Data intelligence for process performance prediction in biologics manufacturing

Nishanthi Gangadharan[1], David Sewell[3], Richard Turner[3], Ray Field[3], Matthew Cheeks[3], Stephen G Oliver[4], Nigel K.H. Slater[1], Duygu Dikicioglu[1, 2**]

[1]Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, CB3 0AS, UK

[2] Department of Biochemical Engineering, University College London, London, WC1E 6BT, UK

[3] Biopharmaceutical Development, BioPharmaceuticals R&D, AstraZeneca, Cambridge, CB21 6GH, UK

[4] Department of Biochemistry, University of Cambridge, Cambridge, CB2 1GA, UK

[**] Address correspondence to: d.dikicioglu@ucl.ac.uk

Abstract

Despite the availability of large amount of data in bioprocess databases, little has been done for its retrospective analysis for process improvement. Historic bioprocess data is multivariate time-series, and due to its inherent nature, is incompatible with a variety of statistical methods employed in data analysis resulting in the lack of a tailored methodology. We present here an integrative framework of knowledge discovery tailored for handling historical bioprocess datasets. The pipeline successfully predicts process performance at harvest from an early time point, and robustly identifies the most relevant process parameters to model process performance. We present the utility of this pipeline on biologics manufacturing data from upstream bioprocess development for antibody production by mammalian cells. The proposed multi-model system that employs machine learning can predict performance at harvest after two weeks of operation with satisfactory accuracy employing data generated as early as on the sixth day of the culture.

Keywords

Biologics manufacturing; culture performance prediction; data mining; time-series analysis; machine learning; two-dimensional modelling

1. Introduction

The biologics manufacturing industry, due to the sensitive nature of its marketed products, is routinely subjected to uncompromising scrutiny by regulatory authorities to ensure product quality. The only practical way of maintaining minimum product variability is by running the processes on pre-defined and well-established trajectories[1]. This requires increased process understanding[2] coupled with advanced process monitoring and comprehensive information management for the resulting data[3]. Although these data hold important information on process dynamics, and thus could potentially assist increasing process knowledge[4], little has been done to harness the power of actionable intelligence through effective data utilisation.

An overarching protocol for bioprocess data mining is required to address this deficiency. Although a number of methods for data mining are discussed in the literature[5,6], a tailored workflow that is adapted to process datasets across different biomanufacturing operations is yet unavailable. This is attributed to the fact that bioprocess data does not fit into conventional continuous time-series regimes, as continuous time points are intervened by discontinuities across cultures and gaps in data collection. Moreover, heterogeneities with respect to timescale, data types, units and scaling all contribute to the distinct nature of these data[7], all of which contribute to the challenges in data mining and modelling.

A typical knowledge discovery process entails data pre-processing (handling of the missing data, visualisation, clustering, and feature selection) and processing (application of various computational methods to discover trends in the data) [5,8]. A critical bottleneck in modelling and analysis of bioprocess data is the incompleteness of the datasets or missing information in the dataset. In bioprocess datasets, the high complexity of mechanisms underlying the biological systems adds an additional layer of challenge to the missing data handling problem since the inherent behaviour of individual parameters and complex interactions between them are important considerations in handling this missing information[9]. There are several reasons why all the data are not recorded in

bioprocess databases in biomanufacturing settings, including: sensor breakdown during operation, inconsistent sampling rates, malfunctions in the data acquisition system, network outages, power blackouts, the use of incorrect or unrecognised format in logging the data, errors in data recording of online parameters creating consecutive gaps in the framework, glitches in the data management software resulting in corrupted file systems, and having samples flagged due to poor quality and subsequently excluded from storage; all cause gaps in the existing database[10,11]. Frequently, gaps arise due to characteristics inherent to the production process itself; the product concentration may not be measured on the initial days of culture due to low product concentration in the extracellular environment, based on ad hoc process knowledge, leading to missing data in early time points in the dataset. Similarly, switching to new data acquisition systems for complying with recent data management standards automatically promotes recording of more cultivation parameters, which renders earlier datasets incomplete when the data used for analyses span the period of change.

Conventional methods, such as imputation-based methods, likelihood-based methods, and similarity measures are all among the options available for handling the missing data existing in bioprocess databases as reviewed in[7]. Identifying the method that would be suitable for a particular dataset is essential, as capturing the complex trends in the temporal and static profiles of different parameters can assist in imputing values in a manner that minimises deviation from existing parameter behaviour and thus minimises bias.

Understanding the nature of the data, the missing data mechanism, the distribution, rate, and pattern of the missing information are all important considerations in selecting a suitable approach for any given dataset. Bioprocess datasets are highly heterogeneous owing to inter-batch variability, diverse production protocols, and the multifarious data collection techniques that are employed during the manufacturing process. They are often characterised by the presence of a combination of time-dependent and time-independent parameters. Those parameters that evolve over time and exhibit a

temporal profile are considered time-dependent parameters and those without such predictable fluctuations are considered time-independent parameters.

Upstream bioprocess development and production datasets generally exhibit a missing data mechanism known as Missing At Random (MAR), which would allow inferences to be made from the observed data, but are independent of the missing values. Occasionally these datasets also exhibit more than one missing data pattern or a combination of different missing data mechanisms and patterns[7]. Any method selected for this task should be sufficiently adaptable to accommodate this complexity. Unsurprisingly, a single method often fails to handle this problem, and versatile hybrid methods, which are combinations of methods that encompass the technical prowess of two or more different methods, are preferred. Such hybrid methods usually comprise of imputation-based approaches that combine statistical and machine learning methods to predict a close approximation to the missing value.

Unlike financing and marketing datasets with large-size and uniformly distributed data with many attributes, those arising from bioprocesses are rank-deficient, i.e. are characterized by relatively small offline data points (n) compared to the number of attributes (p), which makes it a typical $n << p$ problem[12]. Strong correlations exist between recorded parameters and the unit and scaling of the parameters in a bioprocess dataset can vary by several orders of magnitude[7].

Supervised variability in a selection of parameters such as the cell type, batch number, or the culture volume also play important roles in determining the production efficiency, and these translate to data analytics as having a large number of qualitative variables in the dataset along with other numeric information. Typically, multiple output parameters are monitored to characterise product quality in bioprocesses, stipulating the inclusion of more than one response variable in modelling the process outcome, rendering it a multivariate multiple regression problem. All bioreactors in a single batch of runs do not perform equally, and some may fail without any apparent reason. Although failed runs can create a considerable financial disadvantage[13], the data from these failed processes are

invaluable for analysis. Looking at patterns and trends in failed runs can help identifying the reasons for their failure, and these, in return, can be used as indicators of performance degradation, and can assist in building robust models for implementing control actions in the future to avoid such failures.

Here, we propose an integrative framework for knowledge discovery in bioprocess datasets, which addresses all these challenges and considerations. The framework itself is a conglomeration of methods that are handpicked or tailored to address the unique nature of data with such characteristics. We summarise the pre-processing steps of our analysis below in conjunction with some common recommendations on handling data of similar nature.

A novel dual-hybrid methodology, which uses a combination of Fuzzy C-Means clustering (FCM), Support Vector Regression (SVR), and Genetic Algorithm (GA) to deal with time-independent parameters, and a combination of Stineman interpolation and SVR to handle time-dependent parameters was developed to handle missing information in the dataset. The proposed strategy was then benchmarked against a routinely employed method for handling missing data that is integrated into a commercial software package, Simca, in aspects that were comparable. Pre-processing of the data was carried out by initially identifying problems in the data and selecting the suitable measures to address these problems to render set compliant with modelling[5,8]. Conserved behavioural patterns of parameters across cultures, or similarity in the behaviour of two or more parameters were first inspected by employing various visualisation techniques prior to formal model fitting, as previously recommended for handling voluminous and high-dimensional data[14]. Considering the time-series nature of the data, methods that allow a holistic view of the dataset employing superimposed or shared space techniques, which were reported to work better than juxtaposed or split space techniques[15], were used to ensure comparisons to be made in the same space. Shared space techniques make it easier to identify conserved parameter behaviour across cultures as opposed to having separate graphs for each culture that are less informative and makes comparison and hence identification of subtle behaviour patterns harder.

Bioprocess data sets are high dimensional, with typically 20-30 different parameters monitored over the course of time for a relatively few timepoints in order to ensure a high-quality process, which instigates the issue of 'curse of dimensionality'. Learning from a high dimensional feature space using finite number of data samples is a challenging concept in machine learning which typically requires enormous amount of training data. When this is not possible, features need to be selected to ensure superior trainability. Clustering of parameters was employed to reduce the dimensionality of the feature space and assist variable identification for model building by selecting a representative parameter from a cluster and eliminating those parameters bringing in the same information to models built, as recommended previously[16]. This was an especially challenging task in a time-series data setting due to the interdependence between parameter values across time, necessitating the use of dedicated tools[17,18] or developing suitable approaches as those proposed in this work.

The final pre-processing step for handling this high dimensional time-series dataset was to extract the optimally predictive feature subset of minimal size through parameter selection to identify accurate and independent predictors, yield interpretable estimates and to avoid overfitting in this sample-limited biomanufacturing problem[19–22].

