Computational and Structural Biotechnology Journal 23 (2024) 2497–2506

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Antibody glycan quality predicted from CHO cell culture media markers and machine learning

Meiyappan Lakshmanan [a,b,c,d], Sean Chia [a], Kuin Tian Pang [a], Lyn Chiin Sim [a], Gavin Teo [a], Shi Ya Mak [a], Shuwen Chen [a], Hsueh Lee Lim [a], Alison P. Lee [a], Farouq Bin Mahfut [a], Say Kong Ng [a], Yuansheng Yang [a], Annie Soh [a], Andy Hee-Meng Tan [a], Andre Choo [a], Ying Swan Ho [a,*], Terry Nguyen-Khuong [a,*], Ian Walsh [a,*]

[a] *Bioprocessing Technology Institute (BTI), Agency for Science, Technology and Research (A*STAR), 20 Biopolis Way, #06–01 Centros, Singapore 138668, Republic of Singapore*
[b] *Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, India*
[c] *Centre for Integrative Biology and Systems medicinE (IBSE), Indian Institute of Technology Madras, India*
[d] *Robert Bosch Centre for Data Science and AI (RBCDSAI), Indian Institute of Technology Madras, India*

ARTICLE INFO

Keywords:
Machine learning
Capillary electrophoresis
N-glycan prediction
Anti-Her2
Spent media
Media analysis
Antibody quality

ABSTRACT

N-glycosylation can have a profound effect on the quality of mAb therapeutics. In biomanufacturing, one of the ways to influence N-glycosylation patterns is by altering the media used to grow mAb cell expression systems. Here, we explore the potential of machine learning (ML) to forecast the abundances of N-glycan types based on variables related to the growth media. The ML models exploit a dataset consisting of detailed glycomic characterisation of Anti-HER fed-batch bioreactor cell cultures measured daily under 12 different culture conditions, such as changes in levels of dissolved oxygen, pH, temperature, and the use of two different commercially available media. By performing spent media quantitation and subsequent calculation of pseudo cell consumption rates (termed media markers) as inputs to the ML model, we were able to demonstrate a small subset of media markers (18 selected out of 167 mass spectrometry peaks) in a Chinese Hamster Ovary (CHO) cell cultures are important to model N-glycan relative abundances (Regression - correlations between 0.80–0.92; Classification - AUC between 75.0–97.2). The performances suggest the ML models can infer N-glycan critical quality attributes from extracellular media as a proxy. Given its accuracy, we envisage its potential applications in biomaufactucuring, especially in areas of process development, downstream and upstream bioprocessing.

1. Introduction

Glycosylation involves the attachment of N-glycans, a carbohydrate consisting of several monosaccharides, to the amide nitrogen of an asparagine amino acid within the protein. It is an enzymatic, site-specific process that occurs within a cells' endoplasmic reticulum-Golgi complex and is heavily dependent upon the efficiencies of glycotransferases, availability of nucleotide sugar donors and metabolic precursors or cofactors of the glycosylation biosynthetic pathway. In monoclonal antibodies (mAbs), N-glycans attach to the CH2 domain and are a critical quality attribute (CQA) that has a pivotal influence on the drug's efficacy [1,2]. Various cell culture parameters such as media/-feed, pH and temperature can be varied to influence the type, complexity, branching, and topology of N-glycan structures, by altering the host cellular metabolism [3], feeding glycotransferase inhibitors [4] and changing physicochemical parameters of the process [5]. Consequently, the complex glycosylation biosynthetic pathway is not a trivial process to control during the manufacturing of biologics and slight changes in process parameters can result in diverse changes in the N-glycan profiles.

N-glycans associated with mAbs can be broadly categorized into fucosylation, galactosylation, mannosylation and sialylation (Supp Fig. 1). The four groupings are known to be important CQAs for antibody therapeutic efficacy as each have implications for the immunogenicity, half-life, and pharmacokinetics of the therapeutic candidate. For instance, galactosylation, fucosylation and mannosylation, affect the

* Corresponding authors.
E-mail addresses: ho_ying_swan@bti.a-star.edu.sg (Y.S. Ho), terrynguyen@gmail.com (T. Nguyen-Khuong), walshi@bti.a-star.edu.sg (I. Walsh).

https://doi.org/10.1016/j.csbj.2024.05.046
Received 19 February 2024; Received in revised form 22 May 2024; Accepted 28 May 2024
Available online 1 June 2024

complement activity [6], antibody-dependent cellular cytotoxicity (ADCC) [7] and clearance [8] of the antibody respectively. Similarly, sialyation has also been reported to affect the pharmacokinetic and pharmacodynamic properties of glycoprotein drugs [9].

Various studies have examined altering media to change N-glycan levels for improved therapeutic efficacy. For instance, it was demonstrated that supplementing the feed media with differing levels of glucose and amino acids can significantly affect IgG glycan profiles [3]. Other studies have also demonstrated the effect of other supplements, such as uridine, galactose, 2-F-peracetyl fucose, and manganese chloride on cell metabolism and gene expression of key glycosylation-related proteins [10–12]. Changing levels of other media components such as trace metals, productivity enhancers like butyrate, hormones, nucleotide sugars, nucleotide sugar transporters, or changing culture physicochemical parameters were also reported to alter N-glycosylation patterns [5,6,13]. Thus, there is keen interest in developing different media formulations in an effort to deliver consistent or better product glycosylation.

Multivariate data-driven approaches (MVDA) have now become an essential aspect in bioprocess design and development [14,15]. MVDA encapsulates a statistical and rational approach to support the detailed understanding of how bioprocessing parameters affect the quality and yield of the biotherapeutic product. Regulatory bodies such as the U.S. Food and Drug Administration (FDA) have encouraged the Quality by Design (QbD) paradigm in which statistical modeling approaches can be a tool [16]. Within MVDA, most of the research so far has used techniques such as Partial Least Square regression analysis and Principal Component Analysis that implement predictive models for mAb glycan abundance [17]. One reason for the adoption of MVDA algorithms is their simplicity and ability to capture simple relationships in the data. However, moving beyond MVDA to data-driven approaches and harnessing machine learning (ML) requires high-throughput experiments enabling the generation of diverse and large quantities of collectible data. Consequently, while ML offers distinct advantages in specific domains, the applicability of ML-based predictions for product attributes like titer, viable cell density, and glycosylation remains lacking [18]. Among the literature that are available, Artificial Neural Networks (ANNs) seem to be the most popular – ANN approaches have been used to predict cell growth from different media and seeding methods [19], as well as forecasting monoclonal antibody concentrations from fluorescence measurements [20]. In the latter case, ANN was shown to outperform Partial Least Squares likely due to its capacity for non-linear modelling. Other ML algorithms, such as random forest models, have also been proposed to enable real-time process control and product CQA prediction [21]. In particular, they appear to demonstrate superior performance for smaller datasets with a reduced risk of overfitting. Finally, hybrid approaches that combine stoichiometric modelling and ANNs have also been used to predict glycan abundance and provide metabolic insight, though the ANNs were relatively small due to data size constraints and the risk of overfitting [22].

In this study we develop ML models that exploit a high-throughput capillary electrophoresis N-glycan characterization and a corresponding quantitative analysis of both spent media and physiochemical measurements for each sample over time. To the best of our knowledge the number of samples involved in the glycomic analysis and collection of corresponding process parameters represents one of the largest datasets for Chinese Hamster Ovary (CHO) from fed-batch cultures. These data include glycan identities and abundances determined at 12 different operating conditions (OCs), including variations in dissolved oxygen and pH levels, the use of two different commercially available media and the introduction of temperature shifts during the culture, over 12 days of culture in 3 biological replicates, resulting in a large dataset of bioreactor data points. Using this dataset, ML algorithms using the mass spectrometry (MS)- derived spent media markers (MMs) as input could accurately predict N-glycan galactosylation, fucosylation, mannosylation and sialylation abundances.

