Probabilistic model by Bayesian network for the prediction of antibody glycosylation in perfusion and fed-batch cell cultures

Liang Zhang1,2 | MingLiang Wang2,3 | Andreas Castan4 | Håkan Hjalmarsson2,3,5 | Veronique Chotteau1,2,5

1Department of Industrial Biotechnology, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH-Royal Institute of Technology, Stockholm, Sweden
2AdBIOPRO, VINNOVA Competence Centre for Advanced Bioproduction by Continuous Processing, KTH Royal Institute of Technology, Stockholm, Sweden
3Division of Decision and Control System, School of Electrical Engineering and Computer Science, KTH-Royal Institute of Technology, Stockholm, Sweden
4Cytiva, Uppsala, Sweden
5Digital Futures - KTH Royal Institute of Technology, Stockholm, Sweden

Abstract

Glycosylation is a critical quality attribute of therapeutic monoclonal antibodies (mAbs). The glycan pattern can have a large impact on the immunological functions, serum half-life and stability. The medium components and cultivation parameters are known to potentially influence the glycosylation profile. Mathematical modelling provides a strategy for rational design and control of the upstream bioprocess. However, the kinetic models usually contain a very large number of unknown parameters, which limit their practical applications. In this article, we consider the metabolic network of N-linked glycosylation as a Bayesian network (BN) and calculate the fluxes of the glycosylation process as joint probability using the culture parameters as inputs. The modelling approach is validated with data of different Chinese hamster ovary cell cultures in pseudo perfusion, perfusion, and fed batch cultures, all showing very good predictive capacities. In cases where a large number of cultivation parameters is available, it is shown here that principal components analysis can efficiently be employed for a dimension reduction of the inputs compared to Pearson correlation analysis and feature importance by decision tree. The present study demonstrates that BN model can be a powerful tool in upstream process and medium development for glycoprotein productions.

KEYWORDS
Bayesian network, Chinese hamster ovary cells, glycosylation, mathematical modelling, monoclonal antibodies

1 | INTRODUCTION

Monoclonal antibodies (mAbs) have been one of the fastest growing pharmaceuticals over the past decades, and represented 65.6% of the total biopharmaceutical global sales in 2017 (Walsh, 2018). Nearly all the therapeutic mAbs contain two N-linked glycosylation sites at position asparagine 297 in the Fc region (Jefferis, 2009; Lalonde & Durocher, 2017). This common posttranslational modification for mammalian cells includes two main steps; the backbones of the glycans are first assembled in the endoplasmic reticulum (ER),
and further matured in the Golgi apparatus. The glycosylation reaction network in the Golgi is highly branched, including the processes of mannose trimming, and additions of N-acetylgalactosamine (GlcNAc), fucose, galactose and sialic acid with help of catalysis of specialized enzymes (Piancin et al., 2016). Unlike other bioreaction networks, usually only a very small proportion of the glycans reach the end of the network, that is, complete maturation, and most of them terminate halfway by secretion out of the cells (Hossler et al., 2006).

Being a key quality attribute of mAbs, the glycosylation patterns have been widely reported to impact the stability, solubility, half-life, and immunological activities of mAbs (Blondeel & Aucoin, 2018; Raju, 2008). For instance, high mannose glycans have shown a decreased circulation time in vivo (Goetze et al., 2011), and the absence of fucose residues is known to significantly increase the antibody-dependent cellular cytotoxicity (ADCC) activity of the immunoglobulin G (IgG; Shields et al., 2002). Consequently, the regulation of mAb glycosylation has attracted a lot of attention these recent years. A large number of studies have shown that the glycoprofile can be influenced by critical process parameters, including medium components (Surve & Gadgil, 2015; A. Zhang et al., 2016; L. Zhang et al., 2019), temperature (Ahn et al., 2008; Sou et al., 2015), pH (Aghamohseni et al., 2014; Brunner et al., 2017), osmolality (Pacis et al., 2011), and dissolved oxygen concentration (Brunner et al., 2017). A challenge is to ensure that a production process of glycoproteins generates the correct or desired glycoprofile. Thanks to continuous culture medium renewal, perfusion processes have been described to be an operation mode able to sustain the profile of posttranslational modification more efficiently than fed-batch mode, which has a dynamic character (Chotteau, 2015).

