Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Engineering

A decade in review: use of data analytics within the biopharmaceutical sector

Matthew Banner[1], Haneen Alosert[1], Christopher Spencer[2], Matthew Cheeks[2], Suzanne S Farid[1], Michael Thomas[1,3] and Stephen Goldrick[1]

There are large amounts of data generated within the biopharmaceutical sector. Traditionally, data analysis methods labelled as multivariate data analysis have been the standard statistical technique applied to interrogate these complex data sets. However, more recently there has been a surge in the utilisation of a broader set of machine learning algorithms to further exploit these data. In this article, the adoption of data analysis techniques within the biopharmaceutical sector is evaluated through a review of journal articles and patents published within the last ten years. The papers objectives are to identify the most dominant algorithms applied across different applications areas within the biopharmaceutical sector and to explore whether there is a trend between the size of the data set and the algorithm adopted.

Addresses
[1] Department of Biochemical Engineering, University College London, Gower Street, London WC1E 6BT, UK
[2] Cell Culture Fermentation Sciences, Biopharmaceutical Development, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK
[3] London Centre for Nanotechnology, University College London, Gordon Street, London WC1H 0AH, UK

Corresponding author: Goldrick, Stephen (s.goldrick@ucl.ac.uk)

Current Opinion in Chemical Engineering 2021, **34**:xx–yy

This review comes from a themed issue on **Biotechnology and bioprocess engineering: mechanistic and data-driven modelling of bioprocesses**

Edited by **Cleo Kontoravdi** and **Colin Clarke**

https://doi.org/10.1016/j.coche.2021.100758

Introduction

The biopharmaceutical sector has seen significant improvement in bioreactor design, instrumentation and analytical technologies over the last 10 years. One of the most widely adopted technological advancements within the sector is the utilisation of high-throughput automated micro-bioreactors, enabling the parallelisation of experiments using a fraction of the liquid volume required for laboratory-scale experiments [1]. Additionally, there has also been a large push from regulatory bodies to adopt process analytical technology (PAT) for improved monitoring and control of biopharmaceutical processes [2,3]. Recent advances include also the application of omics, such as transcriptomics, proteomics, metabolomics and fluxomics in the sector, allowing for a better understanding of the intracellular workings of cellular processes [4]. These recent advancements have resulted in the generation of larger and more diverse data sets requiring more specialised statistical and modelling tools for efficient data analysis. The process of data analytics can be described as the analysis of raw data to make useful conclusions about the information provided. The application of these tools enables the true potential of data to be fully harnessed for improved process understanding and more informed decision-making [5] facilitating optimisation of existing biomanufacturing processes and hence increased product quantity and quality. The increased use of these advanced sensors and automated bioreactors will play a key role in the advancement of the sector towards the core principles of Industry 4.0 [6]. However, to ensure these 'smart factories' of the future can deliver improved and automated process control, the need for advanced data analytics is paramount.

The process of data analytics is predominantly carried out using machine learning (ML) algorithms. ML is a very broad term which can be defined as the generation of algorithms to perform tasks based on rules learnt from the data rather than explicitly programmed by the user. The application of these algorithms within the biopharmaceutical sector can provide additional insights and identify patterns within data sets enabling improved decision making for process optimisation. Certain ML algorithms such as neural networks (NN) and random forests (RF) are typically considered computationally intensive which has previously hindered their widespread application, however there has been a recent surge in the application of these algorithms due to the increase in computing power available and improvements in algorithm efficiency [7]. ML algorithms have seen also an increase in popularity and impact across other sectors outside the biopharmaceutical sector, such as analysing markets to predict better performing stocks in the finance sector [8] and targeted display adverts in the advertising sector [9]. Traditionally within the biopharmaceutical sector, a

subset of ML referred to as multivariate data analysis (MVDA) has been used to examine variable interactions and is preferred over univariate and bivariate techniques due to its ability to analyse multiple variables and minimise false inferences [10,11]. MVDA algorithms such as partial least squares (PLS) and linear regression (LR) are commonly used in the sector. Examples include the application of a PLS algorithm for the prediction of amino acid concentrations by analysing Raman spectroscopy in mammalian cell cultures [12] and the application of LR for root cause analysis of product quality deviations in therapeutics proteins [13].

With the growing body of ML algorithms, it can be a challenge to decide which algorithm to implement. Within this study, an evaluation of data analysis techniques was carried out over the period of 2010–2020 to identify any trends related to the utilisation of specific algorithms within bioprocessing. Initially, the paper reviews the rising prevalence of data analysis techniques in the sector by evaluating literature and patents published between the period of 2010–2020. A more in-depth analysis of the most cited literature from the period of 2015–2020 was carried out to identify the most dominant algorithms, the size of the data sets utilised within each model and their application area within bioprocessing. This paper outlines a clear increase in the application of data analytics within the biopharmaceutical sector and demonstrates the wide range of algorithms implemented to analyse these complicated bioprocessing data sets. The increased use of these digital methodologies demonstrates the shift of the sector towards Industry 4.0, which envisages fully automated and autonomous biomanufacturing operations.

Material and methods

The methodology required to produce this analysis is split into two parts: a systematic search of all scientific literature containing the key phrases: MVDA, ML and biopharmaceuticals, assessed between the period of 2010–2020 and an in-depth analysis of the most impactful journal articles assessed from the period of 2015–2020.

Literature search (2010–2020)

Google Scholar was queried to search scientific literature published from 1 January 2010 to 31 December 2020 that mentioned the key terms ML or MVDA and biopharmaceutical. The exact search terms for each query were:

- MMVDA - ("Bioprocess" OR "Biopharmaceutical") AND ("Multivariate data analysis" OR "Multivariate analysis")
- ML - ("bioprocess" OR "biopharmaceutical") AND ("machine learning" OR "artificial Intelligence")

Similarly, for patents, Google Patent was queried using the aforementioned search terms to gather patents that contained the key terms MVDA or ML and biopharmaceutical between 1 January 2010 and 31 December 2020. For both patents and journal articles, the total number of results were recorded for each year.