We show that eighteen selected MMs improve model performance significantly compared to commonly used media variables such as glucose, lactate, and amino acids. For galactosylation prediction, ML models showed a statistically significant improvement compared to MVDA approaches. We believe such prediction models could be used to detect N-glycan CQAs from the extracellular spent media as a proxy from selected MS peaks, which can be highly effective for N-glycan CQA analysis and monitoring, in the overall aim of developing mAb products

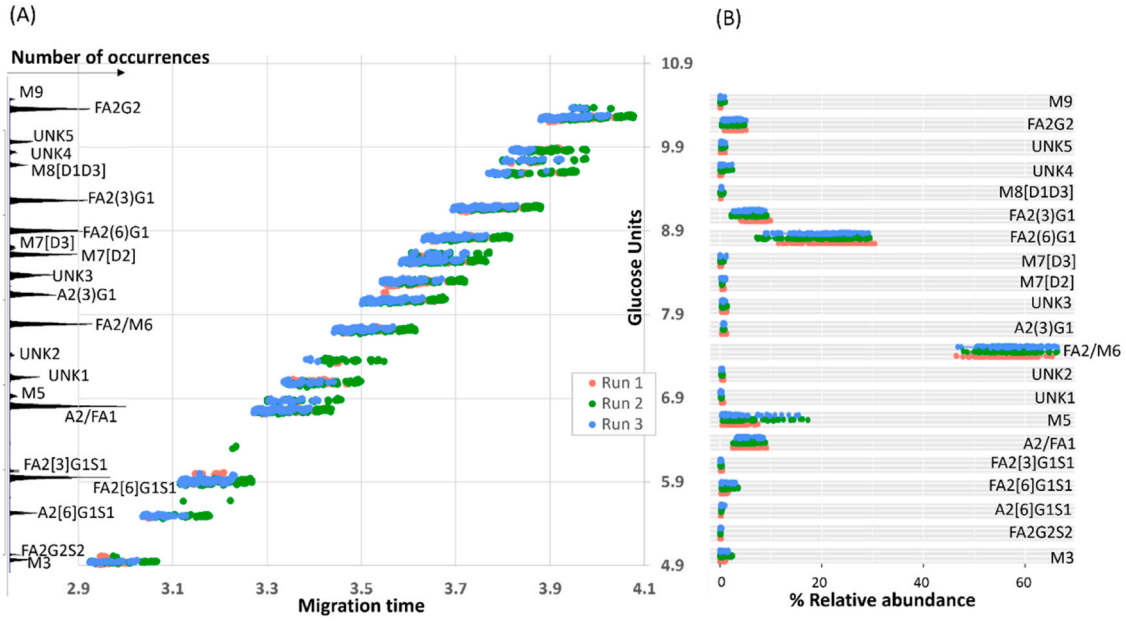

Fig. 1. Qualitative and quantitative glycan analysis using CE. (A) Migration times, Glucose Units (GU) and GU occurrence plot for all glycan samples. (MT,GU) points are averaged from 3 technical replicates resulting in 432 sample points (i.e. 1296 divided by 3 technical replicates). The number of times a glycan is identified are plotted on the left and show the number of times their GU values were observed. Any peaks marked UNK_ could not be matched to the AFTPS GU database. Glycan in text here are represented as SNFG diagrams Supplementary Table 1. (B) The relative abundances of glycans and unidentified peaks – each datapoint corresponds to a (MT,GU) point in 1A.

of high quality.

2. Materials and methods

2.1. Experimental

Detailed experimental protocols are described in the Supplementary material. In the following a brief description is provided.

2.1.1. Bioreactor operation

Our in-house CHO-K1 cell lines producing Anti-HER2 biosimilar (IgG1 subclass) were cultured in 14-day fed-batch cultures using Ambr250 bioreactors (Sartorius, Royston, UK). Operating conditions, pH, dissolved oxygen (dO_2) percent air saturation, temperature shift, were varied each day to bring variability to the glycosylation, media components and culture. For temperature shift experiments, at the end of day 5, temperature was reduced to 33 °C (from the initial 37 °C from day 0–5) and maintained until time of culture harvest. Cell counting was performed on Vi-Cell XR viability analyzer (Beckman Coulter, CA, USA) and glucose, lactate, pH, ammonium ions, sodium ions, and potassium ions were profiled using a Nova bioprofile 100 plus analyzer (Nova Biomedical, Waltham, MA).

2.1.2. Glycan analysis

Cell supernatant was filtered, and antibodies were then purified using protein A HP spin trap columns (CYTIVA, Marlborough, MA, USA). Samples were then desalted using 30 kDa Amicon Ultra Centrifugal Filters, (Merck Millipore, Tullagreen, Co. Cork, Ireland). N-glycans were released from purified antibodies by digesting with recombinant PNGase F (New England Biolabs, Ipswich, MA, USA) and labeling them with 8-Aminopyrene-1,3,6-Trisulfonic Acid (APTS) using the FAST Glycan Kit (SCIEX, Farmingham, MA, USA). Capillary electrophoresis of the released APTS-labeled N-glycans was performed on a CESI8000 CE instrument (SCIEX, Redwood City, CA, USA) equipped with a solid-state laser-induced fluorescent detector (excitation 488 nm, emission 520 nm). Data analysis for N-glycan identification and quantitation was performed as previously described [23]. This resulted in each identified glycan having a relative abundance of all glycans detected measured as a percentage.

2.1.3. Metabolite analysis

Spent media were collected for each condition and filtered through 10 kDa molecular weight cut-off (MWCO) membranes. Filtered spent media were analysed using an ACQUITY Ultra Performance Liquid Chromatography (UPLC) system (Waters Corporation, MA, USA) in tandem with a QExactive Orbitrap mass spectrometer (Thermo Fisher Scientific, CA, USA), and the data processed as previously described [24,25]. Briefly, chromatographic peaks were integrated using the xcms package [26] with preset criteria of a minimum signal-to-noise ratio of 3. A pooled quality-control (QC) mix, which consisted of equal aliquots of all spent media samples, was used for signal correction within and between each batch analysis. Features were filtered based on integrated peak area cut-off of 1000 a.u., and integrated peak area (IPA) in QC mix with coefficient of variation below 30 %. Thus, the IPAs were consistently reproduced and could be considered analytes or features of the media. Based on these peak-picking and peak integration criteria, glucose, lactate and 20 amino acids as well as a total of 145 peak features were quantitated using their IPAs.

2.1.4. Calculation of metabolic consumption rates

The specific metabolic rate was used to represent production or consumption of media markers, with a negative value representing consumption and a positive value for production. The production or consumption of media markers in the culture was calculated based on the difference in levels of each media marker (Eq. 1) and the viable cell density (Eq. 2), across two sampling points on a daily basis. The specific metabolic rate of each media marker was then derived by taking the ratio of normalized concentrations of media markers to the integrated viable cell density across the two time points (Eq. 3).

$$\text{Consumption or production of each media marker} = I_x^t - I_x^{t-1} \tag{1}$$

$$\text{Integrated viable cell density,}\quad IVCD_t \quad (cell \bullet day \bullet ml^{-1}) = \frac{VCD_t + VCD_{t-1}}{2} + IVCD_{t-1} \tag{2}$$

$$\text{Specific metabolic rate}: \quad q_x^t = \frac{I_x^t - I_x^{t-1}}{IVCD^t - IVCD^{t-1}} = \frac{\Delta I_x^t}{\Delta IVCD^t} \tag{3}$$

Where x represents the respective media marker, I represents integrated peak area of the media marker (representative of its concentration in the media), and t represents the sampling time point.

2.2. Description of the Fed-Batch Dataset

The dataset used for our analysis consisted of 12 different cell culture conditions in fed-batch Ambr250 reactors, harvested over 3 biological replicates (BRs) across 12 days (days 3–14) – Table 1. For each culture condition the primary physicochemical parameters such as pH (6.9 - 7.3), dO_2 (30–50 % air saturation), two media platforms and temperature shift strategies were altered (Table 1). Additionally for each condition, other process variables such as viable cell density, pH, temperature, dO_2, potassium ions, sodium ions and ammonium ions were collected.

