate production and productivity values demonstrates the importance of selecting clones that are shown to back-metabolise lactate but also highlights the influence the environmental conditions within the bioreactor on these clones.

The second cluster of clones is located to the left of the scores plots and highlights some very interesting correlations between the four cell culture categories of productivity, growth, process parameters and product quality. The cluster primarily consists of clones 36–48 and the most notable variable defining the characteristics of this cluster is the high levels of TSB concentration (10%–20%) associated with each of these clones, which is evident in Figure 4J). The high TSB concentrations were considered an important CQA and therefore gaining an understanding of the process parameters that minimise TSB concentration is of tremendous value for subsequent scale-up activities. Through the analysis of the loadings plot shown in Figure 5B, it suggests high pH operation leads to significantly increased TSB concentrations. This is evident on the loadings graph based on a similar positioning of the TSB and the pH average Day 0–7 & 7–14, pH average, pH time above S.P with respect to PC-1. This is further validated through the correlation matrix shown in Figure 5C, where the TSB concentration and pH average day 0–7 has an R^2 equal to 0.8 and an R^2 equal to 0.7 for both the average pH and pH standard deviation. High pH levels were previously shown to influence TSB concentrations with cell culture operations (Goldrick et al., 2017). The significant influence of pH on the TSB only becomes obvious when the online pH data of the 48 clones is plotted using a coloured based classification of the end-point TSB concentration as shown in Figure 6. Although all the process operation set-points for each clone were fixed some fluctuations around the set-points was observed, this can be seen in Figure 6 where the pH has a set-point equal to 7 with a dead-band equal to 0.1. It is clear from Figure 6 that the clones that operated for the majority of the culture closer to the upper band of the pH dead-band equal to 7.1 had a higher end point concentration of >15% TSB. Whereas, those clones that operated closer to the bottom of the pH dead-band equal to 6.9 had much lower TSB concentrations of between 0% and 6%. It was previously demonstrated that higher pH levels (above 7) towards the end of the cell culture run could promote TSB formation (Goldrick et al., 2017). Additionally it was suggested by (Nielsen et al., 2011) that TSB formation can result from a nucleophilic attack of the sulfide ion (SH^{-}) on the disulfide bond of the protein resulting in the formation of a trisulfide bond and this reaction requires a pH at or above neutral, similar to the pH ranges shown for high TSB concentrations shown in Figure 6. Therefore, it is evident that the pH environment of the bioreactor plays a significant role on TSB concentrations and highlights the need for a tighter control of pH to minimise these deviations. During conventional cell line selection screening, the on-line variables such as pH are generally not considered and therefore this issue would not have been identified. However, step 3 of CLD_4 allows for these important correlations to be identified and leverages the power of PCA analysis to identify hidden process characteristics that may influence process performance.

3.4 CLD_4 step 4: autonomous report generation

The final step of the CLD_4 methodology is to summarise the insights gained within steps 1–3 in an electronic format enabling better decision making to support long-term business objectives and mitigate any potential risks that may impact scale-up activities. Steps 1–3 considered all the information from the four categories of productivity, growth, process parameters and product quality outlined in Table 1. Step 4 takes into account the metadata related to the therapeutic product's commercial potential that is not fully utilised during the decision making within CLD activities. One of the primary reasons this metadata is not fully harnessed is due to the challenge of extracting valuable and useful information from unstructured text sources. Often important information is communicated through presentations or decisions recorded during a meeting or in an email and is not documented correctly as metadata. Within this work metadata was recorded on target market size for this therapeutic drug in addition to consideration of the available manufacturing capacity within company. The inclusion of this information ensures the required dose and target patient population are considered to ensure the selected lead clone titres can successfully meet production targets given the available manufacturing capacity. Delays in launch of therapeutic drug can be very costly within the biopharmaceutical sector. Therefore, it is imperative that the manufacturing capacity required to meet market demand is considered in this decision making as this may influence the portfolio management of the company and needs to be considered within any budget planning procedures by management (Farid et al., 2020).

In this work, the final step of CLD_4 developed a rule based natural language generation (NLG) algorithm that interpreted all the structured data within the data lake and transformed this information into a human readable report. This report helps contextualise all this information to help the scientists make a more informed data-driven decision when selecting the lead clone. The template-based NLG used a Data-to-Text systems as described in the materials and methods. A sample of the rule-based NLG report is shown in Figure 7, which highlights which text is part of the template, which text requires user input and which text was automatically generated by analysing the data within the data warehouse and using information from steps 1-3 of CLD_4. This NLG method was chosen as it is specifically designed to generate texts from sensor data or other relevant non-linguistic data types. Similar reports using this type of Data-to-Text include generating weather reports from weather data (Goldberg, Driedger, and Kittredge, 1994). The rule-based NLG aims to standardise how the results are effectively communicated and to ensure better transparency over this critical decision. NLG algorithms have also been used within healthcare to improve and streamline communication between healthcare professionals and their patients (Cawsey, Webber, and Jones, 1997). Figure 8 includes the final NLG report and highlights the various steps to generate the report. The rule-based NLG is deterministic in nature and ensures consistent and reproducible reports but could be further expanded to utilise more sophisticated algorithms to generate more stochastic reports that may provide more insights from the data analysis.