To evaluate the overall relevance factor, the number of patents and journal articles for both MVDA and ML were quantified using the formula defined as the 'Relevance Index', R_i:

$$R_i = 0.5\,(J_{Si}) + 0.5\,(P_{Si}) \text{ for i } = 2010, \ldots\ldots, 2020 \quad (1)$$

where J_{Si} is the standardised number of journal article results of the ith year and P_{Si} is the standardised number of patent results of the ith year.

Where the standardised values are defined as:

- Journal articles
  $$J_{Si} = J_i/\max(J) \text{ for i } = 2010, \ldots\ldots, 2020 \quad (2)$$
  where J_i is the number of journal article results at the ith year.
- Patents
  $$P_{Si} = P_i/\max(P) \text{ for i } = 2010, \ldots\ldots, 2020 \quad (3)$$
  where P_i is the number of patent results at the ith year.

Literature analysis (2015–2020)

To assess the most impactful articles published over the last five years, the key journal articles were defined as those that had the highest average citation number per year since publication. Whilst the use of 'citation number' to evaluate the impact of a journal article has its limitations, they are considered to be a good measure of scientific impact and relevance [14]. For the purpose of this paper, this metric enables potential trends to be identified in the biopharmaceutical sector. The key journal articles were identified using the software Harzing's Publish or Perish [15]. Harzing's Publish or Perish scrapes the search results from Google Scholar using the same search terms mentioned previously and tabulates and cumulates the data. For each year between 2015 and 2020, the top 10 most cited journal articles for each year about data analytics were collected. From these 60 journal articles, the metadata was manually extracted. This included:

- the algorithm used,
- the number of experiments used in the analysis,
- the number of variables used in the analysis,
- the application area in terms of problem domain (e.g. fault detection) and process stage (e.g. upstream

processing (USP), downstream processing (DSP) or other manufacturing areas).

The data collected only considered the top 60 journal articles which defined the number of variables and number of experiments analysed within their study. Where journal articles evaluated multiple algorithms for their data analysis, the best performing algorithm was selected for this paper.

Data analysis and visualisation

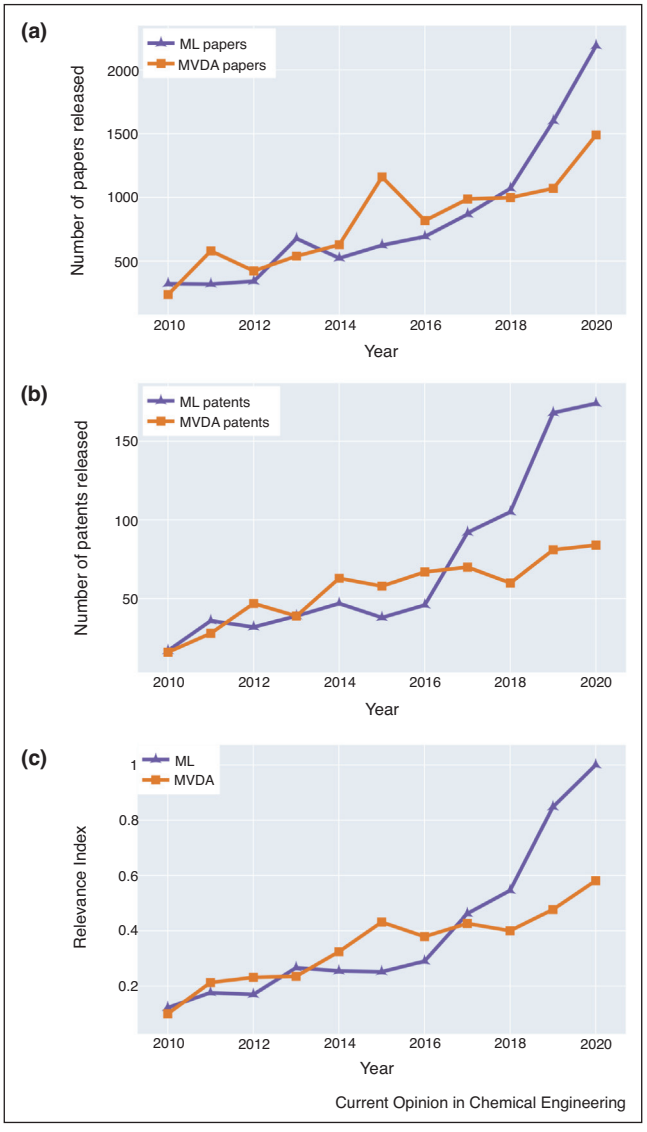
The metadata was imported, analysed and visualised using R 4.0.2 and Python 3.8.3.

Results and discussion

Prevalence of data analytics adoption in the biopharmaceutical sector

As the industry undergoes a digital transformation with the scale and complexity of the data sets increasing, there is a need to better understand how data analytics is used within the biopharmaceutical sector. To quantify the prevalence of MVDA and ML, a literature and patent search was carried out across the period of 2010–2020 using the search terms defined in the material and methods section with the aim of identifying potential trends in the utilisation of these data analysis techniques. The distinction in search terms was needed as some algorithms are exclusively referred to as MVDA while others are labelled exclusively as ML, therefore these search terms provide broader understanding of the utilisation of data analytics across the sector. Figure 1 compares the results recorded for MVDA and ML in terms of number of patents and journal articles released over the period of 2010–2020. To simplify the interpretation of these two metrics, a standardised plot of Relevance Index (Eq. (1)) is presented consolidating both the impact of journal articles and patents (Figure 1c). This metric enables a single evaluation of these search terms and helps determine their utilisation in the sector more broadly.

Figure 1



Prevalence of ML and MVDA data analytics adoption in the bioprocessing and biopharmaceutical sector during the period of 2010–2020 in terms of **(a)** journal articles, **(b)** patents and **(c)** the relevance index. The specific search criteria are provided in the Materials and Methods. The orange squares represent the search terms related to MVDA, and purple triangles represent those related to ML.