2.3. Algorithm design

Random Forests machine learning (RF) algorithms were chosen due to their efficient training time, ability to deal with relatively small sample size and their nonparametric nature [27]. Additionally, RFs were previously shown to be better at predicting mAb quality in continuous manufacturing compared to other ML techniques [21]. RF were trained using the Weka machine learning Java package using default parameters [28]. Features were selected using a simple correlation-based criteria. RFs were optimized to predict the abundance of mAb Fc fucosylation, galactosylation, mannosylation, and sialylation measured as a percentage (i.e. regression). Additionally, RF models were optimized in a classification problem to distinguish abundances in two classes. Table 2 shows the list of variables available for model optimization. Each input and output variable were used in its raw form (e.g. MM rates and N-glycan % abundance). Selected media markers were defined as any consumption rate that correlated with fucosylation, galactosylation, mannosylation or sialylation and had an IPA CV < 30 %.

2.3.1. Model comparison

To compare the ML models with a MVDA approach, the inputs in Table 1 were used in partial least square regression (PLRS) models. Different input combinations were tested in both ML and MVDA: Basic alone, Amino alone, Selected MMs alone, Basic + Amino, Basic + Selected MMs, Amino + Selected MMs, Basic + Amino + Selected MMs, and finally all 167 features.

2.3.2. Evaluation

For regression, Pearson correlation coefficient (CC), mean absolute error (MAE), and root mean square error (RMSE) were used as performance metrics. The CC, MAE, and RMSE metrics were extracted from the Weka ML library [28]. Statistical testing was performed using the t-test and p-values were corrected for multiple testing using Benjamini-Hochberg procedure [29] at 5 % false discovery rate. Classification performance was calculated using area under the receiver operating characteristic (ROC) curve [30]. The area under the ROC (AUC) metric was calculated using the pROC R package [31]. Models

Table 1
Description of operating conditions for the 12 fed-batches. Each were run for 12 days and 3 BRs resulting in 432 data points (12 days x 12 conditions x 3 BRs). Each condition was run in three technical replicates and all variables were averaged.

Media platform 1 GE ActiPro Basal (Cat# SH31037.01) GE Cell Boost 7a Feed (Cat# SH31026.01) GE Cell Boost 7b Feed (Cat# SH31027.07)				Media platform 2 SAFC EX-CELL Advanced CHO Fed-batch System Basal (Cat# G 3126) EX-CELL Advanced CHO Fed-batch Feed 1 with glucose (Cat# 24367 C-1 L)			
Condition No.	pH	dO_2 (% Air saturation)	Temp. Shift	Condition No.	pH	dO_2 (% Air saturation)	Temp. Shift
1	6.9	50	NO	2	6.9	50	NO
3	7.1	50	NO	4	7.1	50	NO
5	7.1	50	YES	6	7.1	50	YES
7	7.1	30	NO	8	7.1	30	NO
9	7.3	50	NO	10	7.3	50	NO
11	7.3	30	NO	12	7.3	30	NO

Table 2
The 167 spent media and 6 physicochemical features available in this study. MMs - media markers. The labels in the input name column are used throughout this work. The number and type of variable for each group is indicated.

Input name	Physicochemical parameters	Definition of media component
Basic	Temperature, pH, Dissolved oxygen,	Potassium ion, Sodium ion, Ammonium ion, Glucose, Lactic acid
Amino	-	Alanine, Cysteine, DL-tyrosine, DL-tryptophan, Glycine, Isoleucine, L-arginine, L-asparagine, L-aspartic acid, Leucine, L-glutamic acid, L-glutamine, L-histidine, Iserine, Lysine, Methionine, Phenylalanine, Proline, Threonine, Valine
Selected MMs	-	18 MS derived rates with IPA CV < 30 % and correlated with at least one of fucosylation, galactosylation, mannosylation, sialylation
Remaining MMs	-	127 MS derived rates with IPA CV < 30 % and not correlated with either fucosylation, galactosylation, mannosylation or sialylation

were optimized, and prediction performance evaluated in a leave one out cross validation (LOOCV) using biological replicates (BRs). Specifically, with three BRs available, the train/test splits were implemented as follows: training set BR 2&3/test set BR 1, training set BR 1&3/test set BR 2, and training set BR 1&2/test set BR 3. All performance metrics were averaged over the 3 test sets and standard error bars calculated.

3. Results

3.1. CHO-K1 culture data

Firstly, to comprehensively characterize how various OCs and media affect the N-glycosylation of Anti-HER2 antibodies, a dataset was generated by analysing spent media, physicochemical parameters, and the glycosylation patterns of the antibodies, from days 3, 5, 7, 9, 11 and 14 in 36 CHO cell culture runs (36 = 12 OCs x 3 BRs). This led to a total of 1296 samples measured, making the dataset one of the largest available (Summary in Supplementary Table 1).

3.2. Defining targets: glycan identification and quantitation

The N-glycans in the Anti-HER2 antibodies were analyzed using a high-throughput Capillary Electrophoresis (CE) and a data extraction approach as previously described [23]. In particular, such analysis allowed the high-throughput and accurate measurements of N-glycan abundances within antibodies, which were necessary as the output target variable. In all 1296 samples, N-glycans were identified based on CE triple standard glucose unit (GU) calculation [32]. Briefly, the calculation involved standardizing the migration time (MT) of the peaks by generating a 'virtual' GU ladder using the MTs of 3 oligosaccharide internal standards. Subsequently, glycans were identified using a CE-GU database. We observed that the glycan structures were identified with a difference between the observed values and the database of at most 0.092 GU (Supplementary Figure 2), suggesting a high degree of accuracy. This allowed us to identify 16 of the most abundant glycans (Fig. 1). Five peaks could not be determined (UNK in Fig. 1); however, as their abundances are very low (Supplementary Table 2) they do not affect the overall fucosylation, galactosylation, mannosylation, and sialylation trends. In particular, this methodology enabled us to determine the number of occurrences of observed GU values, and hence identify and annotate glycans of close MTs that may have otherwise been difficult to annotate (Fig. 1A). For instance, identification and quantitation of A2/FA1 and M5 peaks could be decoupled despite close MTs (Fig. 1A). Some glycan species only occurred occasionally depending on the condition or day of culture while others were always present in all 1296 samples. For example, M5 was only identified in ~7 % of the samples whereas the major FA2/M6 peak was always identified. We determined the relative abundance of the individual glycans (as a percentage of all measured glycans), and further combined their abundances to represent fucosylation, galactosylation, high-mannosylation, and sialylation quantities in a sample (Fig. 1B and Supplementary Figure 2). Totaling the glycan abundance in this way (e.g. fucosylation, galactosylation) allows us to captures the ADCC [33], complement-dependent cytotoxicity [34], clearance [8] and pharmacokinetic/pharmacodynamic properties [9] of mAbs that are related to these glycosylation attributes. While each cell culture underwent exposure to the same 12 OCs, there was notable variation in N-glycan abundance across different biological replicates (BRs) (Supplementary Figures 2, 3, 4, 5). This variability is important as it ensures there is non-redundancy between the train and test set and therefore the predictions on the test set are not easy.