4 Conclusion

The CLD_4 methodology presented in this work represents the next-generation process for lead clone selection through advanced data consolidation, analysis and autonomous report generation. Each of the four steps defined by CLD_4 leverage value from the high volume of data recorded throughout the CLD operation. The CLD_4 evaluated the performance of multiple recombinant CHO cell lines producing high levels of an antibody-peptide fusion with a known product quality issue and recommended selection of the top clone. A key aspect of CLD_4 was the creation of a data lake that stored all productivity, process parameter, growth, product quality and metadata recorded by the process in a structured and accessible format. Additionally, the newly created MI_{CL} metric accessed the performance of 48 clones based on 20 selection criteria. Clone 8 was selected as the lead clone based on its MI_{CL} value of 0.89 and yielded a product titre of 5.2 g L^{-1} whilst meeting all the required product quality specifications. The subsequent ML identified a strong correlation with the pH of the culture and the end-point concentration of TSB. The correlation indicated cell cultures operating at the higher end of pH dead-band resulted in higher TSB concentrations in the range of 10%–15%. This information provided highly valuable insights and recommendations for the subsequent stability and scale-up studies. The final step of CLD_4 automated the generation of a report through a NGL algorithm. This automatically produced report aims to significantly reduce the burden on scientists and engineers through autonomous data analysis in addition to reducing their administrative duties by providing customisable and configurable electronic laboratory notebook entries. Furthermore, this methodology links the laboratory-based activities with the long-term business objectives of the biopharmaceutical company mitigating scale-up risks and ensuring production targets can be met at commercialisation.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

SG, CL, TS, and NB performed experiments and analysed the data. SG, HA, CL, NB, DH, JK, JS, and SF discussed data and helped with drafting article. DH, MC, RT, and SF managed project and acquired funding. All authors contributed to the article and approved the submitted version.

Funding

This research is associated with the joint UCL-AstraZeneca Centre of Excellence for predictive multivariate decision support tools in the bioprocessing sector and financial support from AstraZeneca and UCL is gratefully acknowledged. Financial support from the UK Engineering and Physical Sciences Research Council (EPSRC) and AstraZeneca for the Engineering Doctorate studentship for HA is also gratefully acknowledged (Grant Ref: EP/ S021868/1). This research is aligned with the EPSRC Future Targeted Healthcare Manufacturing Hub hosted by UCL Biochemical Engineering (Grant Ref: EP/P006485/1).

Conflict of interest

Authors CL, NB, TS, DH, JK, MC, RT, and JS were employed by AstraZeneca.

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Becker, F G, Cleary, M, and Team, R. M. (2007). "An architecture for data-to-text systems," in *Proceedings of the eleventh European workshop on natural language generation ENLG 07* (Saarbrücken, Germany: FKI GmbH).

Ben-Tchavtchavadze, M, Perrier, M, and Jolicoeur, M (2007). Small-scale bioreactor platform for bioprocess optimization. *Pharm. Eng.* 27 (5), 28–42.

Cawsey, A J., Webber, B L., and Jones, R B. (1997). Natural Language generation in health care. *J. Am. Med. Inf. Assoc.* 4 (6), 473–482. doi:10.1136/jamia.1997.0040473

Chen, C, Kim, L, Le, H, Daris, K, Neil, S, Stevens, J, et al. (2020). Methods for estimating the probability of clonality in cell line development. *Biotechnol. J.* 15 (2), 1900289. doi:10.1002/biot.201900289

Deshpande, R R, and Heinzle, E (2004). On-line oxygen uptake rate and culture viability measurement of animal cell culture using microplates with integrated oxygen sensors. *Biotechnol. Lett.* 26 (9), 763–767. doi:10.1023/B:BILE.0000024101.57683.6d

European Medicines Agency (2006). *Quality of biotechnological products: Derivation and characterisation of cell substrates used for production of biotechnological/biological products*. ICH Harmonised Tripartite Guideline.