The primary goal of mining time-series bioprocess data was to discover the dynamics unfolding over time, which govern the systems through the use of supervised methods such as Support Vector Machines (SVMs) [23–25]. Models can then be used to predict future process performance, provide a clear understanding of the significance of each parameter, and of how variations in each input parameter influence the final product outcome. We developed a two-step modelling strategy using machine learning to mine historical bioprocess data. Data retrieved from antibody production process development databases were used to test the performance of the proposed pipeline, and was shown to successfully evaluate and predict the final process performance from data collected at an early time point in the process.

The concepts introduced above are preamble to the detailed presentation of the pipeline in Data and Methods. The results are presented and discussed in three separate sections in Results and Discussion. The first part introduces a novel dual hybrid methodology for missing data handling in upstream bioprocess development and production data. This is followed by the introduction of a tailored data-mining pipeline and a two-step modelling strategy developed to analyse historical bioprocess data using data retrieved from antibody production process development databases as an example. The final section explores the compatibility of the proposed pipeline with data possessing heterogeneous characteristics such as combination of qualitative and quantitative parameters, failed and successful runs, multiple output parameters and differences in production scale.

2. Data and methods

2.1 Dataset

The dataset used for this study was extracted from AstraZeneca upstream process development and production databases. All data use Chinese Hamster Ovary (CHO) cell lines in the production of different antibody products. Data were available for 106 cultures representing a heterogeneous operational scale from bench-top (5L volume) to manufacturing (500L volume) garnered across a period of 7 years (2010-2016, both inclusive).

Each culture had a minimum of 25 parameters recorded offline for a period of up to 17 days: Culture Days, Elapsed Culture Time (ECT), Viable cell density (VCD), Total cell density (TCD), Average Cell Compactness (ACC), Average Cell Diameter (ACD), pH, Cell Viability, Elapsed Generation Number (EGN), Average Cell Volume (ACV), Osmolality, Cumulative Population Doubling Level (CPDL), concentrations of glutamine, glutamate, lactate, ammonium, glucose, sodium, potassium, and bicarbonate denoted as [Glutamine], [Glutamate], [Lactate], [NH_3], [Glucose], [Na^+], [K^+], and [HCO_3^-], respectively, temperature, pCO_2, pO_2, monomer content of final product (denoted as monomer percentage) and product concentration ([mAb]). The time-series dataset is normalised and anonymised to protect the proprietary rights (see Supplementary Data).

The repurposed subset of the data for categorical analysis (see section 4.3) comprised of 45 cultures, and included the following qualitative (categorical) variables: Endpoint Day and Midpoint Day (based on the harvest day and the day corresponding to halfway through the process, respectively), Culture Volume, Batch (different batches in which cultures were performed), and cell line. For the quantitative parameters, the single-value readings on the midpoint day and the harvest time point were used. 10% of the batches belonged to failed cultures (i.e. unsuccessful runs). The categorical variables are anonymised to protect the proprietary rights (see Supplementary Data).

2.2 Data Pre-processing

2.2.1 Dual-Hybrid Methodology for Handling Missing Data

Two different hybrid methods were implemented for handling missing data; a hybrid of Fuzzy C-Mean Clustering (FCM)[26], Support Vector Regression (SVR)[27] and Genetic Algorithm (GA)[28] to handle time-independent parameters and a hybrid of Stineman interpolation[29] and SVR to handle time-dependent parameters.

The hybrid method for gap-filling of the time-independent parameters was implemented as described before[30]. Independent, and numerically different, predictions were obtained by FCM and SVR for the same gap. Cluster number (c) values ranging from 2 to 10 and the weighting factor (m) values of 2 and 2.1 were tested in FCM. Cultures that did not have any missing data (i.e. the 'complete' fraction of the dataset) were used to train both algorithms. A GA was then used to minimize the difference between the SVR and the FCM outputs, and to optimise c and m values in FCM, which were then fed into FCM to estimate the missing values. Kernels for the Support Vector Machines (SVM) were selected based on the lowest Root Mean Square Error (RMSE) values. The generalisation of the SVM training was assessed using 10-fold cross validation. All algorithms were implemented in MATLAB R2017b by MathWorks.

For the time-dependent parameters, the missing data in a randomly selected 30% of the dataset was estimated using Stineman interpolation. The data were ranked in increasing order to avoid introducing negatives during implementation. The gap filled data from Stineman interpolation was used as a training set for the SVR algorithms, which then predicted the missing values in the remaining 70% of the dataset. The SVM test and training datasets were ensured to possess culture volume heterogeneity in order to avoid any potential bias. Stineman interpolation was implemented using 'imputeTS' package in R[29] (v3.4.3). SVMs were implemented in MATLAB R2017b.

The high-precision prediction values were rounded off to three significant figures to mimic the precision available for the empirical data in the final complete dataset following gap-filling.

2.2.2 Performance evaluation and benchmarking for gap-filling

Simca (Umetrics), a commercial software, which employs a Nonlinear Iterative Partial Least Squares (NIPALS) algorithm to handle missing data, was used for benchmarking. Simca's available tools needed to be utilised for evaluation. The dataset with gaps, i.e. the 'raw' dataset, and the datasets post-gap-filling, i.e. the 'complete' datasets were compared by Principal Component Analysis (PCA). Only the time-independent partition of the dataset could be used since PCA is unsuitable for time-series data with dependency across time points. The distribution of all principal components and the distance (i.e. similarity) between the cognate principal components in different datasets was evaluated based on the PCA loadings. The impact of imputation on the distribution of component values was evaluated by comparing the spread of the distribution for the datasets.

2.3 Visualisation

The dataset was partitioned based on culture parameters. 'mvtsplot' package in R[31] was used to generate heat maps of multiple time-series. Internal normalisation was employed with all other parameters at their default settings. The discretisation ranges (high/medium/low) are determined by

internal normalisation, and thus are different across cultures; a low range in one time-series may or may not be equal to the low range in a different time-series.

Time-series data of the failed runs (normalised values) were compared to those of successful cultures by visualisation.

2.4 Clustering

The dataset was partitioned into 17 subsets; one for each day of the culture. Parameter values were normalised prior to clustering to avoid parameters with higher values dominating the evaluation. Agglomerative hierarchical clustering in 'ClustOfVar' package in R[16] was used to identify the parameters that are assigned to the same cluster on each day of the culture. Clustering profiles were compared in order to identify the conservation of patterns across different days of culture (i.e. the temporal segments).

Cluster stabilities were tested by generating Adjusted Rand Index (ARI) values using the 'Stability' function (bootstrap value=1000) in 'ClustOfVar'. A plot of the dispersion of ARI was used to evaluate the most stable partition before selecting the cut-off value in the dendrogram.

The stable clusters were colour-coded using 'dendextend' package in R[32] at its default settings. The dendrograms were then partitioned into the desired number of stable clusters using the 'cutreevar' function. The central synthetic variable of each cluster was used to determine the similarity matrices of parameters in each cluster. Tanglegrams were constructed to compare the dendrograms using 'dendextend'. Correlation between clusters of different days was identified using the Cophenetic and Baker method in 'corrplot' package in R[33].

K-means clustering[34] was used to evaluate whether clustering was influenced by the categorical variables. The test involved repeated rounds of clustering by introducing a single qualitative parameter and varying the value of this parameter at a time while maintaining the values of all other

qualitative parameters constant. All qualitative parameters were tested to explore whether the clustering would suggest a direct relationship to the varying qualitative parameter in order to identify its impact.

For the repurposed dataset, clustering was performed (i) for the dataset containing both quantitative and qualitative variables using compatible arguments, (ii) converting all parameters to qualitative format and using the argument for qualitative variables, and (iii) converting all parameters to quantitative format, and using the argument for normalised quantitative variables. The results obtained for the endpoint dataset and the midpoint dataset from the three settings were then compared, and the parameters that remained clustered under all the conditions were identified.

2.5 Feature Selection

The correlation matrix was constructed between the quantitative parameters using the 'findCorrelation' function in 'caret' package in R[35]. The matrix was used to remove multicollinearity. VCD and EGN, expected to be correlated, were mathematically uncorrelated in this analysis due to a reformulation of the calculation of EGN by the industrial partner at some point in time during the collection of data. Both parameters were included in subsequent analysis.

Parameter importance was determined using five different algorithms implemented in R (Fig S1). [mAb] was the response variable and the selected features were the predictors in each case. The default settings of the packages were used in each algorithm unless specified otherwise: (i) 'Cforest' was coupled with 'varImp'[36]; 'Cforest' function ('party' package) for model building, 'varImp' function ('caret' package) for calculating variable importance (difference in out-of-bag (OOB) cross validation accuracy after permuting a variable and immediately after training) (*ntree*=501). (ii) 'Boruta' package was used with default settings[37]. (iii) 'MARS' function was coupled with 'evImp' function from 'Earth' package (*pmethod*=backward, *nprune*=total number of variables in the dataset, *nfold*=10); 'evImp' for estimating variable importance based on *nsubsets*, Residual Sum of Squares

(*RSS*) and Generalised Cross Validation (*GCV*) criteria, with variables included in more subsets and causes a large net decrease in the RSS and GCV being considered important[38]. (iv) 'parRF' function from 'caret' package (with 'trainControl' function, *number*=10, *repeats*=5) was coupled with 'varImp'; 'parRF' for generating models and 'varImp' for determining parameter importance[35]. (v) 'RFE' function was used from 'caret' package (*number of recursive*=10, number of features to be retained=the total number of variables in the dataset) and 'predictors' function was used to identify the predictors used in the final model[35].