Figure 1a–c show a positive trend demonstrating that ML and MVDA are becoming more utilised in the biopharmaceutical sector. The utilisation of ML within the last five years has sharply increased by 250% for journal articles and 357% for patents. However, MVDA has only increased by 28% for journal articles and 44% for patents in the same period. The rise in the application of these techniques is most likely due to an increase in the complexity, size and format (e.g. images) of data available for analysis. As a result, more advanced algorithms are needed to analyse these complex data sets. Additionally, due to the recent advances in computational power and access to high-performance machines, the ability to train and validate ML algorithms on large complex data sets is now much easier. It is worth noting that some techniques that have historically been referred to as MVDA, such as PCA and PLS, are now often labelled as ML due to the increasing popularity of the latter term. The increase in published patents related to data analytics shown in Figure 1b may suggest that ML and MVDA are now becoming more adopted by industry. A recent patent by De Kok *et al.* demonstrated the value of implementing various non-linear ML techniques such as RFs and k–nearest neighbour to predict the performance of large scale systems through experimental design of small scale systems [16]. Furthermore, a recent patent utilising a PLS

algorithm was published by Berry and Moretto for the analysis of Raman spectra to predict a number of culture parameters including glucose, lactate and ammonia for bioreactors ranging from 0.1 L to 100 000 L [17].

An additional consequence of the recent advances in computational power is the increased availability of commercial software that lowers the knowledge barrier to implement ML models. However, this can potentially lead to incorrect applications without full understanding of the detailed assumptions, considerations, and model limitations. For example, models must be able to account for large quantities of noise and variability that is common in biological systems due to their stochastic nature [18]. Another important consideration for model building is the interpretability of the model being utilised. Some ML algorithms, such as NN, are typically much more difficult to interpret due to their complex structures. This is particularly important in drug discovery where the selection of new treatments without justification may hinder regulatory approval. In these cases, further analysis and evaluation of a model's performance and robustness is needed before implementation [19]. If these conditions are not considered, the risk of inaccurate predictions increases resulting in possible process failure and hence, financial loss.

To improve the adoption of ML algorithms within the biopharmaceutical sector, organisations need to ensure the correct architecture of their data storage facilities. Full and easy access to all available data within a structured and queryable database such as a data lake or warehouse will significantly simplify the application of ML algorithms. Some of the more complex ML algorithms often require larger data sets and more computational power is required to build and train these models. Therefore, a company's data infrastructure becomes a priority.

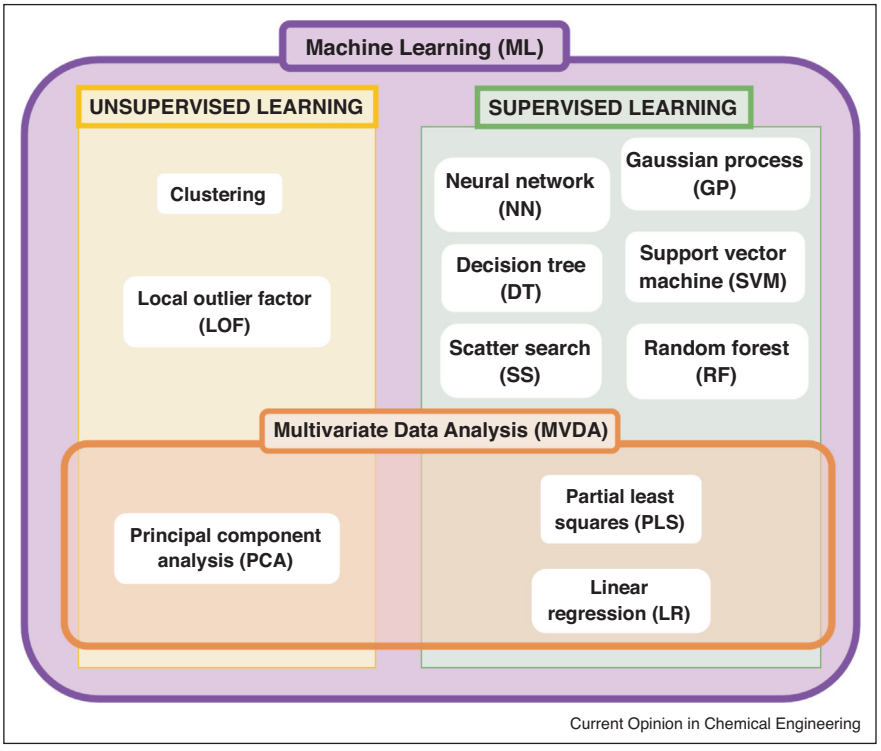
Overall, ML has become a widely adopted technique within the biopharmaceutical sector based upon our analysis of journal articles and patents used in the sector between the period 2010–2020. The increased utilisation of both MVDA and ML algorithms within the biopharmaceutical industry is likely to continue. It is most likely to accelerate as the sector further adopts these algorithms for better decision-making within clinical and commercial manufacturing, however, there is little guidance in the sector as to how much data is needed for each technique or which algorithm will perform better. There are numerous examples where data analytics techniques are directly compared. An extensive analysis by Mendez *et al.* compared multiple non-linear ML and MVDA algorithms to classify ten clinical metabolomics data sets. They concluded that there was no general improvement in predictability between the non-linear and linear algorithms utilised. It was reiterated that '*a model is only as good as the data that is used to train it*' suggesting that the data set used to train models is an equally important factor as the algorithm performing the analysis [20•]. As such, an analysis on data size and algorithms applications is needed to identify potential trends.

Classification and application of ML in the biopharmaceutical sector based on recent literature

To better understand the most prominent application areas and the most dominant algorithms used within the biopharmaceutical sector over the period of 2015–2020, an in-depth evaluation of the most cited journal articles was conducted. The 10 most cited journal articles based upon average number of citations per year was recorded, which resulted in a total of 60 journal articles. Within each of these journal articles: the specific algorithms utilised, the size of the data sets analysed and whether the authors classified the algorithm as either MVDA or ML was documented. Patents were not considered in the in-depth analysis due to a lack of databases that could be queried for information. A summary of all the algorithms utilised within these 60 journal articles during this period is shown in Figure 2, with the learning method of each algorithm defined as either supervised or unsupervised. Supervised learning techniques require a labelled training set to build the model and establish relationships between the inputs and outputs of the given system [21]. Alternatively, unsupervised learning uses unlabelled data and focuses on identifying patterns within the data with the purpose of partitioning the data set into smaller subsets that have similar variable characteristics [21]. In total, there were 11 unique algorithms identified within the top 60 most cited journal articles using data analytics during the period of 2015–2020. As previously discussed within the biopharmaceutical sector, some of the traditional statistical algorithms such as PCA, PLS and LR have been labelled as MVDA within these journal articles. However, these algorithms are more broadly defined as a subset of ML although their exact definition can vary between disciplines.