Mathematical models can be powerful tools for the control and optimization in the biomanufacturing and are receiving a growing interest for this purpose. In the past 20 years, a number of models have been created to describe the glycosylation (Jedrzejewski et al., 2014; Jimenez del Val et al., 2011, 2016; Krambeck & Betenbaugh, 2005), and have contributed to understand the relationship between the cell culture conditions and the glycoprofile (Spahn & Lewis, 2014). However, these detailed kinetic models include a large number of unknown parameters, which significantly increase the model complexity and therefore require a large amount of experimental data for model training. Moreover, some cultivation parameters, like the viable cell density (VCD) or the cell specific perfusion rate (CSPR), cannot easily be used as inputs of these models. This aspect can limit the practical application of these models.

Recently, machine learning has gained popularity in the fields of bioinformatics and computational biology. For instance, artificial neural network (ANN) has been applied to predict the glycoform from secondary protein structures around the glycosylation site (Senger & Karim, 2008). It is well known that ANN, which is a black box model by essence, usually gives accurate models when the data are very numerous and representative of the entire population or model space. Unfortunately, pure black-box models lack biological interpretations, that is, these methods usually cannot provide the intracellular flux changes related to the metabolic reactions, since they use their own networks or algorithm to link the outputs to the inputs instead of metabolic reaction networks.

Bayesian networks (BN) are probabilistic graphic models (PGM) used in a variety of fields including among others climate change and speech recognition (Darwiche, 2010; Marcout & Penman, 2019). By using the knowledge of probability and graph theory, this type of model is able to reflect the interaction of variables in a network (Airoldi, 2007). BN are used to study biological interactions (En Chai et al., 2014; Wilkinson, 2007; Yu et al., 2013). In these studies, each BN node represents a gene expression level or a metabolite concentration of the cells.

The probabilities of nodes in a BN is based on the Bayes’ theorem: \( P(H | X) = \frac{P(X | H) \cdot P(H)}{P(X)} \), where \( P(X) \) represents the probability of observation \( X \), and \( P(H) \) is the prior probability of hypothesis \( H \). \( P(H | X) \) denotes the posterior probability of hypothesis \( H \) given that \( X \) is observed, and \( P(X | H) \) stands for the likelihood that \( X \) is observed given that hypothesis \( H \) holds.

In the present work, rather than using kinetic modeling, we present a BN approach for glycosylation prediction of mAb. The proposed approach relies on the underlying probability distribution and the observed data to provide a probabilistic quantification of the model uncertainty. The benefit of this approach is that this probabilistic quantification is strictly valid. By essence this uncertainty is typical of the biological nature of the glycosylation process, for which the outputs are uncertain. We validate then the model for Chinese hamster ovary (CHO) cells cultures with three different applications: pseudo-perfusion, perfusion culture, and fed-batch culture.

## 2 | MATERIALS AND METHODS

In this study, the PGM was applied to predict the glycosylation profiles in pseudo perfusion cultures (Case study 1), perfusion cultures (Case study 2), and fed batch culture (Case study 3), respectively. The experimental data of the pseudo perfusion cultures and perfusion cultures were taken from our previous study published in Zhang et al. (2020) and from the work of Hutter and collaborators (Hutter et al., 2018), for which the experimental settings have been described in the respective articles. The fed batch cultures were performed according to the set up given below.