The outputs from all five algorithms were sorted by their rank of importance taking the top five, ten or fifteen variables. A score was assigned to each parameter based on the number of occurrences of that parameter in that rank across algorithms. The scores were used to generate a compiled variable importance table.

The predictor importance for multiple response variables were handled either by (i) considering the response variables together (multi-response approach) and assuming a linear relationship between the output parameters, or (ii) considering one response variable at a time (single-response approach). Four of the five different feature selection algorithms were able to handle multivariate multiple regression problems; Boruta[37], MARS[38], parRF[37] and MRF[39], a dedicated software for multivariate regression problems. Qualitative parameters were converted to categorical values and the quantitative parameters were scaled for MRF. Consolidated variable importance from single-response approach and multi-response approach was determined for both the endpoint dataset and the midpoint dataset by comparing the outputs acquired from each algorithm.

3. Calculation

3.1 Time-series Modelling of Parameter Interactions

A two-dimensional multi-model system was designed to predict culture performance [mAb] at an early time point (Fig S2). The dataset was first partitioned based on parameters. For each parameter, SVMs

were trained to predict values on the 15th day of culture by progressively adding the readings of their values on earlier days. Models with the lowest RMSE of prediction for each parameter including [mAb] were selected for the next step (Fig 1a). The individual parameter values for the 15th day of the culture, which were predicted using the first-dimension models, were the inputs for the second-dimension model, which then predicted the [mAb] value on the 15th day of cultivation (Fig 1b). Parameters were included in the model progressively in their order of importance as suggested by the feature selection algorithms, until the prediction accuracy did not improve any further. The prediction RMSE of [mAb] from (i) the first-dimension model, and the second-dimension model using (ii) the observed parameter values on the 15th day and (iii) the predicted parameter values on the 15th day from the first-dimension models, were compared (Fig 2).

Having obtained two predictions on [mAb]; one from the first- and another from the second-dimension model that uses the predicted values, an optimisation model was trained such that the two values were used to predict a single value closer to the observed value (Fig 2). The model with the lowest RMSE was used as a standard to which other model outputs were compared.

The predictive success of the optimised model was evaluated based on a culture performance criterion established using the observed [mAb] values. For this purpose, the measured (i.e. the observed) [mAb] values were sorted in increasing order and the cultures were categorised as poor-performing, moderately performing, and high-performing. The [mAb] predictions and the culture performance assignments using the model with the lowest RMSE and the optimisation model were compared in order to identify the compromise in prediction accuracy (Fig 3a). The proposed strategy was benchmarked against a modelling methodology described by Charaniya *et al*., 2010[6].

3.2 Handling multiple response variables

SVMs were employed to predict the value of response variables using predictors based on the variable importance obtained from single-response approach and multi-response approach. RMSE was used

as the measure of success for the model's prediction ability while employing either the endpoint dataset or the midpoint dataset. The number of parameters required to predict the outcome and the time point at which a superior prediction could be achieved were inferred from the modelling results (Fig 3a). Coding was implemented in MATLAB R2017b.

4. Results and Discussion

4.1 Handling missing data in bioprocess datasets

4.1.1 Understanding the nature of the data and missingness

The data used for this study had readings for 22 parameters throughout 15 to 17 days of culture, making it a rank-deficient matrix. This specific dataset was comprised only of numerical variables and no categorical variables possessing a qualitative nature, whose numerical values could not be evaluated based on their weight but, rather, only indicated which category they belonged. The numerical values for different parameters varied by several orders of magnitude in the dataset (e.g. 6 orders of magnitude difference between pH and VCD). This issue was addressed by normalisation.

Some parameters in the dataset, such as ECT, EGN, VCD, TCD, [mAb], and CPDL, displayed a time-dependent profile, whereas the remaining parameters were time-independent, and did not follow a specific trend in time. Their values either remained within an acceptable range without displaying any trend, as was the case for cell compactness or the diameter of the cells; or these parameters were controlled to allow for fluctuations only within a dead band, as in the case of pH or Temperature of the culture (Fig S3). Some other parameters fluctuated in an unpredictable trajectory and did not follow any temporal patterns.

Viability, a calculated parameter that is a function of both VCD and TCD, was omitted from further analysis. Following the elimination of this correlated parameter, approximately 8.21% of the information was missing in this dataset. However, the distribution of the gaps within the dataset was not uniform (Fig 3b). Some parameters were characterised by more gaps than others were. For instance, one culture series had Osmolality measurements recorded only on three consecutive days

since 2010 (Fig 3b, S4a). Such inconsistent sampling created about 9.6% missing values in that parameter alone. A missing value in one of the calculated and thus mathematically correlated parameters was observed to create a domino effect on other parameters. Furthermore, if the values for the ensuing time points also depended on that missing value, this was observed to aggravate the problem. EGN measurements were such an example, where a missing EGN data on any given day resulted in data being not available for the following days until the end of cultivation as well as missing data in CPDL values, which were calculated from EGN (Fig 3b, S4b). In a standard culture of 15 days with 22 parameters, this alone created around 8.4% of the missing values in their respective parameters. Equipment failure at any point in time also created missing information in the dataset because readings were only available for those parameters that were measured manually, but not for the ones measured automatically such as ACC, ACD and ACV (Fig 3b, S4c).

The gaps in three parameters [mAb], Osmolality and [HCO_3^-] constituted about 47.41% of the total missing data. These were crucial parameters for the analysis and interpretation of the results, and they constituted about 13% of the total dataset. Despite the low fraction of missing data in the complete dataset, the variability of the extent and nature of gaps across different recorded parameters necessitated the use of advanced gap-filling approaches.

In addition to MAR missing data mechanism, the dataset used in this study exhibited a structured missing data pattern. This was because, as a rule, [mAb] was recorded from the 3rd or the 4th day of culture onwards, and then every alternate day until the 10th or the 11th day of culture (Fig 3b, S4d). Reducing the amount of sample drawn from the cultures during the early stages of cultivation would be desirable from an operational perspective. However, introducing a structured missing data pattern of a different mechanism, which could lead up to 1.6% gaps in the data for each culture, instigates a problem in data analysis, rendering the availability of suitable methodologies and tools limited for the handling of the gaps in the dataset.

4.1.2 Evaluation of Existing Methodologies for Handling of the Missing Data

A pilot study was carried out on data from 48 cultivations, which were not included in the dataset described; the cultivations were of a similar nature in terms of the sample collection and parameter detection methods employed. This 'toy' dataset was used to evaluate the performance of different methods for handling missing data. The nature of the data suggested that imputation-based methods should be suitable and, therefore, several multiple imputation (MI) methods were tested. Apart from MI, a few other methods such as Multivariate Imputation by Chained Equations (MICE) and Artificial Neural Networks (ANN) were also evaluated as possible options. The specific requirements of each method, the issues faced while employing them on bioprocess dataset, and the methods available to circumvent these limitations, as well as the results generated are displayed in Table S1. Although likelihood-based methods are known to generate good results, they were not adopted in this study due to their computational cost.

These preliminary analyses emphasized the importance of selecting hybrid methods for handling the missing data in bioprocess datasets. LGPImpute[40] was tested using data from a single representative culture. Due to confidentiality issues, the method was employed by the team who developed LGPImpute on an anonymised test set. Multiple results were generated from one-fold cross validation and ten-fold cross validation (Fig S5). The best result from ten-fold cross validation was acceptably successful in capturing the trends in time-dependent parameters in comparison to one-fold cross validation, and the best performing set was selected by visual inspection. However, this procedure can rapidly become infeasible as the size of the dataset increases, creating a serious problem for the utilization of the tool.

4.1.3 Development of the Dual-Hybrid Methodology

Understanding the shortcomings of the existing methodologies on the 'toy' dataset led to the development of a novel hybrid data processing pipeline to address these problems. FCM-SVR-GA was first applied to the original large dataset without distinguishing between the time-dependent and time-independent parameters. This method, utilising optimised cluster number and weighting factor

values in FCM predictions, allowed the imputation of values that were closer to the observed values for each parameter, thus ensuring less deviation from the original pattern. While SVR enforces systemic learning that is suitable for datasets that exhibits MAR missing mechanism, it alone cannot ensure high intra-column similarity. This is particularly obvious when dealing with parameters such as pH for which a value of 7.5 and 7.9 could elicit different responses in a bioprocess. By using FCM, which employs centroids to establish similarity, it achieves minimised intra-cluster dissimilarity. Thus, by combining SVR and FCM, we obtain values that are compliant in a systemic setting as well as ensuring high intra-column similarity. This could not be achieved by employing just one method. This strategy, relying on the concept of minimising intra-cluster dissimilarity, proved very advantageous in achieving successful imputations for the missing values observed in time-independent parameters. However, it failed to ensure a monotonic increase in the time-dependent parameters (Fig S6a).

In order to overcome this problem, the data documented for the time-dependent parameters were segregated from those reported for the time-independent parameters, and the two subsets were processed separately, leading to the development of a dual methodology.

The time-dependent data were processed separately using a different hybrid approach, which was a combination of Stineman interpolation and SVR. The performance of various interpolation approaches were also tested on the pilot dataset. Spline interpolation introduced negative values with large magnitudes on the early days of culture at the interpolation stage itself, further increasing the risk of introducing values that are even more negative in the following SVR prediction stage than those introduced during interpolation. Kalman smoothing, which uses a structural model fitted by maximum likelihood, was not successful in picking up trends (Fig S6b). Stineman interpolation generated satisfactory results by ensuring a monotonic increase in time-dependent parameters, as well as employing a robust interpolation function near abrupt spikes or steps, based on a second-degree interpolating polynomial (Fig S6a), and hence was adopted as a suitable method. Stineman interpolation performed equally well on the original dataset as on the pilot dataset.