3.3. Media marker selection

A media marker is selected if MS peaks have a CV less than 30 % and its rate has a correlation coefficient (CC) greater than or equal to 0.6 to fucosylation, mannosylation, sialylation, or galactosylation abundances (Fig. 2A). The criteria of CC≥ 0.6 as a feature selector is derived from cut-off thresholds on the training sets where Pearson correlation (R), mean absolute error and root mean squared error start to decline (Supplementary Figure 6). Thus, the criteria for CV ensured consistent reproducibility of the peaks/MMs while the criteria for CC allowed the reduction of selected features while still attaining high training performance (Supplementary Figure 6). From this feature selection criteria, 18 MM variables were selected for fucosylation and mannosylation from 145 possibilities (Fig. 2B; selected MMs). In the case of galactosylation and sialylation, a smaller number of 9 and 10 MMs respectively were selected (Fig. 2B). (Supplementary Figure 6). Although, each of the 18 selected MMs were not characterized in detail, we could conclude that 3

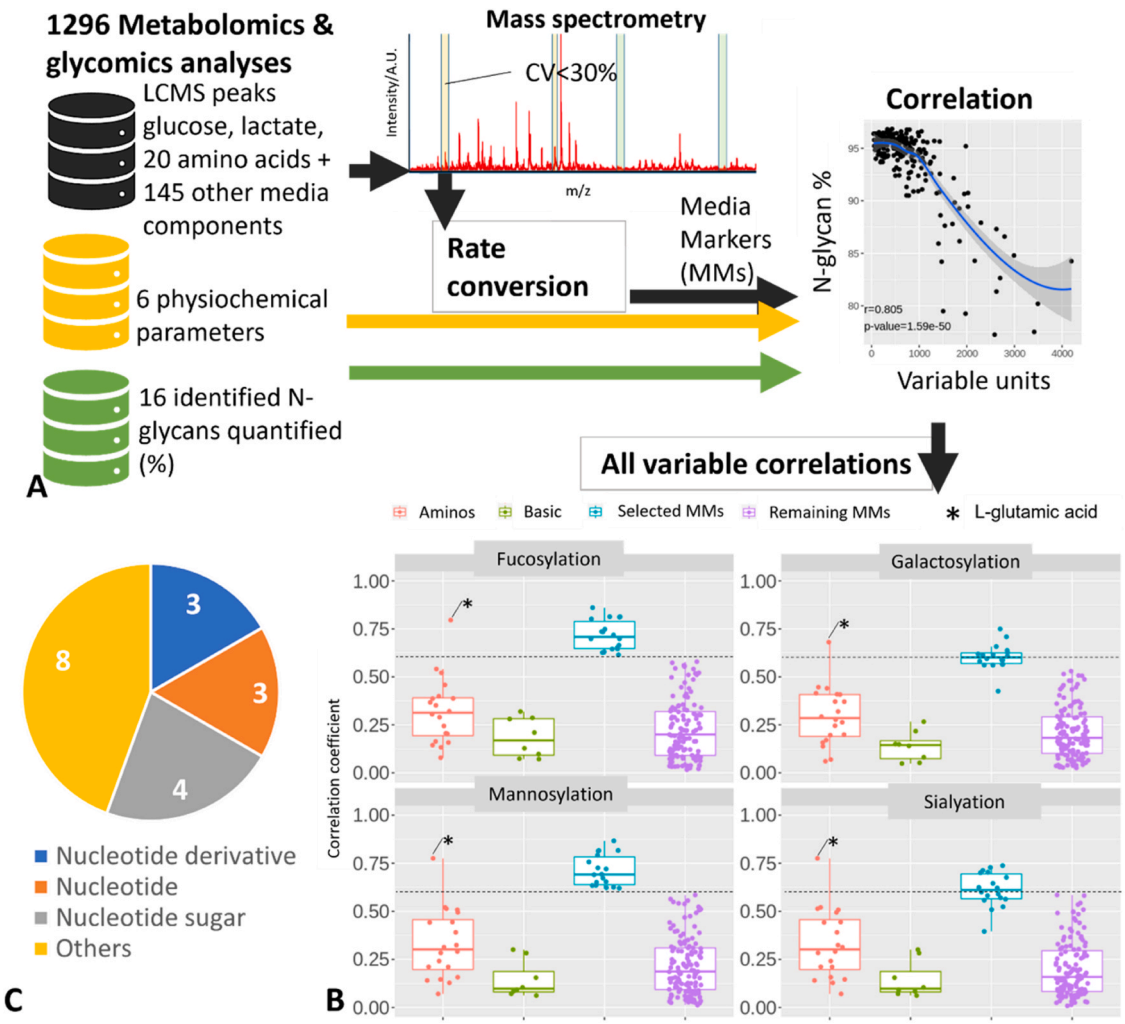

Fig. 2. Feature importance of variables. (A) The workflow for feature selection. MS peaks with coefficient of variation CV < 30 % are converted to rates, known as media markers. Media markers and physicochemical variables are checked for correlation to fucosylation, galactosylation, mannosylation and sialylation N-glycan abundance. Media markers are selected if the correlation coefficient is ≥ 0.6. (B) In the boxplots, each variable's correlation coefficient is plotted, the dashed line represents a correlation coefficient ≥ 0.6. (C) The 18 selected MMs grouped into four categories.

were nucleotide derivatives, 3 were nucleotides and 4 were nucleotide sugars (Fig. 2C). Furthermore, the selected MMs did not contain an amino acid, glucose, lactate, sodium ion, ammonium ion, or potassium ion. Additionally, a total of 127–137 MM variables did not fulfil the selection criteria (i.e. CC<0.6) (Fig. 2B; remaining MMs).

Glucose, lactate and physicochemical parameters like pH, temperature, dissolved oxygen are known to be important for bioreactor optimization [35]. However, we observed that such factors, including pH, temperature, dissolved oxygen, glucose, lactate, sodium ion, ammonium ion, and potassium ion (termed basic), all had poor correlation to the glycosylation patterns of the antibodies, with none having CC≥ 0.6 (Fig. 2B). Recently, CHO-based mechanistic models have capitalized on amino acid metabolism for various bioprocessing applications [36–38], in our dataset we found that only L-glutamic acid had a CC≥ 0.6 to each fucosylation, galactosylation, mannosylation and sialylation output variables (Fig. 2B).

3.4. Grouping variables

While a high correlation coefficient (CC) does not imply causation, optimizing models would be challenging when input variables exhibit only low CCs with the N-glycan output target. Nevertheless, incorporating a mixture of low CC and high CC variables in a model can also be advantageous. Thus, we sought to use different input combinations to understand their corresponding predictive performance of N-glycan abundances using RFs ML.

From the list of all 173 variables available (Table 2), eight different input combinations were used to optimize the models: basic (8 variables; Table 2), amino (20 variables; Table 2), selected MMs (18 variables; Fig. 2B), basic+amino (28 variables), 'basic+selected MMs' (26 variables), 'amino+selected MMs' (38 variables), 'basic+amino+selected MMs' (46 variables), and 'all features' (173 variables).

3.5. Predicting glycan CQA using selected MMs

Predicting individual N-glycan abundances for all 16 N-glycans (Supplementary Figure 1) would require the training of 16 models and the reporting of all 16 output variables. Given that it is the total fucosylation, mannosylation, sialyation and galactosylation levels that are associated with the efficacy of mAbs [39], we summed the individual N-glycan abundances to their overall glycosylation attributes as shown in Supplementary Figure 1. In particular, models were trained to predict four percentage values: total fucosylation, galactosylation, high-mannosylation, and sialylation abundances, which are the main characteristics of N-glycosylation (Supplementary Figure 2). Additionally, classification models were optimized to predict extreme N-glycan

outliers.

3.5.1. Regression – predicting N-glycan abundance

We explored the potential for accurate N-glycan abundance prediction using the selected MMs as input variables to multivariate models such as PLSR (MVDA) and RF (ML) (Fig. 3). We observed significant differences in the performance of the prediction models using different input combinations in both MVDA and ML. The RF algorithm, in particular, generally had a higher correlation coefficient compared to MVDA approaches (Fig. 3). For instance, the correlation coefficient of the ML model using selected MMs alone was greater by 5 %, 32 %, 8 % and 9 % for fucosylation, galactosylation, mannosylation and sialylation, respectively compared to MVDA approaches. Similar differences were observed in the mean absolute error and root mean squared error metrics between both models (Supplementary Figure 7 and 8). For galactosylation, the RF regression model demonstrated significantly greater correlation coefficients ($p < 0.001$; t-test) compared to MVDA across all input variable combinations (values of 0.84 RF vs. 0.52 MVDA, 0.85 RF vs. 0.46 MVDA, 0.86 RF vs. 0.53 MVDA, and 0.85 RF vs. 0.54 MVDA for selected MMs, basic + selected MMs, amino + selected MMs, and basic + amino + selected MMs, respectively). This suggests that ML may be a superior approach compared to the commonly used MVDA in this type of prediction task – particularly when using the 18 selected MMs.