There is a long history of utilising conventional MVDA algorithms such as PCA and PLS, which assume linear relationships between inputs and responses. PCA has also been extensively used to better understand root cause of batch-to-batch variations and has been successfully implemented as far back as 1987 [22] and is commonly used today [1]. PLS algorithms have been employed historically to monitor end-point quality of fermentations based on the golden batch concept. This allows for operators to identify the source of process deviations or disturbances from an ideal trajectory and take corrective action quickly [23]. More recently, they have proven useful for spectral analysis involving PAT applications [24••]. Other established ML algorithms include NN and support vector machines (SVM). These algorithms allow for non-linear relationships to be modelled, which can be particularly useful for capturing non-linearities within

Figure 2



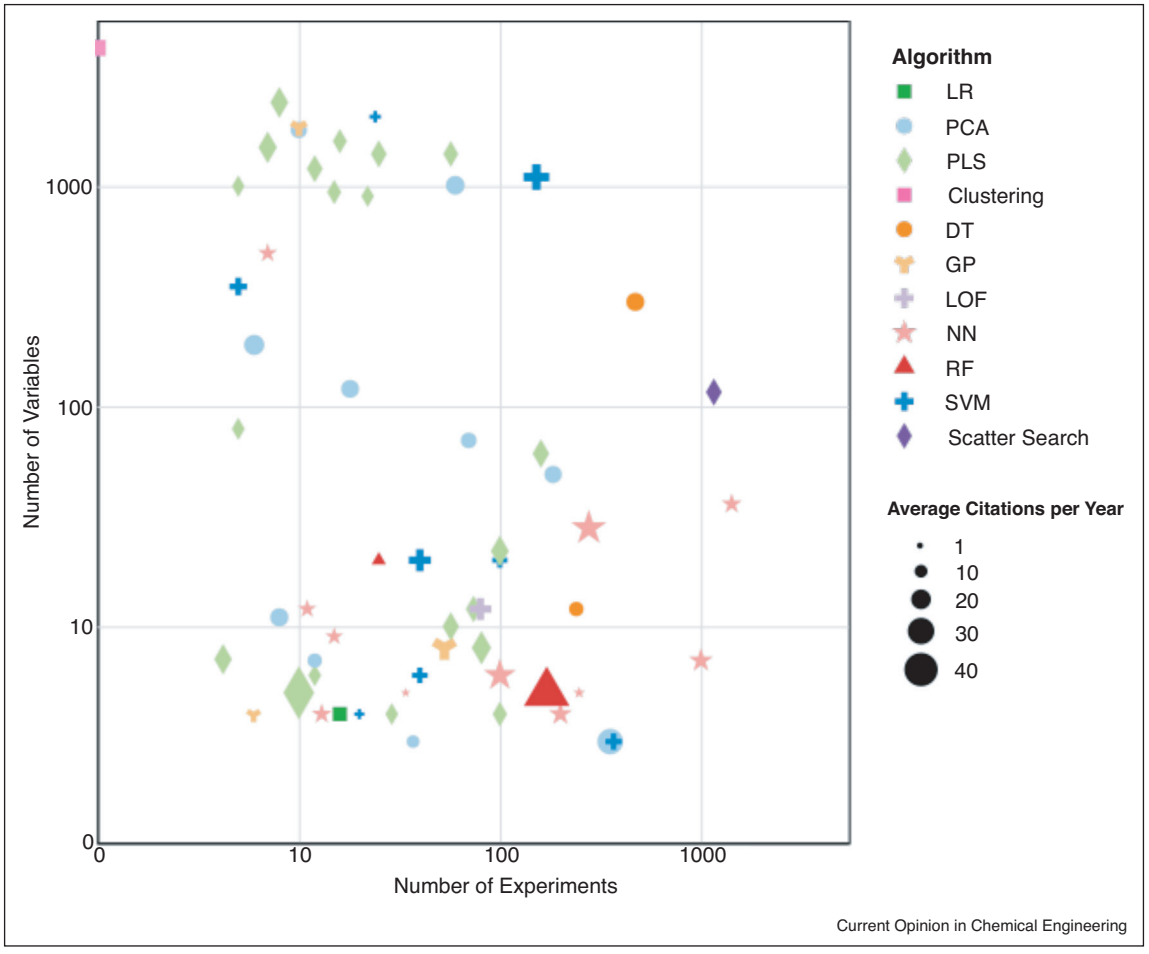
Classification of each algorithm employed in the top 60 journal articles based upon average citation number from the period of 2015–2020 identified as MVDA or ML and supervised or unsupervised learning technique. MVDA algorithms labelled as: Linear Regression (LR), Partial Least Squares (PLS) and Principal Component Analysis (PCA). ML algorithms labelled as: Clustering, Decision Tree (DT), Local Outlier Factor (LOF), Neural Networks (NN), Random Forests (RF), Support Vector Machines (SVM), Gaussian Process (GP) and Scatter Search (SS).

biological systems. For example, NN were applied successfully to predict loading capacity of depth filtration filters based upon non-linear functional correlations between inputs and outputs [25]. SVM are also advantageous due to their strong ability to generalise properties on unseen data which has resulted in its application across biological areas [26]. For example, SVM have been used to address the non-linear problem of predicting key process variables over time in penicillin production [27]. In-depth descriptions of the majority of the ML algorithms shown in Figure 2 can be found in the literature [28,29].

Within each journal article, the information related to the size of the data sets analysed and the application area was extracted. The two major factors that were recorded from each journal article was the number of variables and the number of experiments within each of the data sets analysed by these algorithms. For this analysis, the definition of a variable is any information that was recorded throughout the experiment and an experiment was considered as a single independent process run. The extracted information regarding the number of experiments, number of variables, algorithm type and average citation per year of the top 60 data analytics journal articles is summarised in Figure 3.