In order to identify a suitable fraction of the dataset to be used for SVR training, a series of training set sizes ranging from 10% to 40% were tested. With increasing percentage of the training set size, the trend-capturing ability of SVMs, and hence the prediction accuracy, improved (Fig S6c) and fewer negative, thus unrealistic, values were imputed (Fig S7a). A 30% training set size was selected as it provided a reasonably good prediction, and also avoided the risk of overfitting due to large training set size. The training exercise was repeated with other randomised 30% fractions of the total dataset, and the number of negative values imputed was in the range, 21+2.6x (x= -3.46 to 12.3). The resultant dataset with no gaps (complete dataset), and the lowest number of negative values (12) was used for further analysis. Even though having negative values imputed for these parameters would not be interpretable physically, keeping these values as imputed introduced less bias than would employing a methodology that would impute values in such a way that it would not be able to capture the trends or could disrupt the pattern or the temporal data structure.

The absolute value of the average of these negative values imputed in all the randomised trials was found to be 43.1, with the highest negative value imputed being -0.009 and the lowest being -324.8. A substantial number of the negative imputed values had low weightage indicating that most of the negative imputed values were close to zero, thus avoiding any large bias (Fig S8). The incidence of negative values could be avoided by selecting a different SVM kernel, irrespective of the RMSE values. However, this study aimed to employ the kernel with the lowest RMSE value to ensure minimal deviation from the original pattern. The results showed that even in the randomised trials, the gap-filling protocol was able to provide satisfactory results (results for a representative culture are provided in Fig S9). Any predictable (i.e. satisfactory) patterns in parameter behaviour, such as those obtained here, can also be used to identify unexpected behaviour in the data, such as any intolerable deviations in the data pattern that are introduced by missing data handling methods.

As a final step in our dual strategy, the outputs from both hybrid methods were combined to produce a complete dataset with no gaps (see Supplementary Data).

4.1.4 Evaluation of the performance of the dual hybrid methodology against commercially available software

The performance of the methodology was then benchmarked against a commercially available data analysis and handling platform, Simca, which is routinely employed in data analysis by the bioprocess industry. Principal Component Analysis (PCA) was performed by Simca on the raw dataset after handling the missing information via its inbuilt methods, while the complete dataset had no missing information as it was already handled using the dual-hybrid methodology. PCA analysis was conducted separately for each day of the culture for both the complete and the raw datasets, and the distribution of the parameters was compared by overlapping the loadings plots for the two datasets. In order to compare the loadings for all days of culture, as suggested by PCA for both complete and raw datasets, a 'spread of distribution' was defined mathematically. The difference between the values of parameters in the culture lying at the lower and the upper bound was designated as the bandwidth, and this bandwidth was calculated for each day of the culture using the Euclidean distance of each data point from the origin. The bandwidth of the complete dataset (0.377) was very similar to, and even slightly narrower than that of the raw dataset (0.444). This indicated that the dual hybrid missing data handling mechanism did not result in the imputation of values that disrupted the original distribution of the components, and therefore was a satisfactory choice. The distribution of the loadings of parameters post-PCA in both the cases were comparable, as indicated by the coordinate proximity of the cognate data points, providing additional evidence that the dual-hybrid methodology did not introduce substantial bias (Fig S7b, S7c, Table S2).

Simca reports that it employs a modified version of the NIPALS algorithm to accommodate the missing values, and was recommended for use only when the missing data pattern was random, and not structured[41]. For structured missing patterns, such as the dataset employed in this work, this could have potentially created a bias in gap filling. The dual-hybrid missing data handling method proposed here avoided such bias in handling structured missing patterns. Furthermore, Simca is not fully

equipped with the methods and approaches, which are specifically suitable for handling time series datasets. Therefore, it cannot take into account the temporal nature of bioprocess datasets. The dual-hybrid methodology, on the other hand, is specifically designed to handle such scenarios. The dual-hybrid pipeline was more flexible, and thus more appealing, than its commercial alternative.

4.1.5 Concluding remarks

Handling of missing data is a crucial step in data analysis because incomplete databases are incompatible with a variety of statistical methods, making the data unusable. It is also a challenging step as inefficient and incompatible methodologies can introduce bias into the dataset. This dual-hybrid methodology, which made use of a combination of (i) FCM-SVR-GA for the analysis of time-independent parameters and of (ii) Stineman interpolation-SVR for the analysis of time-dependent parameters, handles the missing offline data in biologics manufacturing datasets successfully (Fig 4).

We also highlighted the factors that need to be considered in the selection of an efficient method for handling the missing information in bioprocess datasets (Fig 4). Although a wide variety of methods ranging from very simple techniques such as conventional arithmetic methods to highly complex and computationally heavy likelihood-based methods are available, the key elements were to be able to address the nature of both the dataset and the missing information. This, to a great extent, was shown to minimise bias.

The analyses proposed in this methodology were conducted blindly until cell culture experts, who did not take part in the data analysis themselves, assessed the outcome. The performance of the methodology did not rely on *a priori* knowledge of the dataset, nor on the physical or biological meaning of the parameters. Regardless, the cell culture scientists regarded the gap-filling as acceptable. This dual-hybrid pipeline can be applied more widely than just within the domain of the biologics manufacturing process industry, for data where heterologous time-dependency was observed in the parameters.

4.2 Conjunctive Framework for Historical Bioprocess Data Analysis

4.2.1 Data Pre-Processing – Visualisation

Visualisation was employed after gap-filling as a preliminary step to identify any possible inherent patterns in parameter behaviour (Fig 3a). Because the shared space technique for the visualisation of large numbers of time-series can create clutter problem compounded by limited colour acuity of human visual system[15], time-series were initially aggregated and the aggregates were visualised[42]. We extracted the information on the empirical distribution of values in individual time-series as well as the average trend across all time-series to identify higher level trends and patterns in the data. The values of each time-series were discretised using quantiles of the time-series values, and then each distinct category (low, medium, and high) were represented using different colours (purple, grey, and green, respectively). This discretisation acted as a simple smoother in visualising the variation in the data[14].

Heat maps were used to visualise the recurring patterns in parameter behaviour across cultures. Although the cultures were conducted years apart and the cultured cells produced different antibody products, process parameters possessed a level of similarity in behaviour (Fig S10-S14). Upon visual inspection, only ACD and ACV were observed to behave similarly. Although [K^+], osmolality, TCD (but not VCD) and [mAb] were also observed to behave reasonably alike, no direct biological inference could be made from this relationship. However, these recurring trends were useful identifiers to distinguish irregularities in parameter behaviour in the dataset.

4.2.2 Data Pre-Processing – Clustering

Clustering of parameters whose values change over time as in time-series datasets would create inaccurate results because parameter relationships change over time (Fig 3a). To address this challenge, similarity relationships between parameters were identified on each day of the culture, and how those relationships changed over the course of the process were investigated collectively in this analysis. Different cultures had identical values for a number of parameters on the first day and the last day of the cultivations rendering clustering analysis conducted on those days impractical. Thus,

the analysis excluded the first and the last days of the cultures. The correlated or dependent parameters that clustered together on any other day of the culture were identified. Parameters that cluster together at different time points did not always form true clusters nor essentially had similar behaviour; these clusters are designated as false clusters from this point forward. Due to this problem, it became inevitable to identify which parameters formed true clusters in this dataset, as using representative parameters from false clusters could result in render the predictive performance of constructed models poor. Stability and homogeneity were used as two separate measures to evaluate the "trueness" of a cluster.

The stability of partitions (i.e. the dendogram cut-off) was tested and the adjusted RAND index was used to identify clusters that remained intact when the dataset was resampled by bootstrapping. The index value greater than 0.85 indicated highly stable clusters, which were likely to be "true" clusters (Fig S15). The 2^{nd}, 3^{rd}, 5^{th}, 7^{th}, and the 12^{th} days of the cultivations were identified to have fewer clusters with more parameters assigned to each cluster than what was observed for the clustering of the remaining days. In the rest of the analysis, the dendrograms had to be partitioned further to achieve stability indicating similar behaviour for only a few parameters at a time.

The homogeneity of a cluster was investigated as an additional measure to ensure that all variables in a cluster brought in the same information, and it was said to be at its highest when all the quantitative parameters were correlated or anti-correlated with the central synthetic variable. This was observed to be the case for the following pairs of parameters on all the days of culture: TCD & VCD, ACD & ACV, and pCO_2 & [$HCO3^-$]. All these correlations or anti-correlations were higher than 0.75, with the exception of one instance where the correlation of pCO_2 and [HCO_3^-] with the central synthetic variable were identified as -0.649 and -0.669, respectively, on the 12^{th} day. The absolute difference in the correlation values of these parameters with their respective central synthetic variable was close to zero indicating the tightness of the clustering of these parameter pairs (Fig S16a).