Moreover, the performance of both RF and MVDA models significantly improved when selected MMs were incorporated (Fig. 3 - enriched vs. baseline input). The RF model using selected MMs solely as input demonstrated high correlation coefficients of 0.94, 0.84, 0.94 and 0.80 for fucosylation, galactosylation, mannosylation and sialylation respectively. In contrast, the best baseline input containing 'Basic+Amino' had only correlation coefficients ranging from 0.71 to 0.76. In the case of RF models, correlation coefficients with enriched input combinations were significantly higher compared to those with baseline input combinations ($p < 0.05$ for all glycosylation attributes except sialylation; t-test with Benjamini-Hochberg correction). On the other hand, in MVDA models, we observed that the correlation coefficients were significantly improved only in the cases of fucosylation and mannosylation ($p < 0.05$ for all comparisons; t-test with Benjamini-Hochberg correction).

A final observation was that the performance of both RF and MVDA appeared to decline when using all features (All features; Fig. 3, supplementary Figure 7 and 8), this was especially the case in MVDA. In RF, the decrease with more features could be attributed to overfitting, a situation in which feature selection approaches are known to alleviate

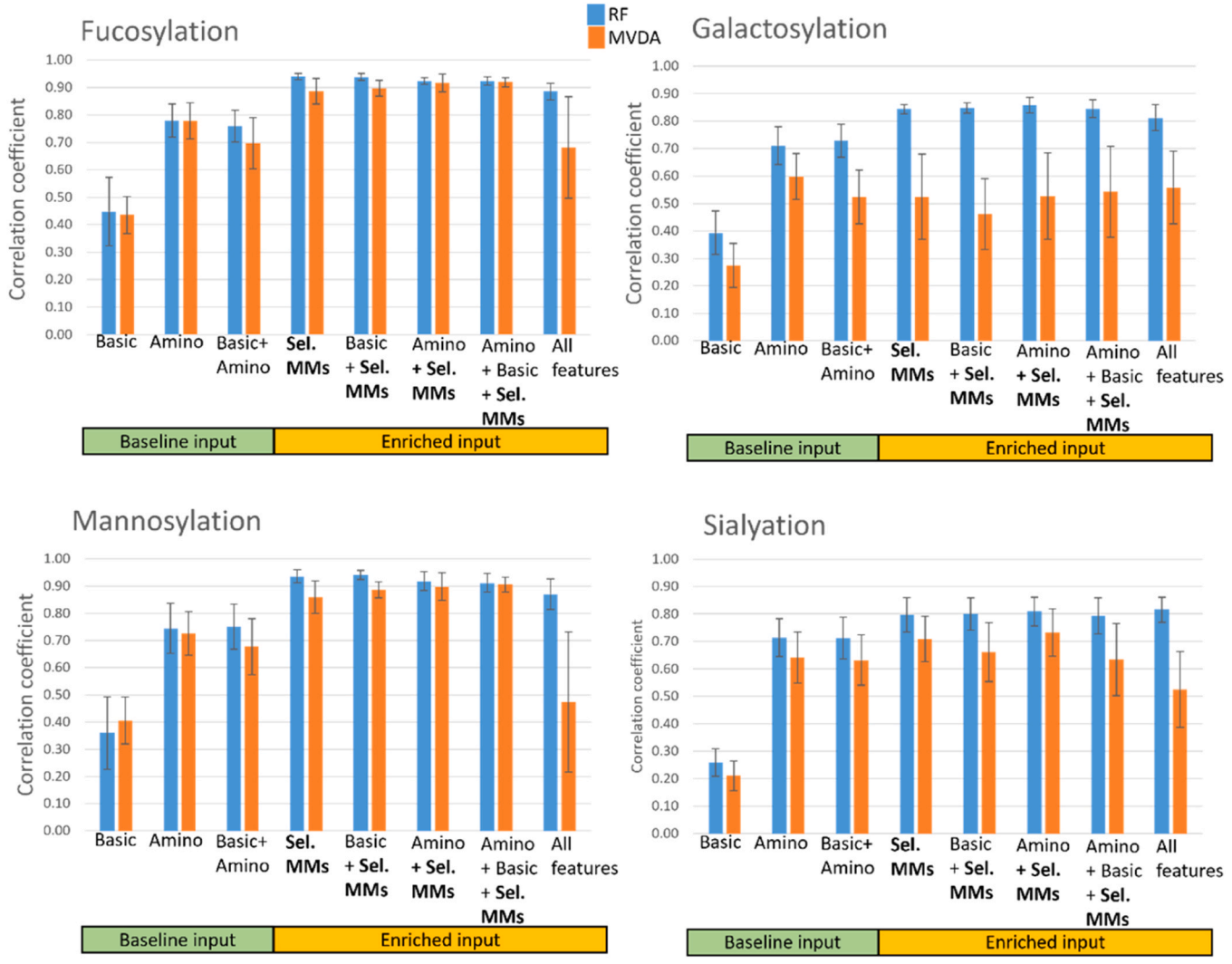

Fig. 3. Comparing regression correlations between RF and MVDA algorithms with different input combinations. Moving left to right along the x-axis the number of input variables increases – in bold are models that use the 18 selected MMs. RF models (orange) were compared to MVDA models (blue) between different input combinations of baseline input (green) and enriched input using selected MMs (yellow). Correlation coefficient is averaged over the three independent test sets and error bars are shown as standard errors.

[40]. In MVDA, the large performance decrease (Fig. 3, Supplementary Figures 7 and 8) when using all features may be due to its inability to find non-linear relationships in large input dimensions. Similar performances were also observed during boot strapped validation performance assessment (Supplementary Table 3 and Supplementary Table 4).

3.5.2. *Classification – predicting outliers*

Besides the prediction of continuous glycosylation abundance, we also sought to develop models that classify N-glycan abundance into outliers or non-outliers (Supplementary Figure 9 A). The usefulness of these models is to analyze whether glycan levels have hit critical configurations through a cut-off threshold (Supplementary Figure 9B). For example, models can be defined to monitor < 90 % fucosylation, < 25 % galactosylation as a critical condition in the media. Moreover, the modelling of classification decision boundaries is easier to achieve than modeling the relationship between input variables and a continuous output variable (e.g. 5.6 % of regression predictions had >5 % MAE; Supplementary Figure 10).

In this case, the RF model using selected MMs as input (i.e. either Selected MMs, Basic+Selected MMs, Amino+Selected MMs, Basic+Amino+Selected MMs) showed better AUC performance compared to MVDA when varying the cut-off thresholds by 1 % increments (Fig. 4). The increase in AUC between RF and MVDA was particularly noticeable for galactosylation predictions, where cut-off thresholds at 17 %, 18 %, 45 %, and $23 - 40$ % produced statistically significant RF classifiers ($p < 0.05$; t-test with Benjamini-Hochberg correction). Choosing just one example, the RF/ML model was able to predict galactosylation outliers that fell below 30 % abundance significantly better than the MVDA model (85.82 vs. 93.79 AUC). Further, the use of the 18 selected MMs (enriched input) significantly improved AUC compared to using basic, amino acid or basic and amino acid combinations (baseline input) (Fig. 5). Thus, the addition of the 18 selected MMs as input is crucial for the accuracy of RF/ML models. In fact, the use of selected MMs was found to also increase the AUC substantially in the case of MVDA also increases AUC substantially models (Fig. 4). Similar performances were also observed during boot strapped validation performance assessment (Supplementary Table 3 and Supplementary Table 4).

4. Discussion

Creating N-glycan prediction models crucially relies on a high throughput and accurate characterization method to generate large training and benchmarking datasets. The work presented here shows that capillary electrophoresis, a labeling approach using APTS, a triple internal standardization of retention times [32] and a computational analysis with database matching [23] enabled the N-glycan characterization of 1296 samples (432 in triplicate). In contrast, such throughput may be challenging for other experimental approaches, such as liquid chromatography mass spectrometry methods. The dataset enabled the optimization of a supervised machine learning approach where experimentally derived fucosylation, galactosylation, mannosylation, and sialyation abundances were the labels used in training. Additionally, to train the supervised RF models each output label must have a corresponding input feature set. To this end, we identify and quantitate the levels of glucose, lactate, sodium ion, potassium ion, ammonium ion and the 20 amino acids as such input features, with an additional 145 MMs that were quantitated. To our knowledge, this dataset is one of the largest available for training models to predict N-glycan abundance.