Farid, S S., Baron, M, Christos, S, Nie, W, and Coffman, J (2020). Benchmarking biopharmaceutical process development and manufacturing cost contributions to R&D. *MAbs* 12 (1). doi:10.1080/19420862.2020.1754999

Fishburn, P C. (1967). Letter to the editor—additive utilities with incomplete product sets: Application to priorities and assignments. *Operations Res.* 15 (3), 537–542. doi:10.1287/opre.15.3.537

Frye, C, Deshpande, R, Scott, E, Francissen, K, Joly, J, Anthony, L, et al. (2016). Industry view on the relative importance of 'clonality' of biopharmaceutical-producing cell lines. *Biologicals* 44 (2), 117–122. doi:10.1016/j.biologicals.2016.01.001

Gagnon, M, Hiller, G, Luan, Y T, Amy, K, Defelice, J, and Drapeau, D (2011). High-end PH-controlled delivery of glucose effectively suppresses lactate accumulation in CHO fed-batch cultures. *Biotechnol. Bioeng.* 108 (6), 1328–1337. doi:10.1002/bit.23072

Garcia-Ochoa, F, and Gomez, E (2009). Bioreactor scale-up and oxygen transfer rate in microbial processes: An overview. *Biotechnol. Adv.* 27 (2), 153–176. doi:10.1016/j.biotechadv.2008.10.006

George, E, Titchener-Hooker, N J., and Farid, S S. (2007). A multi-criteria decision-making framework for the selection of strategies for acquiring biopharmaceutical manufacturing capacity. *Comput. Chem. Eng.* 31 (8), 889–901. doi:10.1016/j.compchemeng.2006.12.009

Goldberg, E., Driedger, N., and Kittredge, R. I. (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert* 9 (2), 45–53. doi:10.1109/64.294135

Goldrick, S. H, Bond, N J., Lewis, G, Kuiper, M, Turner, R, Suzanne, S., et al. (2017). Advanced multivariate data analysis to determine the root cause of trisulfide bond formation in a novel antibody–peptide fusion. *Biotechnol. Bioeng.* 114 (10), 2222–2234. doi:10.1002/bit.26339

Hartley, F, Walker, T, Chung, V, and Morten, K (2018). Mechanisms driving the lactate switch in Chinese hamster ovary cells. *Biotechnol. Bioeng.* 115 (8), 1890–1903. doi:10.1002/bit.26603

Hong, J K, Lakshmanan, M, Goudar, C, and DongLee, Y (2018). Towards next generation CHO cell line development and engineering by systems approaches. *Curr. Opin. Chem. Eng.* 22, 1–10. doi:10.1016/j.coche.2018.08.002

Huang, Y, and Kwiatkowski, C (2015). The role of high-throughput mini-bioreactors in process development and process optimization for mammalian cell culture. *Pharm. Bioprocess.* 3 (6), 397–410. doi:10.4155/pbp.15.22

Kim, B J, Diao, J, Michael, L., and Shuler, U. (2012). Mini-scale bioprocessing systems for highly parallel animal cell cultures. *Biotechnol. Prog.* 28 (3), 595–607. doi:10.1002/btpr.1554

Konakovsky, V, Clemens, C, Müller, M, Jan, B, Berger, M, Schlatter, S, et al. (2016). Metabolic control in mammalian fed-batch cell cultures for reduced lactic acid accumulation and improved process robustness. *Bioengineering* 3 (1), 5. doi:10.3390/bioengineering3010005

Krebs, L. E., DanielBowden, M. C, and Frye, C C. (2018). Effective and efficient characterization of Chinese hamster ovary production cell lines using automated intracellular staining and statistical modeling. *Biotechnol. Prog.* 34 (3), 570–583. doi:10.1002/btpr.2667

Le, H, Kabbur, S, Pollastrini, L, Sun, Z, Mills, K, Johnson, K, et al. (2012). Multivariate analysis of cell culture bioprocess data-lactate consumption as process indicator. *J. Biotechnol.* 162 (2–3), 210–223. doi:10.1016/j.jbiotec.2012.08.021

Li, F, Natarajan, V, Shen, A. Y, Kiss, R, and Amanullah, A (2010). Cell culture processes for monoclonal antibody production. *MAbs* 2 (5), 466–479. doi:10.4161/MABS.2.5.12720

Lin, P, Chan, K, Kiess, I. A., Tan, J, Shahreel, W, Wong, S Y, et al. (2019). Attenuated glutamine synthetase as a selection marker in CHO cells to efficiently isolate highly productive stable cells for the production of antibodies and other biologics. *MAbs* 11 (5), 965–976. doi:10.1080/19420862.2019.1612690

Millecam, T., Jarrett, A. J., Young, N., Vanderwall, D. E., and Della Corte, D. (2021). Coming of age of allotrope: proceedings from the fall 2020 allotrope connect. *Drug Discov. Today* 26 (8), 1922–1928. doi:10.1016/j.drudis.2021.03.028

Munro, T P., Kim, L, Le, H, Zhang, L, Stevens, J, Neil, S, et al. (2017). Accelerating patient access to novel biologics using stable pool-derived product for non-clinical studies and single clone-derived product for clinical studies. *Biotechnol. Prog.* 33 (6), 1476–1482. doi:10.1002/btpr.2572