Once the true clusters on each day of culture were determined, the next step was to identify how each parameter interacted with other parameters on any day of the culture. For this purpose, a similarity matrix of parameters was constructed for each cluster employing the parameter correlation values. The parameter pair TCD & VCD displayed nearly 100% similarity until the 9th day of the culture, and this similarity decreased progressively starting from the 10th day onward (Fig S16b). The similarity measures served as an example to demonstrate how the relationship between parameters changed over time. Although TCD and VCD remained clustered together on all days, the similarity between the two parameters decreased as culture time progressed due to a decrease in VCD as the population reached stationary phase. Such patterns could even lead to two parameters to cluster together only on specific days, and not on others. For example, ACC, ACV and ACD clustered together until the 8th day of the culture, after which, ACC diverged from this behaviour and separated from the cluster, except for on the 12th day. On the 12th day, more parameters were populated in each cluster and a less stringent partition criterion needed to be employed than on any other day (Fig S15).

Tanglegrams were constructed to compare the dendrogram of each day with every other day, hence by comparing the clustering profiles across days (Fig S17-30). Several parameter combinations were conserved across different days of culture. A correlation matrix was generated in order to visualise the similarity between the clusters of each day of culture and of the other days (Fig S16c). Clusters of the 2nd and 3rd days, the 4th and 5th days, the 5th and 6th days, the 5th and 7th days, the 5th and 8th days, 7th and 8th days and the 11th and the 13th days had more than 50% similarity. VCD and TCD, ACD and ACV, and [$HCO3^-$] and pCO2 clustered together throughout the cultivation indicating similarity in their temporal behaviour. The identification of these parameter pairs was elemental for dimensionality reduction as a final step in data pre-processing.

4.2.3 Data Pre-Processing – Feature Selection / Handling Multicollinearity

Once the similarities between different parameters were identified, we then explored whether these similarities were caused by existing mathematical correlations between the parameters, or whether

any uncorrelated parameters displayed similar behaviour highlighted by the clustering analysis (Fig 3a). Correlations were calculated and those having coefficient values $\geq |\pm 0.75|$ were identified from the correlation matrix (Fig S31). Culture days and ECT, TCD and VCD, ACV and ACD, as well as [HCO_3^-] and pCO_2 were correlated (correlation coefficients of 0.998, 0.926, 0.920, and 0.762, respectively). Culture days and ECT were also correlated with Osmolality (respective correlation coefficients of 0.759 and 0.755), and with EGN (respective correlation coefficients of 0.886 and 0.888). [mAb] was highly correlated with culture days, ECT and [Glutamine]. All mathematical correlations between parameters indicated by these results were removed; parameters ECT, Culture Days, pCO_2, TCD and ACD were excluded, and the remaining 17 independent parameters were employed in further analysis.

Although EGN and CPDL were expected to be identified as correlated in a conventional sense, this relationship was not evident in the correlation matrix (0.158) as indicated by the values available in the dataset. This could possibly be an error in data retrieval from the databases or any differences in parameter relationship formulations undisclosed due to propriety rights; as far as the analysis was concerned, the two parameters were mathematically uncorrelated, and thus both were included in further analyses.

4.2.4 Data Pre-Processing – Feature Selection: /Identifying parameter importance

We use ‘feature selection’ in this context to describe a methodology, which helps reduce the number of parameters that would be used in the subsequent stages. Unlike in a conventional feature selection procedure, where we obtain a linear or non-linear transformation of the raw parameters, here we are only identifying parameters that are important for prediction and manually eliminating the ones that are not. This helps in understanding the influence of each independent parameter on the culture in case a control mechanism needs to be implemented in the future as it might be challenging to decide control action on transforms.

The feature selection algorithms employ different modelling methods to evaluate parameters in a systemic setting before prioritising one parameter over the other, and thus provides a reliable measure of importance. Making use of this principle, the remaining independent parameters were then sorted in the order of importance based on the results obtained by employing five different feature selection algorithms. The parameters were ranked in the order of decreasing importance as follows: [Glutamine], [K^+], Osmolalilty, VCD, Temperature, EGN, [Lactate], [Glutamate], ACV, [HCO_3^-], [Na^+], CPDL, ACC, pH, [NH_3], pO_2, and [Glucose] (Table S3). Although in a biological sense glucose is an important factor for CHO cell survival, [Glucose] was identified as the least important feature. This is likely to be due to the nature of the process where the cultures were ensured to maintain [Glucose] at around 5-8g/L *via* process control. Since the glucose feeding regime was not available as an additional parameter in this analysis, the controlled parameter profiles did not deliver further information about the effect of glucose on the response variable ([mAb]). Similarly, pH, which, too, was a controlled parameter, ranked lower than most parameters because in all these cultures pH was maintained at a value between 7±0.15. However, temperature, yet another controlled parameter, was among the top five most important parameters. Different temperature profiles were observed among the cultures; 34.9% of the cultures had a switch from 36.5°C to 33°C at sometime within the 17-day culture period. It was likely that that the models picked up a variation in the [mAb] in response to this temperature shift, designating temperature to be an important parameter.

[Glutamine] was identified as the most important parameter in this analysis based on its effect on the response variable. Indeed, glutamine is a key parameter for cell cultures since it acts as an alternate source of energy[43]. On the other hand, [NH_3], a by-product of glutamine metabolism, was ranked among the least important parameters. [NH_3] was reported to have inhibitory effects on cell growth at concentrations above 8mM[44]. Indeed, less than 2% of the [NH_3] measurements in the dataset were above this threshold value providing further evidence that below inhibitory concentrations, it did not affect the culture performance, [mAb], extensively.

pO_2 was another parameter that did not have a major contribution to the process outcome although oxygen is indispensable for energy generation in animal cell cultures. However, the effects of its limitation were reported to become evident only when dissolved oxygen dropped below 5% of air saturation[45]. The dataset did not have any cultures with severe oxygen limitation, and therefore there was not sufficient information to evaluate the effect of oxygen availability on the response variable, rendering the parameter ineffectual.

4.2.5 Data Processing – Strategy for Modelling Parameter Interactions

A novel two-dimensional multi-model system was developed to predict process performance at an earlier time point (Fig 3a). The first dimension accounted for the intra-cluster variability by incorporating the temporal behaviour of parameters. It generated multiple models for each individual parameter and [mAb], which could predict the value on the 15^{th} day of the culture using their values on earlier days. The second dimension accounted for the inter-cluster variability and generated a single model that used the minimum number of parameters to predict the [mAb]. The input dimension for the models were decided by the pre-processing steps.

We used this two-dimensional model to extract process knowledge from manufacturing data pertaining to upstream bioprocess development and pilot-scale runs of mammalian cell lines expressing different monoclonal antibodies collected at five different scales over the course of seven years from different projects (Fig 3c).Two breakpoints were observed in prediction accuracy suggested by the first-dimension model Root Mean Square Error (RMSE) values: (i) before the sixth day of the culture, and (ii) after the 11^{th} day (Table S4). The parameter values on the 15^{th} day of the culture could be predicted with reasonable accuracy using the data before the first breakpoint. Predictions made using the values beyond the second breakpoint had either similar or improved performance. For example, RMSE for the prediction of [Glutamate] were 1.772 and 1.535 using its values up to the first and the second breakpoints, respectively, exhibiting only slight improvement in prediction ability. The marginally reduced prediction accuracy achieved at an early time point is an acceptable trade-off

based on the observations made on all 17 parameters. This analysis indicated that [Glutamine], VCD, Temperature, EGN, [Na^+], CPDL, ACC and [Glucose] could be predicted with low RMSE on the sixth day of the culture; [Glutamate] and [HCO_3^-] could be predicted on the fifth day, [K^+], Osmolality, pH and [mAb] on the fourth day, and [Lactate], ACV, [NH_3] and pO_2 as early as on the second or the third day of the process (Table S4). Although prediction is possible on later days than the one identified in this exercise (Day 6), it might only improve the accuracy marginally and without any obvious advantages that one would expect from waiting for longer to make a prediction. This would also drive the cost of running the bioreactor up for every additional day it is kept running. Furthermore, the prediction accuracies of the models at the time of the breakpoints were tested on independent and randomly sampled datasets that have not been employed in the model construction, thus indirectly implying similar success rates for any additional datasets that would be included in the analysis.

The first-dimension model was used to predict the final day values of all 17 parameters using their values at or before the first breakpoint. These values were then employed in the second-dimension model to predict the [mAb] value. The second-dimension modelling showed that a combination of the top 14 important parameters provided the lowest RMSE and hence highest accuracy of [mAb] prediction (Table S5). Inclusion of the parameters [NH_3], pO_2 and [Glucose] into the models did not improve prediction accuracy further. These parameters, which were also highlighted by the feature selection algorithms as having little to no effect on the response variable, were excluded from model construction.

The resultant 14-parameter model with the lowest RMSE was then used to predict [mAb] values of 15^{th} day of the culture employing two different types of predictor values; the observed values of all 14 parameters and the predicted values of the same parameters given by the first-dimension models. The RMSE of prediction for these models were compared to select the better performing model (Fig 1). The first-dimension model gave the best prediction accuracy, i.e. the lowest RMSE. However, the second-dimension model using the predicted parameter values from the first-dimension had an RMSE, which was 103% higher than that for the first-dimension model.