We demonstrate that regardless of whether ML or MVDA algorithms was employed, the best performing models were those that used 18 selected MMs as input features (Fig. 3 and Fig. 5). In typical media formulations, components can include glucose (the carbon source), amino acids, vitamins, lipids, nucleotides, and nucleotide sugars. While we did not identify the selected MMs in depth for this study, we observed that nucleotide derivatives, nucleotides, and nucleotide sugars were overly represented in these selected MMs (Fig. 2C). Nucleotide derivatives and nucleotides are involved in nucleotide sugar biosynthesis

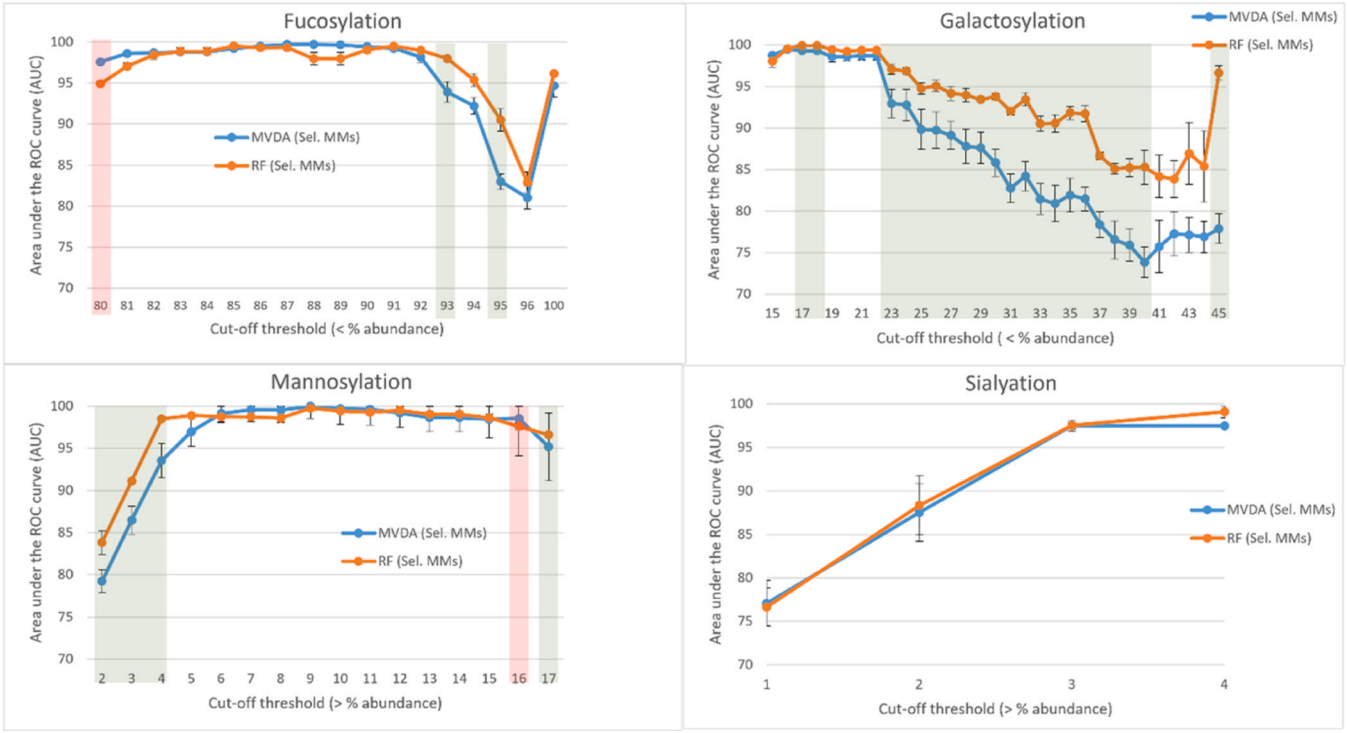

Fig. 4. Leave-one-out cross-validation for RF and MVDA AUC comparison. The AUC is the average of all models that use selected MMs as input (i.e. Selected MMs, Basic+Selected MMs, Amino+Selected MMs, and Basic+Amino+Selected MMs). Statistically significant differences are highlighted by the shaded region ($p < 0.05$; t-test with Benjamini-Hochberg correction; MVDA statistically significant in red shading, RF statistically significant in green shading). The error bars show the standard error between all models in the three test sets.

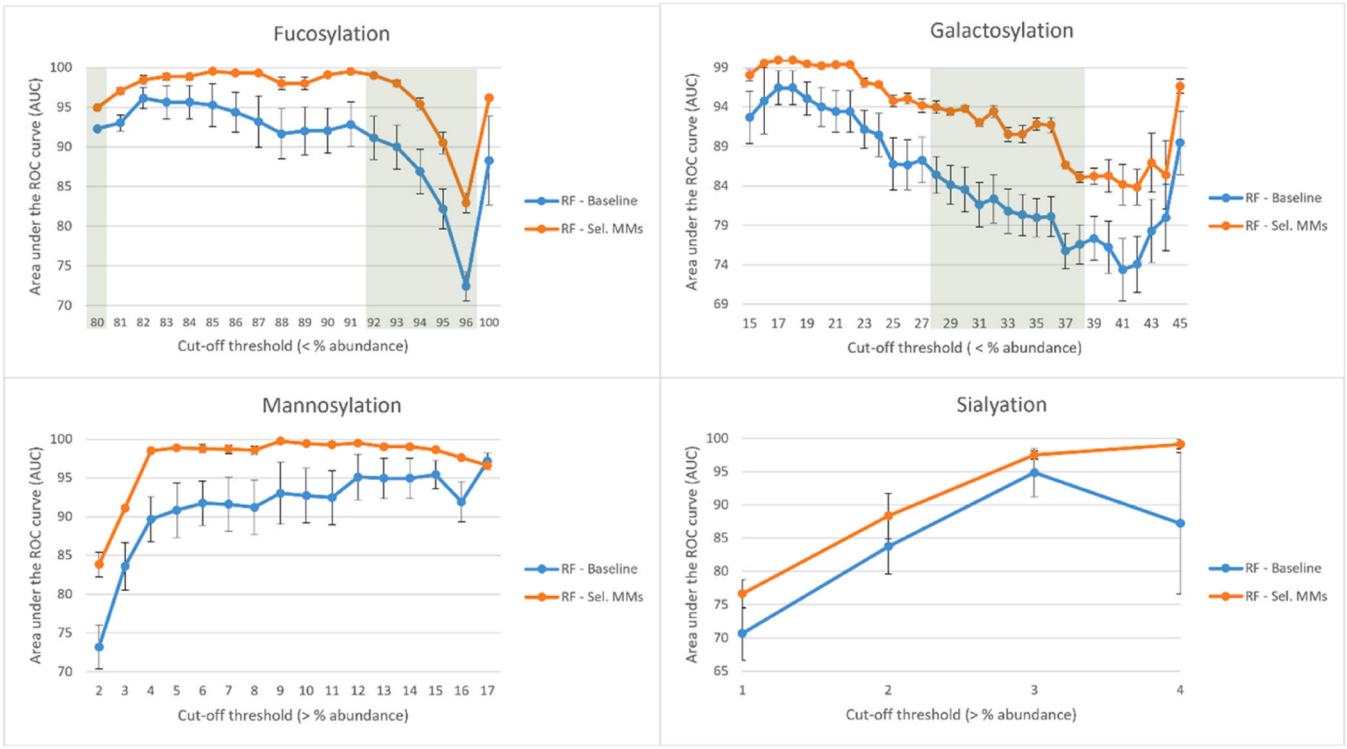

Fig. 5. Leave-one-out cross-validation for RF input type comparison. The AUC is the average of all RF models that use enriched input (either Selected MMs, Basic+Selected MMs, Amino+Selected MMs, and Basic+Amino+Selected MMs) or baseline input (either 'Basic', 'Amino', 'Basic+Amino'). Statistically significant differences are highlighted by the shaded region ($p < 0.05$; t-test with Benjamini-Hochberg correction). The error bars show the standard error between all models in the three test sets.

pathways and nucleotide sugars themselves are well known to be the building blocks for glycan synthesis in the Golgi and to a lesser extent the endoplasmic reticulum [41]. Studies have described possible associations between nucleotide sugar metabolism and CHO N-glycosylation, and how supplementing growth medium with nucleotide-sugar precursors can influence N-glycan abundance Supporting their importance [42,43] Our study has also observed possible associations, though further studies are warranted to unravel such molecular associations.