Narayanan, H, Luna, M F., von Stosch, M, Polotti, G, Morbidelli, M, Butté, A, et al. (2020). Bioprocessing in the digital age: The role of process models. *Biotechnol. J.* 15 (1), 1–10. doi:10.1002/biot.201900172

Nielsen, R W, Tachibana, C, Hansen, N. E, and Jakob, R. W. (2011). Trisulfides in proteins. *Antioxidants Redox Signal.* 15 (1), 67–75. doi:10.1089/ars.2010.3677

Papathanasiou, M M., Quiroga-Campano, A L., Steinebach, F, Elviro, M, Mantalaris, A, and Pistikopoulos, E N. (2017). Advanced model-based control strategies for the intensification of upstream and downstream processing in MAb production. *Biotechnol. Prog.* 33 (4), 966–988. doi:10.1002/btpr.2483

Pollock, J, Coffman, J, Ho, S V., and Farid, S S. (2017). Integrated continuous bioprocessing: Economic, operational, and environmental feasibility for clinical and commercial antibody manufacture. *Biotechnol. Prog.* 33 (4), 854–866. doi:10.1002/btpr.2492

Priola, J J., Nathan, C, Baumann, M, Borth, N, Tate, C G., and Betenbaugh, M J. (2016). High-throughput screening and selection of mammalian cells for enhanced protein production. *Biotechnol. J.* 11 (7), 853–865. doi:10.1002/biot.201500579

Rameez, S, Mostafa, S S., Miller, C, and Shukla, A A. (2014). High-throughput miniaturized bioreactors for cell culture process development: Reproducibility, scalability, and control. *Biotechnol. Prog.* 30 (3), 718–727. doi:10.1002/btpr.1874

Ravindran, S., Singh, P., Nene, S., Rale, V., Mhetras, N., and Vaidya, A. (2019). "Microbioreactors and perfusion bioreactors for microbial and mammalian cell culture," in *Biotechnology and Bioengineering* (IntechOpen), 1–13. doi:10.5772/intechopen.83825

Roush, David J., and Lu., Yuefeng (2008). Advances in primary recovery: Centrifugation and membrane Technology. *Biotechnol. Prog.* 24 (3), 488–495. doi:10.1021/bp070414x

Sartorius, D. (2020). Multi-parallel bioreactors | sartorius. Available at: https://www.sartorius.com/en/products/fermentation-bioreactors/ambr-multi-parallel-bioreactors (Accessed February 9, 2020).

Silk, N. J., Denby, S., Lewis, G., Kuiper, M., Hatton, D., Field, R., et al. (2010). Fed-batch operation of an industrial cell culture process in shaken microwells. *Biotechnol. Lett.* 32 (1), 73–78. doi:10.1007/s10529-009-0124-0

Steinwandter, V, Borchert, D, and Herwig, C (2019). Data science tools and applications on the way to pharma 4.0. *Drug Discov. Today* 24 (9), 1795–1805. doi:10.1016/j.drudis.2019.06.005

Warr, S R. C., Betts, J P. J., Ahmad, S, Katy, V. N, and Finka, G B. (2013). Streamlined process development using the Micro24 bioreactor system. *BMC Proc.* 7 (S6). doi:10.1186/1753-6561-7-s6-p36

Whitford, W G. (2006). Fed-batch mammalian cell culture in bioproduction. *BioProcess Int.* 4, 30–40.

Wilkinson, M D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “Comment: The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Sci. Data* 3: 1–9. doi:10.1038/sdata.2016.18

Yang, Y, Velayudhan, A, Thornhill, N F., and Farid, S S. (2017). Multi-criteria manufacturability indices for ranking high-concentration monoclonal antibody formulations. *Biotechnol. Bioeng.* 114 (9), 2043–2056. doi:10.1002/bit.26329

Glossary

Agg	Aggregate
CHO	Chinese hamster ovary
CLD	Cell line development
$\mathbf{CLD_4}$	Cell line development Methodology for Industry 4.0
CPP	Critical process parameters
CQA	Critical quality attributes
Csv	Comma-separated values
$\mathbf{DO_2}$	Dissolved oxygen
DSP	Downstream processing
EP	End point
ETL	Extract transform load
FAbs	Fragment antibodies
Gluc	Glucose
GS	Glutamine synthetase
HT	High throughput
HPLC	High pressure liquid chromatography
Lact	Lactate
mAbs	Monoclonal antibodies
MCB	Master cell bank
MI_{CL}	Cell line manufacturability index
MSX	Methionine sulfoxamine
ML	Machine learning
MLR	Multiple linear regression
MTP	Microtitre plates
NLG	Natural language generation
PCA	Principal component analysis
PLS	Partial least squares
S.P	Set point
Std	Standard deviation
TSB	Trisulfide bond
VCD	Viable cell density
Viab	Viability