This discrepancy could possibly have been caused by either of the two issues: (i) inherent problems associated with the second-dimension model structure, or (ii) the input data fed into the model. In order to assess the validity of these hypotheses, the second-dimension model was used with observed parameter values as input variables instead of the parameter values predicted from the first-dimension model. For this new test case, the calculated RMSE was similar to that of the first-dimension model. This indicated that it was not the model structure, but the data that was fed into the model that affected prediction accuracy, indicating potential for improvement for the proposed modelling approach *via* optimisation. Although a simple alternative solution to this problem would have been to use the observed values on the 15^{th} day as input, this is not desirable, since the objective of this modelling exercise is to predict the process performance from an earlier time point than the 15^{th} day, when those cognate values would actually be available. However, using the observed parameter value with second dimension model helped us decide on an acceptable level of RMSE as a target for the predictive models.

The objective of the optimisation was to train an SVM to take into account values from the first dimension and the second dimension and predict a single value, which was close to the observed [mAb] value. In this way, optimisation acted as an additional step to ensure superior predictive capability of the developed models. This optimisation step lowered the RMSE by 52.07% (from 1363.96 to 710.21). The new RMSE was comparable to the lowest RMSEs attained in predicting [mAb] and these values for the "successful" models remained within ±5.7% of one another. Data from seven cultures, which were not employed for test nor training purposes in the study, were then used to evaluate the performance of the optimised model. The [mAb] predictions made by the optimised model were classified as poor-, moderately, or high-performing, as in the case of the observed culture performance values (Table S6, Fig S32a). Two of these cultures (28%) were predicted as moderately performing as indicated by the classification of the optimised model output, whereas, these cultures were poor-performing as indicated by their observed values. The optimised model that used values from before the first breakpoint could not distinguish the moderately performing cultures from the

poor-performing cultures in these two instances (Fig S32b). This zone where the predictions could not be correctly assigned into observed performance classifications was discerned as the grey zone.

The optimised model was then used with the values from the second breakpoint to test whether the misclassified performance assignments could be corrected. Both cultures were correctly classified as poor-performing when the values from the second breakpoint were used (Table S6). These results suggested that if the predicted [mAb] fell within the grey zone, the decision making needed to be postponed until the second breakpoint before any sound process evaluations can be made.

The only other model-based analysis of bioprocess data available in the literature employed an overall similarity matrix-based approach for predicting process performance[6]. Therefore, the modelling strategy proposed in this paper was compared to the single other available methodology. The weights of each parameter were recalculated and assigned as their Spearman's correlation coefficient with the product titre. The importance of the parameters was observed to be substantially different from those obtained in this study *via* feature selection algorithms (Table S7). This difference was caused by the fact that feature selection algorithms assigned importance by considering the effect of each parameter on the final product titre in the presence of all the other parameters in a systemic setting. In contrast the weights were assigned by only considering the effect of a single parameter on the final product titre ignoring the interactions between parameters in this strategy[6]. In the real-world scenario, all parameters interact with each other, and consequently affect the final product titre. Unlike any existing schemes, the methodology proposed in this work can successfully account for such variability. Furthermore, unlike SVMs that are trained to predict culture performance based on overall similarity matrices[6], the modelling scheme proposed here is adaptable since it takes into account the systemic interaction of all the different parameters involved, and also accounts for both inter-cluster and intra-cluster variability.

4.2.6 Concluding remarks

Although general principles of data mining has been established for decades, the unique nature of data arising from specific processes pose a substantial challenge that necessitates the development of tailored methodologies for each application. Unlike its other conventional counterparts, data from biologics manufacturing databases are not continuous in a conventional sense as continuous time points are followed by discontinuities across cultures. They are also incoherent to a certain extent as the data comes from experiments conducted by different people, using different sets of equipment, over a span of several years. The methodology proposed above (Fig 5) is specifically tailored for data arising from the bioprocess industries. The process was split into simpler steps with clear objectives. The output of each step then became the input for the next, making it a process chain. Prediction accuracy of the final model was dictated by the proficiency of the methods employed during different stages of data pre-processing and processing and facilitated back-tracking. The proposed method enables trainability and expansion. Furthermore, the platform presents a pipeline to process rank deficient multivariate time-series datasets, which is currently a venue in very much need for improvement. Also, the methodology is platform-independent, which enables the user to select a platform of their choice and provides them control over the knowledge discovery process. Although the individual methods employed in this framework are themselves not novel, there are no specific frameworks currently available that addresses the unique nature of bioprocess data. Furthermore, the methodology described here is not limited to bioprocess datasets alone, but can be applied to any multivariate time-series dataset with discontinuities, heterogeneous information, parameters with different magnitudes, strong correlations, rank deficiency and MAR as missing data mechanism, rendering it a pipeline with broad application areas not limited to industrial biotechnology.

4.3 Handling Categorical Variables, Multiple Process Outputs and Failed Runs in Bioprocess Datasets

4.3.1 Dealing with failed runs, qualitative parameters and multiple response variables

The dataset investigated in the first two subsections was also categorically classified by four parameters, which were not expected to contribute to the success or failure of the process, and should

therefore be excluded from modelling. However, the impact of these parameters on the dataset needed to be evaluated, and any potential differences implicated by any of 'Culture Volume', 'Batch', 'Day of Harvest' and 'Cell Line' needed to be identified. In the event of any differences highlighted due to these parameters, the runs that fall into specific classes should be analysed separately from the rest of the dataset (Fig 3d). The possibility of any distinguishable patterns embedded within the data was tested using K means clustering[46]. Clustering results indicated that none of the above four differences reflected in the intrinsic nature of the dataset, indicating no impact on performance (Fig S33).

Even though a standardised protocol would be followed in handling all bioreactors, this does not prevent problems arising during the course of operation. These failures often are beyond the control of the operator. These "failed" runs were particularly interesting and extremely valuable as it is very rare that information was systematically recorded and is available on failed systems in an industrial setting. Time series data of failed cultures (3) in this dataset were used to compare them with time series data of successful cultures (104) in the original dataset to identify any differences in parameter behaviour. Time-series profiles of the failed runs (normalised) were investigated using a control by taking the average value at each time point from normalised successful cultures, which could then be employed as early indicators of process failure. A comparative investigation of the time-series profiles of these critical quality parameters of failed and successful bioprocesses provided potential insights into the arrangement of events that caused failure. Our analysis associated failed runs with either one or more of the following features including an increase in ACC, high glutamine concentration, inefficient glucose utilisation at the beginning of the culture, depletion of Na^+ ions, and an increase in pO_2 (Fig S34-S37).

One of the challenges that need to be addressed when working with datasets containing both quantitative and qualitative parameters is that data of different nature require different treatment protocols. The mathematical approaches that are available for handling quantitative parameters are

more diverse than those that are available for handling qualitative parameters or categorical variables. Clustering was once such aspect. The clustering method used for this analysis successfully distinguished between quantitative parameters and qualitative parameters. A comparison of treating the categorical variables as numerical values or numerical values as categorical values yielded different results. When all parameters were treated as qualitative, there was low variability in the culture data, which affected the stability of clusters, resulting in fewer clusters with more number of parameters in them for different timepoints corresponding up to halfway through cultivation denoted as the midpoint set, and up to the harvest represented by the endpoint set (Figure S38). The clustering results on the midpoint dataset and on the endpoint dataset both indicated the presence of three highly stable clusters. VCD and TCD, ACD and ACV, HCO_3^- and pCO_2 remained clustered in all scenarios for both midpoint and endpoint sets (Table S8).

The dataset did not hold high similarities between the parameter profiles across different cultures, and therefore the employed methodology, which utilises a homogeneity measure of variables to a central synthetic variable (as the clustering criterion), worked satisfactorily. Datasets with highly similar profiles for different parameters are more likely to suffer from cluster instability; this would particularly impose a risk in the analysis of categorical variables with too few input arguments, as the bootstrapping conducted to evaluate the stability will lead to the selection of data subsets that have highly similar (or exact) values. Under such circumstances, the stability analysis should be restricted to the evaluation of quantitative parameters alone with less likelihood (not being used in the statistical sense) of encountering this problem.

Following the identification of the parameters that remain clustered and display similar behaviour at halfway into the process operation time and at the harvest, we then explored whether the correlation between the parameters evolved over time and demonstrated that the correlation between the parameters did not display any substantial change (Figure S39). Nearly 25% of the parameter interactions were identified to be highly correlated ($|\pm 0.70|$). In light of this, TCD, [Glutamate], ACC,

ACD, [K^+], [Lactate], [NH_3] and [HCO_3^-] were removed from the midpoint set while TCD, ACD, [Na^+], [Lactate] and [HCO_3^-] were removed from the endpoint set.

4.3.2 Modelling Parameter Interactions

Bioprocesses can have multiple output parameters measured to ensure product quality, which necessitates process performance to be described by more than one such parameter in the dataset. Multivariate multiple regression models are used to address such problems. One of the most important challenges imposed by working with datasets containing multiple output parameters is that many feature selection algorithms are incompatible with multivariate multiple regression problems. Furthermore, most of those algorithms that are compatible with assuming a model structure where a linear relationship exists between the output parameters, which may not necessarily hold in reality. The four algorithms that were successfully employed all intrinsically assumed linear relationship between response variables to build models and then to decide on the predictor importance. Predictor importance was also calculated for each individual response variable separately (Table S9). The parameter importance for both the midpoint and the endpoint datasets was evaluated in conjunction with the modelling and reiterated as deemed necessary.