The average number of variables within the data set analysed using data analysis techniques was 444 ± 801 (minimum 1) demonstrating that there is a large range in the number variables being analysed within each data set. As the number of variables increase, PLS becomes the most commonly used algorithm. This can be seen in Figure 3, where for 500 or more variables, approximately 60% of the journal articles utilised PLS. PLS is the one of the most commonly used algorithms for the analysis of spectral data sets, based on its proven ability to correlate large numbers of variables (i.e. spectral wavelengths) with either the critical quality attributes or critical process parameters of interest. The maximum number of variables shown in Figure 3 was taken from Lempp *et al.* who analysed 4242 transcripts related to different genes of *Escherichia coli* [30]. This data was recorded using a single 1 L bioreactor with high-frequency transcriptomics data measuring a total of 29 different time points across the 20-hours experiment. Within their analysis, they selected a hierarchical clustering algorithm and used this to classify how different gene pools affected the transcription

Figure 3



Characteristics of algorithms in the top 60 most cited data analytics journal articles from the time period of 2015–2020 in terms of number of variables and experiments (log scale), algorithm type and average citations per year. Each symbol represents a different algorithm. The size of the symbol is proportional to the average number of citations per year for the journal articles, where a bigger symbol indicates a higher number of citations.

concentrations during the bioreactor run. This demonstrates that successful models can be built using a single bioreactor run provided that high frequency analytics are implemented to ensure sufficient data is available to validate the model. However, the information captured on a singular run may not be generalised for the process as it will not account for batch-to-batch variations or differences in bioreactor operation. One of smallest numbers of variables shown by Figure 3 was taken from Villain *et al.* who used three variables to build an SVM model which modelled a quantitative structure activity relationship (QSAR) for the acute toxicities of algae [31]. To avoid overfitting, the data set of 368 experiments was used to cross validate the model using a threefold cross validation technique. While the number of variables is small, this demonstrates that models built from smaller number of variables can still have high predictive power. Typically, having more variables within a data set allows for more complex models to be built as there are more degrees of freedom within the data set allowing for more interactions and relationships between the variables to be defined. However, complex models are not always desirable as they can be difficult to interpret and also the risk of these models being overfitted becomes higher [32]; therefore, independent external validation data is advised to ensure the model is robust. The average number of experiments used for the analysis within each of these journal articles was equal to 128 $\pm$ 267 (minimum 1) with the largest number of experiments identified in Figure 3 from Riba *et al.* [33]. Riba *et al.* utilised 1423 images to train an NN to classify dispensed cells as viable or dead [33]. For the purpose of this analysis, each individual image was classified as an individual experiment. Each image consisted of a 50 by 50-pixel grid which was used to train and validate the model. To avoid overfitting, the prediction performance of the model was evaluated using a 10-fold

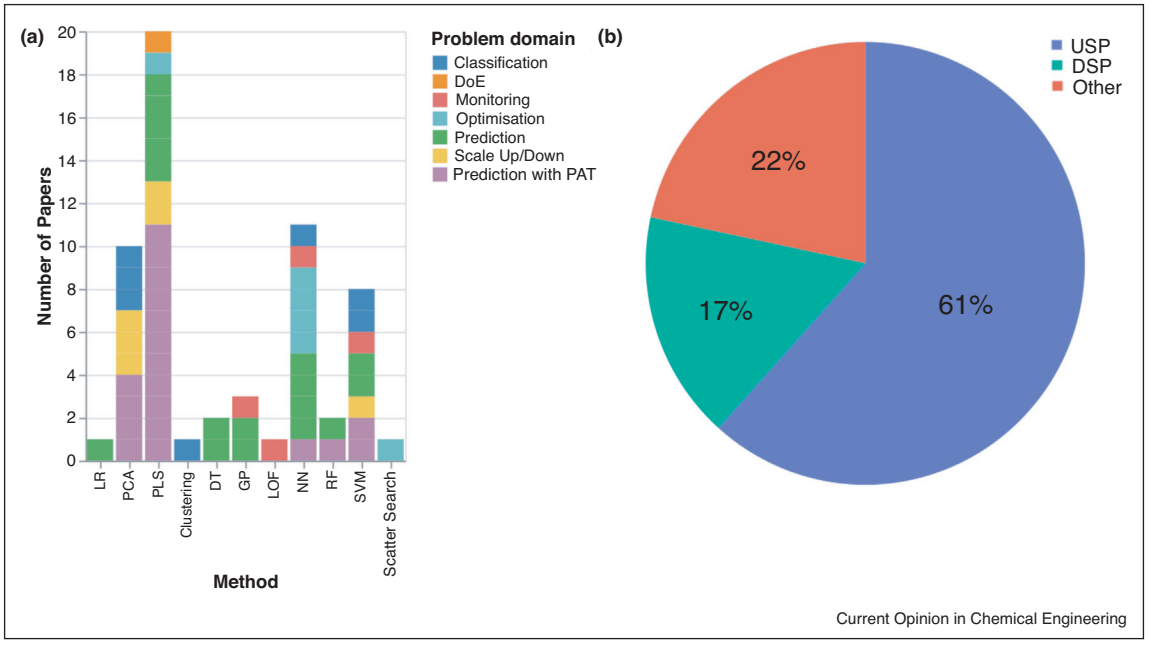
cross validation procedure. The translation of images to information results in large quantities of data being produced which is well suited for algorithms such as NN as these algorithms are more data hungry and generally require more data to build accurate models [34]. However, it is observed in Figure 3 that NN have been implemented also to applications with much smaller data sets, which demonstrates the robustness of this algorithm to varying volumes of available data.