There is a clear performance boost when predicting galactosylation in ML compared to MVDA using the selected MMs as input. This is observed in both regression (Fig. 3; $p < 0.001$) and classification (Fig. 4; all $p < 0.05$). The enhanced prediction performance in galactosylation may be attributed to the ML's capacity to capture non-linear relationships among input variables and N-glycan abundance. For instance, the levels of the four nucleotide sugars (Fig. 2C) involved in N-glycosylation may be combined to predict galactosylation in a non-linear manner rather than the linear relationships that MVDA can capture. For fucosylation, sialylation and mannosylation, we also observed the superior performance of ML approaches, though to a lesser extent (Fig. 3 and Fig. 4). In summary, RF/ML is the preferred choice over MVDA and the use of the 18 selected MMs is crucial for best prediction performance – particularly for galactosylation.

Model applications. One potential application involves the monitoring of N-glycan CQAs during bioprocessing operations through a proxy analysis that specifically targets 18 mass spectrometry features of the media. Another application of the model could be to prioritize batches before passing them to downstream processing for purification. For example, the classification model could discern cell cultures of highly abundant galactosylation, saving cost by prioritizing the time-consuming N-glycan analytics thus improving complement activation of the mAb [44].

In silico changes to the MMs can used as input to the models to simulate the effects on the glycosylation. This would enable development of optimized feeding algorithms (OFAs) for better N-glycosylation, similar to a previous work [45]. In the context of control, the classification model could oversee a process and identify suboptimal glycosylation quality. Subsequently OFAs can be employed to simulate what to feed and the quantity to feed to rectify the batches quality. Nevertheless, the intricate glycosylation biosynthetic pathway and the influences of cell metabolism pose challenges in controlling glycosylation and further research is imperative. Real-time monitoring applications would be contingent on the development of real-time assays for the 18 peaks associated with the selected MMs. This could involve a real time MS system, microfluidic sample preparation device and automated peak picking combination. That capability would offer an alternative to Raman and Near-infrared spectroscopy real-time monitoring approaches. For instance, offering an orthogonal approach to Raman chemometric models that have shown promise to monitor glycosylation profiles in real-time albeit with scale up issues [46].

Further studies are also warranted to apply such methods beyond CHO cell lines that express Anti-Her2 type mAb products as a general approach to predicting glycosylation patterns. Finally, it was not our aim to create a sophisticated ML algorithm, instead we wanted to show that an "off-the-shelf" RF method can effectively use the 18 selected MMs for glycan profile prediction. This may possibly be further strengthened with the testing of more sophisticated machine learning and feature selection approaches using our dataset.

5. Conclusion

We showed that 18 media markers could be used to accurately predict N-glycan abundance in CHO cell expression of antibodies. The 18 media markers significantly improved both regression and classification performance and were superior to models that used traditional process variables such as pH, temperature and media components like glucose, lactate and amino acids. Notably, ML techniques outperformed MVDA in

improving predictions especially for galactosylation. The models could potentially be used in a bioprocess to determine glycosylation quality offline using spent media as a proxy and to prioritize batches to pass to downstream processing. We believe such approaches would be highly effective for N-glycan CQA analysis and monitoring and can contribute significantly to the overall digitalization efforts in bioprocessing to develop novel therapeutic products for human health.

Ethical considerations

The research described in this manuscript complies with all relevant ethical guidelines and regulations.

Funding

This research is supported by A*STAR (C222812028 awarded to K.T. P.; C210112057 awarded to I.W.) and the Singapore Ministry of Health's National Medical Research Council under its Open Fund-Young Individual Research Grant (OF-YIRG) (MOH-001132-00) (awarded to S.C.).

CRediT authorship contribution statement

Shi Ya Mak: Methodology. **Ying Swan Ho:** Conceptualization, Data curation, Investigation, Methodology, Supervision, Writing – original draft. **Gavin Teo:** Methodology. **Andre Choo:** Methodology, Supervision. **Lyn Chiin Sim:** Methodology. **Farouq Bin Mahfut:** Methodology. **Alison P. Lee:** Methodology. **Hsueh Lee Lim:** Methodology. **Shuwen Chen:** Methodology. **Andy Hee-Meng Tan:** Methodology. **Kuin Tian Pang:** Data curation, Methodology. **Annie Soh:** Methodology. **Sean Chia:** Formal analysis, Methodology, Writing – review & editing. **Yuan sheng Yang:** Methodology. **Meiyappan Lakshmanan:** Conceptualization, Formal analysis, Investigation, Project administration, Visualization, Writing – original draft. **Say Kong Ng:** Methodology. **Ian Walsh:** Conceptualization, Data curation, Formal analysis, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Terry Nguyen-Khuong:** Conceptualization, Data curation, Formal analysis, Methodology, Supervision, Writing – original draft.

Conflict of Interest

The authors declare that they have no competing interests.

Acknowledgements

We thank all members of the GlycoAnalytics at BTI group for useful discussions. We acknowledge support from the Agency for Science, Technology, and Research, Singapore (Z.P., Y.S.Y., Z.W., Y.S.H., I.W., S. C.).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.05.046.