The maximum number of variables required to produce the lowest RMSE were identified using both single-response and multi-response modelling approaches (Fig 3g, S40a). The results indicated different RMSE values for the models generated using the multi-response approach and the single-response approach (Fig 3g, S40b) for both endpoint dataset and the midpoint dataset. This indicated that the variable importance determined by assuming linear relationships between the response variables was different from when no such relationship between them was assumed, strongly in favour of the concerns discussed above as to the interpretation of modelling based on misleading assumptions. In this case, our analysis validated the presence of an underlying non-linear relationship between the response variables, i.e. the performance measures, if any existed at all. This enforced us to employ the single-response approach to determine variable importance. It is important to note that

the dataset for which this approach was successful had a limited number of predictors. A similar assumption may not hold true if additional parameters were to be added to the dataset, or in the event of working with other datasets with a large number of predictors. Therefore, the selection of a suitable method for handling the multivariate multiple regression problem should be left to the discretion of the data scientist.

An evaluation of the models that employed the midpoint dataset and the endpoint dataset indicated that the midpoint dataset was a better predictor of process performance than the endpoint dataset, utilising fewer parameters (denoting Cell Line, VCD, Culture Volume and Batch number) while achieving superior predictions. This was a highly desirable outcome as the main objective of the exercise was to predict the harvest performance of the process employing the data collected at an early time point. Interestingly, some of the categorical variables were identified as important for predicting process output, indicating the centrality of categorical variables to carry important information. Ignoring them altogether thus can lead to substantial loss of information.

4.3.3 Concluding remarks

In this exercise, we observed that qualitative or categorical parameters had substantial impact on bioprocess performance. Substandard treatment of data possessing such nature may yield grossly inaccurate results. Furthermore, many biomanufacturing protocols dictate the monitoring of more than one parameter to determine the performance of the process, and this imposes an additional challenge on modelling as the task then evolves into handling of a multiple-response multivariate problem, which we address here. Dealing with failed runs in an attempt to identify the deviations from the expected parameter behaviour, was proposed as an early indicator of process failure. The analysis suggested that a method for identifying the upper and lower bounds of expected parameter behaviour at each time point in a successful run, thus proposing a systematic way of defining a dynamic design space to comply with the Quality by Design requirements. Comparison of the expected performance

with those of the failed processes also provide interesting insights into understanding the root causes of failure.

5. Conclusions

Using this methodology (Fig 5), we were able to predict the final performance of the biologics manufacture upstream processes successfully utilising only 65% of the process parameters within the first half of the process duration. Missing data, even when they constituted a substantial fraction of the total dataset, could successfully be imputed without disturbing the original distribution of the data (Fig 3e). We could distinguish poor-performing, moderately performing, and high-performing cultures, and also identify "grey zones" of prediction, where the decision on the process performance should be delayed until a later time point (Fig 3f). We were also able to accurately predict the process performance at harvest halfway through the duration of the process by employing only four process parameters. Some of these were categorical variables, thus indicating the relevance of the information contributed by them. Process performance was predicted based on single performance parameters or a combination of them. We showed that the dependent relationship between these output parameters was the primary challenge in the decision-making involved in model construction (Fig 3g). The analyses demonstrated the adaptability of our integrative framework for positively influencing the modelling of bioprocess datasets of different structure and preventing any technical *faux pas* due to the utilisation of incompatible statistical procedures.

Streamlining the data exploitation process brings us one step closer to building automated, and self-adapting, systems for bioprocess control that ensure reproducibility and improved process performance. In this work, we present a platform-independent computing framework for extracting information from bioprocess datasets, which provides the user with flexibility and control over the knowledge discovery process. Each step of the process chain in the workflow can be evaluated independently for backtracking purposes. Furthermore, the machine-learning methods used in the workflow allow sufficient scope for trainability and expansion with accumulating data. We adopted

this framework to mine datasets of varying inherent structures to demonstrate its versatility in successful handling different datasets. Although, the modelling approach was conducted on the dataset that we described above, the application itself is not limited to this dataset, nor to bioprocesses for that matter. Data arising from different sources, which have a discontinuous time series structure that render them incompatible with standard time series analysis, can rely on this framework for knowledge discovery. This framework provides a workflow comprised of compatible approaches, which were demonstrated to work successfully in conjunction to one another, in order to address various challenges associated with datasets, which suffer from similar challenges to those described above. This approach is especially useful in circumstances when a black box predictive model is required owing to limited mechanistic understanding of the system. Due to the inherent complexity, bioprocess data makes a good example for the broader problem that is discussed. The models built using this methodology will enable the acquisition of process knowledge from previously unexploited datasets to lead to improvement in performance through efficient process control for both bioprocess industries and beyond through the development of digital monitoring platforms with self-adaptability and controllability.

Acknowledgements

The authors thank Fabio Manoel Franca Lobato and Ibrahim Berkan Aydilek for discussions on LGPImpute and the FCM-SVR-GA hybrid algorithm, respectively. This work was funded by MedImmune as part of the University of Cambridge – MedImmune Beacon Collaborative project and by Biotechnology and Biological Sciences Research Council, Grant number: BB/ K011138/1 to SGO and NKHS and BB/L013770/1 to DD. DD gratefully acknowledges the funding from the Leverhulme Trust and the Isaac Newton Trust (ECF-2016-681).

References

[1] Gnoth S, Jenzsch M, Simutis R, Lübbert A. Process Analytical Technology (PAT): batch-to-batch reproducibility of fermentation processes by robust process operational design and

control. J Biotechnol 2007;132:180–6. https://doi.org/10.1016/j.jbiotec.2007.03.020.

[2] Pörtner R, Platas Barradas O, Frahm B, Hass VC. Advanced Process and Control Strategies for Bioreactors. Curr. Dev. Biotechnol. Bioeng. Bioprocesses, Bioreact. Control., Elsevier Inc.; 2016, p. 463–93. https://doi.org/10.1016/B978-0-444-63663-8.00016-1.

[3] Craven S, Whelan J. Process Analytical Technology and Quality-by-Design for Animal Cell Culture, 2015, p. 647–88. https://doi.org/10.1007/978-3-319-10320-4_21.

[4] Craven S, Becken U. A Quality-by-Design Approach to Upstream Bioprocess Interrogation and Intensification. Eng J 2014.

[5] Charaniya S, Hu W-S, Karypis G. Mining bioprocess data: opportunities and challenges. Trends Biotechnol 2008;26:690–9. https://doi.org/10.1016/j.tibtech.2008.09.003.

[6] Charaniya S, Le H, Rangwala H, Mills K, Johnson K, Karypis G, et al. Mining manufacturing data for discovery of high productivity process characteristics. J Biotechnol 2010;147:186–97. https://doi.org/10.1016/j.jbiotec.2010.04.005.

[7] Gangadharan N, Turner R, Field R, Oliver SG, Slater N, Dikicioglu D. Metaheuristic approaches in biopharmaceutical process development data analysis. Bioprocess Biosyst Eng 2019;42:1399–408. https://doi.org/10.1007/s00449-019-02147-0.

[8] Pratama I, Permanasari AE, Ardiyanto I, Indrayani R. A review of missing values handling methods on time-series data. 2016 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2016 - Proc., 2017. https://doi.org/10.1109/ICITSI.2016.7858189.

[9] Del Rio-Chanona EA, Cong X, Bradford E, Zhang D, Jing K. Review of advanced physical and data-driven models for dynamic bioprocess simulation: Case study of algae–bacteria consortium wastewater treatment. Biotechnol Bioeng 2018;116:bit.26881. https://doi.org/10.1002/bit.26881.

[10] Imtiaz SA, Shah SL. Treatment of Missing Values in Process Data Analysis 2008.

https://doi.org/10.1002/cjce.20099.

[11] Severson K, Molaro M, Braatz R, Severson KA, Molaro MC, Braatz RD. Principal Component Analysis of Process Datasets with Missing Values. Processes 2017;5:38. https://doi.org/10.3390/pr5030038.

[12] Rommel S, Schuppert A. Data mining for bioprocess optimization. Eng Life Sci 2004;4:266–70. https://doi.org/10.1002/elsc.200420059.

[13] Crater JS, Lievense JC. Scale-up of industrial microbial processes. FEMS Microbiol Lett 2018;365. https://doi.org/10.1093/femsle/fny138.

[14] Peng RD. A Method for Visualizing Multivariate Time Series 2008;25:1–17.

[15] Javed W, Member S, Mcdonnel B, Member S, Elmqvist N. Graphical Perception of Multiple Time Series 2010;16:927–34.

[16] Chavent M, Kuentz V. Journal of Statistical Software ClustOfVar : An R Package for the Clustering of n.d.;VV.

[17] Montero P. TSclust : An R Package for Time Series Clustering 2014;62.

[18] Fidaner IB, Cankorur-Cetinkaya A, Dikicioglu D, Kirdar B, Cemgil AT, Oliver SG. CLUSTERnGO: A user-defined modelling platform for two-stage clustering of time-series data. Bioinformatics 2015;32:388–97. https://doi.org/10.1093/bioinformatics/btv532.

[19] Tsagris M, Lagani V, Tsamardinos I. Feature selection for high-dimensional temporal data 2018:1–14. https://doi.org/10.1186/s12859-018-2023-7.

[20] Hmamouche Y, Casali A, Lakhal L, Hmamouche Y, Casali A, Lakhal L, et al. A Causality Based Feature Selection Approach for Multivariate Time Series Forecasting To cite this version : HAL Id : hal-01467523 2018.

[21] Liu T-Y, Trinchera L, Tenenhaus A, Wei D, Hero AO. Jointly Sparse Global SIMPLS Regression 2014.