It was observed in Figure 3 that data analysis techniques have been successfully applied with experiment numbers as low as one. Another example of an application which utilised a small number of experiments to build a valid model was by Brestich *et al.*, who successfully applied PLS to UV-vis spectroscopy to monitor preparative chromatography [35]. This was achieved through the utilisation of three experiments to train the PLS model with one for validation, in addition to employing cross validation to avoid overfitting the model. While the number of experiments was small, the dimension of these data sets used was large. For each of the four different chromatography experiments, a spectrum from 240 nm to 300 nm with 2 nm resolution was obtained. As a result, the data set produced had large dimensionality, that is, a large number of variables, allowing for more complex models to be built that could yield accurate predictions. This is a common feature of the majority of other PATs in which large amounts of information can be recorded in a small number of experiments. Additional data analysis applications involving a smaller number of experiments has also been shown in Figure 3, however not all small data sets contain enough information to build valid models. Indeed, the study by Vodopivec *et al.* analysed a data set of five experiments and applied SVM to compare metabolic profiling of 350 different metabolites between bioreactors of different sizes. The authors concluded that the models produced were of low quality, potentially due to the lack of data regarding the various metabolites which were not considered during the bioreactor runs [36]. Furthermore, this indicates that the data sets used to train these models were not representative of the whole bioreactor system. Another journal article with a small number of experiments was by Wang *et al.* who applied Gaussian process (GP) multilinear regression to infer the modulation effect of four metabolites using six experiments as a training set and one for validation. A smaller number of experiments was sufficient as GP uses the generated estimates of modulation effects to estimate parametric models, which generate more data which resulted in more accurate predictions [37]. These applications demonstrate that smaller numbers of experiments can produce robust and accurate models, but the data needs to be representative of the whole process being modelled. It must be noted that for the analysis of any data set, there are numerous algorithms that can be applied and should yield similar prediction errors or find the same correlations. Therefore, the choice of algorithm is most likely highly dependent on the familiarity and experience of the user analysing the data. They will most likely apply their preferred algorithm first and if the results are satisfactory, this algorithm will be utilised in the model building process.

A common perception from the sector is that developing a robust model requires a large number of experiments. However, the majority of processes in the biopharmaceutical industry suffer from something referred to as the 'Low-N' problem, where there is a limited number of historic experiments available for modelling a particular process [38•]. The Low-N scenario is common particularly with new biopharmaceutical products which have often only one or two experiments or runs at manufacturing scale. In these cases, it is common that the number of experiments is less than the number of variables being considered. Building models using these data sets are at risk of being overfitted [39], particularly when the ratio of variables to experiments is large. Possible solutions to provide larger data sets necessary for some of these data-hungry models include using algorithms such as GP to generate artificial data based on small data sets [40]. The method mentioned by Tulsyan *et al.* assumes that the initial data sets being replicated are representative of the whole process, which may not be accurate. Other solutions involve the use of Digital Twins to generate unlimited simulated data sets. This data can be used to develop and evaluate ML algorithms for process optimisation and speed up the readiness of these algorithms to implement once experimental data becomes available [41]. Other challenges have been raised by Mowbrey *et al.* about the data produced in the biopharmaceutical sector. This includes the sparsity of high dimensional data sets that often do not contain sufficient information about each individual dimension (variable) for model building [29]. The authors proposed a solution to effectively utilise these data sets by filling in the knowledge gaps using first principle models.

Figure 4 shows the breakdown of each application area by problem domain and process stage. More specifically, Figure 4a shows the frequency of each problem domain per algorithm type for the 60 most cited journal articles from the period of 2015–2020 and highlights PLS was the most widely implemented and diverse algorithm in the sector. This was utilised in 33% of journal articles within five different problem domains. This is most likely due to its proven success within the sector over the last 30–40 years in analysing noisy data with strongly correlated variables [42], which are a common feature within biopharmaceutical data sets. Comparatively, NN were used in 18% of journal articles in Figure 4a and utilised within five different problem domains. For the purpose of clarification in this paper, 'Prediction with PAT' was defined as soft sensing using an external non-standard device such

Figure 4



Applications of data analytics in the most cited journal articles by **(a)** problem domain per algorithm type and **(b)** process stage in terms of upstream processing (USP), downstream processing (DSP) or other.

as Raman spectroscopy, while 'Prediction' referred to the use of a model to develop a soft sensor using currently available variables. The main application areas in terms of problem domain where data analytics have been applied are in 'Prediction with PAT' and 'Prediction' which, in total, account for 62% of the top 60 journal articles. With the increasing uptake of PAT across the sector, large amounts of data are being recorded, which can be better exploited using the available ML algorithms. Interestingly, the PAT applications in all 60 journal articles screened were focused on process monitoring with no demonstrations of control which, may indicate the utilisation of PAT within the biopharmaceutical sector is still in the early stages of deployment. A similar trend was identified by Armstrong *et al.* in bioprocess chromatography systems where there was a clear gap between the number of PAT applications used for monitoring in comparison to control. One of the challenges they discussed was related to lack of confidence in the application of these technologies from a regulatory approval perspective in comparison to the standard off-line quantification methods [43•]. While there are no control applications appearing in the analysis of the top cited literature, ML is still being utilised in the sector to optimise the performance of existing processes. NN have been able to identifying the optimal fermentation conditions of biopharmaceutical product [44], while also being utilised to increase the speed of parameter estimation in mechanistic modelling of chromatography runs [45].

Within the topmost cited 60 journal articles during this period, most techniques focused on upstream processing (USP) which accounts for 61% of the overall journal articles as seen in Figure 4b. This is due most likely to a greater number of variables recorded during USP operations in comparison to downstream processing (DSP). This emphasises an opportunity to further explore these techniques in DSP for process optimisation. It is clear that ML can provide additional insight into existing processes as seen in the volume of applications in Figure 4. The majority of these ML applications are generated within research and development environments and one of the remaining challenges will be simplifying the transfer of these models between different scales and processes. Craven *et al.* investigated and compared mechanistic and statistical model transferability across bioreactors of different scales and modes of operation in mammalian cell bioprocessing [46]. The authors found that for the prediction of viable cell density between batch, fed batch and continuous operations, the ML models prediction quality was lower compared to the mechanistic models. This was attributed to the ML model's inability to incorporate feeding into its formulation, indicating that ML models may struggle to extrapolate from datasets which are widely different from the dataset it was trained and validated on. This research demonstrates that data analytics will continue to be an integral part of biopharmaceutical process development, but additional work is required to further exploit the benefits of these tools for

process control and optimisation within commercial biomanufacturing.