References

[1] Zhang P, Woen S, Wang T, Liau B, Zhao S, Chen C, Yang Y, Song Z, Wormald MR, Yu C. Challenges of glycosylation analysis and control: an integrated approach to producing optimal and consistent therapeutic drugs. Drug Discov Today 2016;21:740–65.
[2] Cobb BA. The history of IgG glycosylation and where we are now. Glycobiology 2020;30:202–13.
[3] Fan Y, Jimenez Del Val I, Müller C, Wagtberg Sen J, Rasmussen SK, Kontoravdi C, Weilguny D, Andersen MR. Amino acid and glucose metabolism in fed-batch CHO cell culture affects antibody production and glycosylation. Biotechnol Bioeng 2015;112:521–35.
[4] Zhou Q, Shankara S, Roy A, Qiu H, Estes S, McVie-Wylie A, Culm-Merdek K, Park A, Pan C, Edmunds T. Development of a simple and rapid method for producing non-fucosylated oligomannose containing antibodies with increased effector function. Biotechnol Bioeng 2008;99:652–65.
[5] Ivarsson M, Villiger TK, Morbidelli M, Soos M. Evaluating the impact of cell culture process parameters on monoclonal antibody N-glycosylation. J Biotechnol 2014;188:88–96.
[6] Thomann M, Reckermann K, Reusch D, Prasser J, Tejada ML. Fc-galactosylation modulates antibody-dependent cellular cytotoxicity of therapeutic antibodies. Mol Immunol 2016;73:69–75.
[7] Pereira NA, Chan KF, Lin PC, Song Z. The "less-is-more" in therapeutic antibodies: afucosylated anti-cancer antibodies with enhanced antibody-dependent cellular cytotoxicity. MAbs, 2018 (Taylor Fr) 2018:693–711.
[8] Goetze AM, Liu YD, Zhang Z, Shah B, Lee E, Bondarenko PV, Flynn GC. High-mannose glycans on the Fc region of therapeutic IgG antibodies increase serum clearance in humans. Glycobiology 2011;21:949–59.
[9] Chia S, Tay SJ, Song Z, Yang Y, Walsh I, Pang KT. Enhancing pharmacokinetic and pharmacodynamic properties of recombinant therapeutic proteins by manipulation of sialic acid content. Biomed Pharmacother 2023;163:114757.
[10] Wells E, Song L, Greer M, Luo Y, Kurian V, Ogunnaike B, Robinson AS. Media supplementation for targeted manipulation of monoclonal antibody galactosylation and fucosylation. Biotechnol Bioeng 2020;117:3310–21.
[11] Gramer MJ, Eckblad JJ, Donahue R, Brown J, Shultz C, Vickerman K, Priem P, van den Bremer ETJ, Gerritsen J, van Berkel PHC. Modulation of antibody galactosylation through feeding of uridine, manganese chloride, and galactose. Biotechnol Bioeng 2011;108:1591–602.
[12] Yin B, Wang Q, Chung CY, Bhattacharya R, Ren X, Tang J, Yarema KJ, Betenbaugh MJ. A novel sugar analog enhances sialic acid production and biotherapeutic sialylation in CHO cells. Biotechnol Bioeng 2017;114:1899–902.
[13] Ha TK, Kim D, Kim CL, Grav LM, Lee GM. Factors affecting the quality of therapeutic proteins in recombinant Chinese hamster ovary cell culture. Biotechnol Adv 2022;54:107831.
[14] Sokolov M, Morbidelli M, Butté A, Souquet J, Broly H. Sequential multivariate cell culture modeling at multiple scales supports systematic shaping of a monoclonal antibody toward a quality target. Biotechnol J 2018;13:1700461.
[15] Powers DN, Trunfio N, Velugula-Yellela SR, Angart P, Faustino A, Agarabi C. Multivariate data analysis of growth medium trends affecting antibody glycosylation. Biotechnol Prog 2020;36:e2903.
[16] Rathore AS. Roadmap for implementation of quality by design (QbD) for biotechnology products. Trends Biotechnol 2009;27:546–53.
[17] Luo Y, Kurian V, Ogunnaike BA. Bioprocess systems analysis, modeling, estimation, and control. Curr Opin Chem Eng 2021;33:100705.
[18] Walsh I, Myint M, Nguyen-Khuong T, Ho YS, Ng SK, Lakshmanan M. Harnessing the potential of machine learning for advancing "quality by design" in biomanufacturing. 1 (Taylor Fr) 2022:2013593.
[19] Rodriguez-Granrose D, Jones A, Loftus H, Tandeski T, Heaton W, Foley KT, Silverman L. Design of experiment (DOE) applied to artificial neural network architecture enables rapid bioprocess improvement. Bioprocess Biosyst Eng 2021;44:1301–8.
[20] Chiappini FA, Teglia CM, Forno ÁG, Goicoechea HC. Modelling of bioprocess non-linear fluorescence data for at-line prediction of etanercept based on artificial neural networks optimized by response surface methodology. Talanta 2020;210:120664.
[21] Nikita S, Thakur G, Jesubalan NG, Kulkarni A, Yezhuvath VB, Rathore AS. AI-ML applications in bioprocessing: ML as an enabler of real time quality prediction in continuous manufacturing of mAbs. Comput Chem Eng 2022;164:107896.
[22] Antonakoudis A, Strain B, Barbosa R, del Val IJ, Kontoravdi C. Synergising stoichiometric modelling with artificial neural networks to predict antibody glycosylation patterns in Chinese hamster ovary cells. Comput Chem Eng 2021;154:107471.
[23] Walsh I, Choo MSF, Chiin SL, Mak A, Tay SJ, Rudd PM, Yuansheng Y, Choo A, Swan HY, Nguyen-Khuong T. Clustering and curation of electropherograms: an efficient method for analyzing large cohorts of capillary electrophoresis glycomic profiles for bioprocessing operations. Beilstein J Org Chem 2020;16:2087–99.
[24] Xu S, Huo J, Huang Y, Aw M, Chen S, Mak S, Yip LY, Ho YS, Ng SW, Tan AH-M. von Hippel-Lindau protein maintains metabolic balance to regulate the survival of naive B lymphocytes. Iscience 2019;17:379–92.
[25] Wang Z, Yip LY, Lee JHJ, Wu Z, Chew HY, Chong PKW, Teo CC, Ang HY-K, Peh KLE, Yuan J. Methionine is a metabolic dependency of tumor-initiating cells. Nat Med 2019;25:825–37.
[26] Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 2006;78:779–87.
[27] Qi Y. Random forest for bioinformatics. Ensemble Mach Learn: Methods Appl 2012:307–23.
[28] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explor Newsl 2009;11:10–8.
[29] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc: Ser B (Methodol 1995;57:289–300.
[30] Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27:861–74.
[31] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinforma 2011;12:1–8.
[32] Jarvas G, Szigeti M, Chapman J, Guttman A. Triple-internal standard based glycan structural assignment method for capillary electrophoresis analysis of carbohydrates. Anal Chem 2016;88:11364–7.

[33] Kanda Y, Yamane-Ohnuki N, Sakai N, Yamano K, Nakano R, Inoue M, Misaka H, Iida S, Wakitani M, Konno Y. Comparison of cell lines for stable production of fucose-negative antibodies with enhanced ADCC (and others) Biotechnol Bioeng 2006;94:680–8.
[34] Natsume A, Niwa R, Satoh M. Improving effector functions of antibodies for cancer treatment: enhancing ADCC and CDC. Drug Des, Dev Ther 2009:7–16.
[35] Li F, Vijayasankaran N, Shen A, Kiss R, Amanullah A. Cell culture processes for monoclonal antibody production. MAbs, 2010 (Taylor Fr) 2010:466–79.
[36] Yeo HC, Park S-Y, Tan T, Ng SK, Lakshmanan M, Lee D-Y. Combined multivariate statistical and flux balance analyses uncover media bottlenecks to the growth and productivity of Chinese hamster ovary cell cultures. Biotechnol Bioeng 2022;119:1740–54.
[37] Hong JK, Nargund S, Lakshmanan M, Kyriakopoulos S, Kim DY, Ang KS, Leong D, Yang Y, Lee D-Y. Comparative phenotypic analysis of CHO clones and culture media for lactate shift. J Biotechnol 2018;283:97–104.
[38] Hefzi H, Ang KS, Hanscho M, Bordbar A, Ruckerbauer D, Lakshmanan M, Orellana CA, Baycin-Hizal D, Huang Y, Ley D. A consensus genome-scale reconstruction of Chinese hamster ovary cell metabolism (and others) Cell Syst 2016;3:434–43.
[39] Wada R, Matsui M, Kawasaki N. Influence of N-glycosylation on effector functions and thermal stability of glycoengineered IgG1 monoclonal antibody with homogeneous glycoforms. 2 (Taylor Fr) 2019:350–72.
[40] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. bioinformatics 2007;23:2507–17.
[41] Freeze, H.H., Boyce, M., Zachara, N.E., Hart, G.W., Schnaar, R.L. (2022). Glycosylation precursors.
[42] Naik HM, Majewska NI, Betenbaugh MJ. Impact of nucleotide sugar metabolism on protein N-glycosylation in Chinese Hamster Ovary (CHO) cell culture. Curr Opin Chem Eng 2018;22:167–76.
[43] Blondeel EJ, Aucoin MG. Supplementing glycosylation: a review of applying nucleotide-sugar precursors to growth medium to affect therapeutic recombinant protein glycoform distributions. Biotechnol Adv 2018;36:1505–23.
[44] Nimmerjahn F, Vidarsson G, Cragg MS. Effect of posttranslational modifications and subclass on IgG activity: from immunity to immunotherapy. Nat Immunol 2023;24:1244–55.
[45] Kotidis P, Jedrzejewski P, Sou SN, Sellick C, Polizzi K, Del Val IJ, Kontoravdi C. Model-based optimization of antibody galactosylation in CHO cell culture. Biotechnol Bioeng 2019;116:1612–26.
[46] A Gibbons L, Rafferty C, Robinson K, Abad M, Maslanka F, Le N, Mo J, Clark K, Madden F, Hayes R. Raman based chemometric model development for glycation and glycosylation real time monitoring in a manufacturing scale CHO cell bioreactor process. Biotechnol Prog 2022;38:e3223.