[22] Yoo W, Mayberry R, Bae S, Singh K, Peter He Q, Lillard JW, et al. A Study of Effects of MultiCollinearity in the Multivariable Analysis. Int J Appl Sci Technol 2014;4:9–19.

[23] Song H, Zhang Z, Song H. Analyzing Multiple Multivariate Time Series Data Using Multilevel Dynamic Factor Models Analyzing Multiple Multivariate Time Series Data Using Multilevel Dynamic Factor Models 2014;3171. https://doi.org/10.1080/00273171.2013.851018.

[24] Sree Dhevi AT. Imputing missing values using Inverse Distance Weighted Interpolation for time series data. 6th Int Conf Adv Comput ICoAC 2014 2015:255–9. https://doi.org/10.1109/ICoAC.2014.7229721.

[25] Niu X, Yang C, Wang H, Wang Y. Investigation of ANN and SVM based on limited samples for performance and emissions prediction of a CRDI-assisted marine diesel engine. Appl Therm Eng 2017;111:1353–64. https://doi.org/10.1016/j.applthermaleng.2016.10.042.

[26] Ahmed MN, Yamany SM, Mohamed N, Farag AA, Moriarty T. A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data. IEEE Trans Med Imaging 2002;21:193–9. https://doi.org/10.1109/42.996338.

[27] Basak D, Basak D, Pal S, Ch D, Patranabis R. Support vector regression. NEURAL Inf Process Lett Rev 2007:203--224.

[28] Abdella M, Marwala T. The use of genetic algorithms and neural networks to approximate missing data in database. IEEE 3rd Int. Conf. Comput. Cybern. 2005. ICCC 2005., IEEE; n.d., p. 207–12. https://doi.org/10.1109/ICCCYB.2005.1511574.

[29] Moritz S, Bartz-Beielstein T. imputeTS: Time Series Missing Value Imputation in R. R J 2017;9:207–18.

[30] Aydilek IB, Arslan A. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Inf Sci (Ny) 2013;233:25–35. https://doi.org/10.1016/J.INS.2013.01.021.

[31] Peng RD. Package “mvtsplot” Title Multivariate Time Series Plot. 2015.

[32] Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics 2015;31:3718–20. https://doi.org/10.1093/bioinformatics/btv428.

[33] Taiyun Wei M, Taiyun Wei cre A, Simko aut V, Levy ctb M, Xie ctb Y, Jin ctb Y, et al. Package “corrplot” Title Visualization of a Correlation Matrix. 2017.

[34] Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. Appl Stat 1979;28:100. https://doi.org/10.2307/2346830.

[35] Max Kuhn Contributions from Jed Wing A, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. Package “caret” Title Classification and Regression Training Description Misc functions for training and plotting classification and regression models. 2019.

[36] Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics 2008;9:307. https://doi.org/10.1186/1471-2105-9-307.

[37] Kursa MB, Rudnicki WR. Feature Selection with the **Boruta** Package. J Stat Softw 2010;36:1–13. https://doi.org/10.18637/jss.v036.i11.

[38] Stephen Milborrow M. Multivariate Adaptive Regression Splines 2019. https://doi.org/10.1214/aos/1176347963.

[39] Maintainer R, Rahman R. Package “MultivariateRandomForest” Type Package Title Models Multivariate Cases Using Random Forests 2017. https://doi.org/10.1093/bioinformatics/btw765.

[40] Resende DCO De, Santana ÁL De, Lobato FMF, L. JJAF, Lobato FMF. Time series imputation using genetic programming and Lagrange interpolation. An Do XLVIII SBPO Simpósio Bras Pesqui Operacional 2016. https://doi.org/10.1109/BRACIS.2016.30.

[41] Philip C Nelson BR, Eng M. THE TREATMENT OF MISSING MEASUREMENTS IN PCA AND PLS

MODELS. 2002.

[42] Van Wijk JJ, Van Selow ER. Cluster and Calendar based Visualization of Time Series Data. n.d.

[43] Zhang F, Sun ÆX, Yi ÆX. Metabolic characteristics of recombinant Chinese hamster ovary cells expressing glutamine synthetase in presence and absence of glutamine 2006:21–8. https://doi.org/10.1007/s10616-006-9010-y.

[44] Schneider M, Marison IW, von Stockar U. The importance of ammonia in mammalian cell culture. J Biotechnol 1996;46:161–85.

[45] Heidemann R, Lütkemeyer D, Büntemeyer H, Lehmann J. Effects of dissolved oxygen levels and the role of extra- and intracellular amino acid concentrations upon the metabolism of mammalian cell lines during batch and continuous cultures. Cytotechnology 1998;26:185–97. https://doi.org/10.1023/A:1007917409455.

[46] Raykov YP, Boukouvalas A, Baig F, Little MA. What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm. PLoS One 2016;11:e0162259. https://doi.org/10.1371/journal.pone.0162259.

Supplementary Materials

Supplementary Figures

Supplementary Tables

Supplementary Data: Time-series dataset sourced from upstream bioprocessing of mammalian cell cultures producing monoclonal antibodies

Figure Legends

Figure 1 Methodology for the implementation of the first- and second-dimension modelling (a) First dimension modelling. For (i) complete dataset of 106 cultures, each containing 15 days (D) of readings

for 17 parameters (P) and [mAb]: (ii) Separate tables with 15 days of readings for each independent culture parameter (P1 to P17) and [mAb] in 106 cultures, (iii) train SVMs to predict parameter values on the 15th day of culture from the values collected on previous days by progressively adding them into the model, for each parameter and [mAb], and (iv) select the model with the lowest RMSE for the prediction of values for each P and [mAb] on the 15th day of culture. 70% of the dataset was used for training and the rest for testing. **(b)** Methodology for the implementation of the second-dimension modelling. For (i) complete dataset of 106 cultures, each containing 15 days (D) of readings for 17 parameters (P) and [mAb]: (ii) Train SVMs to predict [mAb] using the "important" parameters (starting with P1 and progressively adding the rest) in subsequent models, and (iii) select the model with the lowest RMSE as the best model for prediction of [mAb]. 70% of the dataset was used for training and the remaining fraction for testing.

Figure 2 Methodology for the implementation of the two-dimensional model compilation and model optimisation Parameters, days, the predicted values and the observed values are denoted as P, D, 'Pred' and 'Obs', respectively. For (i) observed values of each parameter on the 15th day of the culture: (ii) Predict values of each P on 15th day of the culture using their respective values on earlier days and the best model from the first dimension, (iii) use [mAb] prediction model from the second dimension to predict the [mAb] value on the 15th day of the culture and (iv) use the best [mAb] prediction model from the first dimension to predict the [mAb] value on the 15th day of the culture. (v) [mAb] values on the 15th day of the culture predicted using observed parameter values on the 15th day of the culture, (vi) the [mAb] value on the 15th day of the culture predicted by the second-dimension model using parameter values for the 15th day of the culture as predicted by the first-dimension model, and (vii) [mAb] values on the 15th day of the culture predicted by its values on previous days and the best first-dimension [mAb] prediction model will be used to (viii) compare models based on RMSE of prediction. Model optimisation will be carried out by (ix) training an SVM to predict the observed [mAb] values from (vi) and (vii).

Figure 3 Missing data handling in bioprocess data from antibody producing CHO cell cultures. Schematic representation of the important factors that need to be taken in to account while selecting an appropriate method for handling missing data in datasets arising from different sources are presented. This challenge is coupled with the proposed schema of the dual hybrid methodology used for handling the missing data in bioprocess datasets. Branching represents splitting of the dataset based on the parameter types for implementing different missing data handling methods.

Figure 3 Implementation of the pipeline (a) Proposed framework with the pipeline embedded within an efficient biomanufacturing process monitoring and control architecture **(b)** Sample bioprocess data structure displaying rank deficiency and missing data (toy data). Imputation challenges due to missing measurement creating a gap in a calculated parameter (brown), equipment failure creating missing values for the whole culture (dark blue), structured missing pattern due to process-related reasons (pink), value dependency on the availability of data at first time point (light blue), always missing a value at first time point (green), values for related parameters missing (yellow). **(c)** Data distribution across different years and culture volumes. **(d)** Data distribution across different projects and culture volumes. **(e)** The impact of imputation on the distribution of PCA loadings as indicated by spread of the distribution for gap-filled (red) and raw (blue) datasets. Radius represents the loadings on a polar coordinate; angles in the radial plot are neither scaled to size nor have any mathematical relevance. **(f)** Culture performance for biologics manufacturing. Zones of poor, moderate and high performance (increasing performance from left to right) were identified based on the observed data. Concentration values marking the initiation and the end of each zone are denoted for reference. “Grey zone” (blue shading) between the poor and moderately performing cultures marks a novel zone identified by model predictions that necessitate delays in decision-making. **(g)** Model predictions for non-linear multivariate data on biologics manufacturing. The performance by employing only four process parameters or all parameters in the models, which predict all response variables individually (single-response) or together (multi-response), are compared based on Root Mean Square Errors (RMSE) of

model predictions. The predictions made halfway through the process (Midpoint set) and those made at the end of the process (Endpoint set) are also compared.

Figure 5 Proposed pipeline for mining bioprocess datasets STAGES shows the fundamental sequence of any typical data analysis process and METHODOLOGY summarises the cognate methodologies employed in this pipeline within the context of bioprocess data mining.