Conclusion

Over the last decade there has been a plethora of algorithms implemented for the analysis of highly diverse biopharmaceutical data sets. This work highlighted some interesting trends within the data set investigated. Between 2010 and 2020, PLS emerged as the most frequently applied algorithm within the sector, according to citation frequency. PLS represented 33% of such journal articles, rising to 60% when there were a large number of variables (>500). This accounted for almost 50% of the algorithm usage. This is most likely due to the high implementation of PLS for the analysis of PAT applications that contain large number of variables due to nature of the spectral data files. The second most cited algorithm was NN, with approximately 22% of journal articles published utilising this technique during this period. This may be due to the ability of this algorithm to capture complex relationships, which may yield more accurate predictions of non-linear variables such as viable cell densities or amino acid consumption rates. Within the journal articles evaluated, it was found that the majority of ML applications were focused on analysing data related to USP applications, accounting for 61% of the journal articles investigated. This is likely due to the large number of variables available for analysis compared to DSP or other application areas. There was no clear trend between the size of the data set analysed and the algorithm applied. This outcome demonstrates that the data set size, in terms of number of variables or experiments, is independent of the algorithm utilised. The appropriate algorithm should be based on the specific problem to be analysed. Therefore, the amount of data required for the development of useful models within the biopharmaceutical sector is most likely dependent on the complexity of both the data set and the problem to solve.

Based on the growing trend observed in the use of ML algorithms, it is clear that the sector will continue to explore and take advantage of insights and model predictions to optimise process development and manufacturing operations. Significant improvements are expected in current manufacturing operations with increased adoption of advanced data analytics, enabling soft sensor and PAT integration and hence more advanced control strategies. Furthermore, as the industry adopts the core principles of Industry 4.0, it will move towards the digitisation of all their recorded data within a queryable and structured centralised repository such as a data lake or data warehouse. This digital revolution will simplify data consolidation and accessibility, enabling ML algorithms to be applied to all data recorded from multiple sites across different scales and unit operations. This will help facilitate the ultimate goal of having a fully automated data-driven biopharmaceutical manufacturing facility of the future.

Conflict of interest statement

Nothing declared.

Acknowledgements

Financial support from the UK Biotechnology and Biological Sciences Research Council (BBSRC) and AstraZeneca for the Industrial CASE PhD studentship for Matthew Banner is gratefully acknowledged (BB/V509607/1). This research is associated with the joint UCL-AstraZeneca Centre of Excellence for predictive decision-support tools in the bioprocessing sector and is aligned with the EPSRC Future Targeted Healthcare Manufacturing Hub hosted by UCL Biochemical Engineering.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Manahan M, Nelson M, Cacciatore JJ, Weng J, Xu S, Pollard J: **Scale-down model qualification of ambr® 250 high-throughput mini-bioreactor system for two commercial-scale mAb processes**. *Biotechnol Prog* 2019, **35**:1-13.
2. FDA: *Guidance for Industry Guidance for Industry PAT — A Framework for Innovative Pharmaceutical*. 2004.
3. Mercier SM, Diepenbroek B, Wijffels RH, Streefland M: **Multivariate PAT solutions for biopharmaceutical cultivation: current progress and limitations**. *Trends Biotechnol* 2014, **32**:329-336.
4. Griffin TJ, Seth G, Xie H, Bandhakavi S, Hu WS: **Advancing mammalian cell culture engineering using genome-scale technologies**. *Trends Biotechnol* 2007, **25**:401-408.
5. Goldrick S, Umprecht A, Tang A, Zakrzewski R, Cheeks M, Turner R, Charles A, Les K, Hulley M, Spencer C *et al.*: **High-throughput Raman spectroscopy combined with innovate data analysis workflow to enhance biopharmaceutical process development**. *Processes* 2020, **8**:1179.
6. von Stosch M, Portela RM, Varsakelis C: **A roadmap to AI-driven in silico process development: bioprocessing 4.0 in practice**. *Curr Opin Chem Eng* 2021, **33**:100692.
7. Kuniavsky M: **Introduction - the middle of Moore's law**. *Smart Things: Ubiquitous Computing User Experience Design*. Elsevier; 2010:3-11.
8. Rundo F, Trenta F, di Stallo AL, Battiato S: **Machine learning for quantitative finance applications: a survey**. *Appl Sci* 2019, **9**.
9. Perlich C, Dalessandro B, Raeder T, Stitelman O, Provost F: **Machine learning for targeted display advertising: transfer learning in action**. *Mach Learn* 2014, **95**:103-127.
10. Beckett C, Eriksson L, Johansson E, Wikström C: **Multivariate data analysis (MVDA)**. *Pharmaceutical Quality by Design*. 2018:201-225.
11. Todorov H, Searle-White E, Gerber S: **Applying univariate vs. multivariate statistics to investigate therapeutic efficacy in (pre)clinical trials: a Monte Carlo simulation study on the example of a controlled preclinical neurotrauma trial**. *PLoS One* 2020, **15**:1-20.
12. Bhatia H, Mehdizadeh H, Drapeau D, Yoon S: **In-line monitoring of amino acids in mammalian cell cultures using raman spectroscopy and multivariate chemometrics models**. *Eng Life Sci* 2018, **18**:55-61.
13. Goldrick S, Holmes W, Bond NJ, Lewis G, Kuiper M, Turner R, Farid SS: **Advanced multivariate data analysis to determine the root cause of trisulfide bond formation in a novel antibody-peptide fusion**. *Biotechnol Bioeng* 2017, **114**:2222-2234.

14. Aksnes DW, Langfeldt L, Wouters P: **Citations, citation indicators, and research quality: an overview of basic concepts and theories**. *SAGE Open* 2019, **9**.

15. Harzing AW: *Publish or Perish*. . available from https://harzing.com/resources/publish-or-perish 2007.

16. De Kok S, Enyeart P, Richard H, Hauck T, Humphries C, Lieder S: Downscaling parameters to design experiments and plate models for micro-organisms at small scale to improve prediction of performance at larger scale (Similar to 17) (Patent: https://patents.google.com/patent/WO2020227299A1/en?oq=WO2020227299A1).

17. Berry B, Moretto J: *Cross-scale Modelling of Bioreactor Cultures Using Raman Spectroscopy*. 2020.

18. McAdams HH, Arkin A: **It's a noisy business! Genetic regulation at the nanomolar scale**. *Trends Genet* 1999, **15**:65-69.

19. Fan F-L, Xiong J, Li M, Wang G: **On interpretability of artificial neural networks: a survey**. *IEEE Trans Radiat Plasma Med Sci* 2021, **5**:741-760.

20. • Mendez KM, Reinke SN, Broadhurst DI: **A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification**. *Metabolomics* 2019, **15**:1-15

This study evaluates multiple machine learning algorithms on the classification of metabolomic datasets. They propose that the quality of the dataset used by the algorithm is an important factor for model building.

21. Alloghani M, Al-Jumeily D, Mustafina J, Hussain A, Aljaaf AJ: **A systematic review on supervised and unsupervised machine learning algorithms for data science**. In *Supervised and Unsupervised Learning for Data Science.* Edited by Berry MW, Mohamed A, Yap BW. Springer International Publishing; 2020:3-21.

22. Wold S, Esbensen K, Geladi P: **Principal components analysis**. *Chemom Intell Lab Syst* 1987, **2**:37-52.

23. Nomikos P, MacGregor JF: **Multi-way partial least squares in monitoring batch processes**. *Chemom Intell Lab Syst* 1995, **30**:97-108.

24. •• Wasalathanthri DP, Rehmann MS, Song Y, Gu Y, Mi L, Shao C, Chemmalil L, Lee J, Ghose S, Borys MC *et al.*: **Technology outlook for real-time quality attribute and process parameter monitoring in biopharmaceutical development—a review**. *Biotechnol Bioeng* 2020, **117**:3182-3198

The review highlights the importance of process analytical 488 technology and machine learning in the automation of biopharmaceutical manufacturing and its role in Industry 4.0. While discussing potential challenges with its implementation in the sector.

25. Agarwal H, Rathore AS, Hadpe SR, Alva SJ: **Artificial neural network (ANN)-based prediction of depth filter loading capacity for filter sizing**. *Biotechnol Prog* 2016, **32**:1436-1443.

26. Yang ZR: **Biological applications of support vector machines**. *Brief Bioinform* 2004, **5**:328-338.

27. Li Y, Yuan J: **Prediction of key state variables using support vector machines in bioprocesses**. *Chem Eng Technol* 2006, **29**:313-319.

28. Dey A: **Machine learning algorithms: a review**. *Int J Comput Sci Inf Technol* 2016, **7**:1174-1179.

29. Mowbray M, Savage T, Wu C, Song Z, Cho BA, Del Rio-Chanona EA, Zhang D: **Machine learning for biochemical engineering: a review**. *Biochem Eng J* 2021, **172**:108054.

30. Lempp M, Farke N, Kuntz M, Freibert SA, Lill R, Link H: **Systematic identification of metabolites controlling gene expression in *E. coli***. *Nat Commun* 2019, **10**.

31. Villain J, Minguez L, Halm-Lemeille MP, Durrieu G, Bureau R: **Acute toxicities of pharmaceuticals toward green algae. mode of action, biopharmaceutical drug disposition classification system and quantile regression models**. *Ecotoxicol Environ Saf* 2016, **124**:337-343.

32. Ying X: **An overview of overfitting and its solutions**. *J Phys Conf Ser* 2019, **1168**.

33. Riba J, Schoendube J, Zimmermann S, Koltay P, Zengerle R: **Single-cell dispensing and 'real-time' cell classification using convolutional neural networks for higher efficiency in single-cell cloning**. *Sci Rep* 2020, **10**:1-9.

34. Adadi A: *A Survey on Data-efficient Algorithms in Big Data Era*. Springer International Publishing; 2021.

35. Brestich N, Rüdt M, Büchler D, Hubbuch J: **Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression**. *Chem Eng Sci* 2018, **176**:157-164.

36. Vodopivec M, Lah L, Narat M, Curk T: **Metabolomic profiling of CHO fed-batch growth phases at 10, 100, and 1,000 L**. *Biotechnol Bioeng* 2019, **116**:2720-0002729.

37. Wang M, Risuleo RS, Jacobsen EW, Chotteau V, Hjalmarsson H: **Identification of nonlinear kinetics of macroscopic bioreactions using multilinear Gaussian processes**. *Comput Chem Eng* 2020, **133**:106671.

38. • Tulsyan A, Garvin C, Ündey C: **Advances in industrial biopharmaceutical batch process monitoring: machine-learning methods for small data problems**. *Biotechnol Bioeng* 2018, **115**:1915-1924

This study highlights a reoccurring problems within the sector defined as the Low N problem where minimal historic data is available for model building. They propose a solution using machine learning and hardware exploitation to generate arbitrarily large numbers of data sets to improve monitoring accuracy.

39. Verleysen M, François D: **The curse of dimensionality in data mining and time series prediction**. *Analysis* 2005, **3512**:758-770.

40. Tulsyan A, Garvin C, Undey C: **Industrial batch process monitoring with limited data**. *J Process Control* 2018, **77**:114-133.

41. Goldrick S, Duran-Villalobos CA, Jankauskas K, Lovett D, Farid SS, Lennox B: **Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process**. *Comput Chem Eng* 2019, **130**:106471.

42. Wold S, Sjostrom M: *PLS-regression: A Basic Tool of Chemometrics*. 2001.

43. • Armstrong A, Horry K, Cui T, Hulley M, Turner R, Farid SS, Goldrick S, Bracewell DG: **Advanced control strategies for bioprocess chromatography: challenges and opportunities for intensified processes and next generation products**. *J Chromatogr A* 2021, **1639**:461914

This review evaluates the current status and future directions of bioprocess chromatography control in the sector. They include popular research and development areas such as process analytical technology and machine learning and discuss its limitations currently.

44. Chen F, Li H, Xu Z, Hou S, Yang D: **User-friendly optimization approach of fed-batch fermentation conditions for the production of iturin A using artificial neural networks and support vector machine**. *Electron J Biotechnol* 2015, **18**:273-280.

45. Pirrung SM, van der Wielen LAM, van Beckhoven RFWC, van de Sandt EJAX, Eppink MHM, Ottens M: **Optimization of biopharmaceutical downstream processes supported by mechanistic models and artificial neural networks**. *Biotechnol Prog* 2017, **33**:696-707.

46. Craven S, Shirsat N, Whelan J, Glennon B: **Process model comparison and transferability across bioreactor scales and modes of operation for a mammalian cell bioprocess**. *Biotechnol Prog* 2013, **29**:186-196.