### 2.1 | Culture conditions

A CHO-K1 cell line with glutamine synthetase-based (GS) expression (Cytiva) producing a recombinant mAb (IgG1) was used in this experiment. The cells were cultured in 2 L shake flasks with a starting working volume of 200 ml, and a seeding cell density of 0.3 million viable cells per ml (MVC/ml). The culture temperature set point was 37°C and the agitation was 160 rpm. ActiPro medium (Cytiva) was used for the culture. A volume of 20 ml fresh medium without sugars
was fed each day from Day 2 to 8, and 10 ml from Day 9 to 10. Glucose was fed as the only carbon source in the first culture, “Glc,” and both glucose and galactose were fed in the second culture, “Glc + Gal.” The initial glucose concentration was 32 mM in both cultures, and the initial galactose concentration was 4 mM in the second culture. Glucose was fed on Day 8 and 9 when it was depleted, and galactose was fed on Day 5 and 8 when its concentration was low. Both cultures were duplicated.

### 2.2 Antibody purification

To perform the glycosylation analysis, the IgG from the fed batch culture supernatant samples was purified using Protein A magnetic beads (Magic Bioprocessing/Lab On A Bead). 1% (v/v) Protein A magnetic beads was added to the culture supernatant and incubated at room temperature for 2 h. The IgG was then eluted with 60 mM sodium citrate pH = 3.0 from the beads. Finally, the purified IgG was neutralized by 10% 0.2 M (4-2-hydroxyethyl)-1-piperazineethanesulfonic acid) buffer at pH = 7.4.

### 2.3 Analytical methods

The cell density and viability were measured by the cell counter NORMA XS (iPRASENSE). The glucose, galactose, lactate, NH4, and IgG concentrations were determined by Cedex (Roche). The glycosylation profile was analysed using the GlycoWorks Rapi Fluor-MS N-Glycan Kit (Waters) on an Ultra Performance Liquid Chromatography (Waters) with an ACQUITY UPLC Glycan BEH Amide column (Waters).

### 2.4 Data analysis

The feature selection analysis was performed by Python 3.7 with using the function of ExtraTreesRegressor in the sklearn package. The Pearson correlation analysis and the principal component analysis (PCA) were performed by using the function of corrcoef and pca in MATLAB 2018b.

### 3 MODELLING—BN MODEL OF N-LINKED GLYCOSYLATION

In mammalian cells, the glycosylation in the Golgi starts with glycans containing nine mannoses, Man₉GlcNAc₂ (M9), and the glycan mature by trimming of the mannoses followed by addition of the glycan residues GlcNAc, galactose, fucose, and sialic acid. The present BN model approach describes the maturation occurring during the glycosylation by a network of reactions joining critical nodes, which are the glycoforms, denoted “nodes” in the text for reading simplification. During the glycosylation process, a certain percentage of the glycans stop their evolution at various stages before full maturation. Here, to decrease the number of unknown parameters, a simplified glycosylation network described by Hutter et al. (2017, 2018) has been used for the modelling. This model considers only the main glycoforms present on the IgG produced by CHO cells, that is, Man₅/₆, Man₅F+N, G0, G0-N, G0F-N, G1, G0F, G1F, and G2F (Figure 1a), while the glycoforms rarely detected in CHO cell cultures, such as G2 and sialylated glycoforms, are not represented.

In the network, a glycoform, that is, node, can either move to the next node or can be secreted out of the cell. Here by definition, secretion from a node is given the value 0, while evolution to the next glycan is given the Value 1. Figure 1b illustrates our approach zooming into a few reactions of the network. For instance, at node G0F ($N_{G0F}$), a portion of the G0F glycoforms is secreted from the Golgi, while the rest of the G0F receives one galactose and becomes G1F. The percentage of secreted G0F is denoted $P(N_{G0F} = 0)$, and the percentage of G0F further evolving is denoted $P(N_{G0F} = 1)$, thus,

$$P(N_{G0F} = 0) + P(N_{G0F} = 1) = 100\%.$$  \hspace{1cm} (1)

Similarly, a portion of G1F stops evolving and is secreted, while the rest of G1F glycoforms receives another galactose generating G2F. The conditional probability of secreted G1F, that is, the proportion of secreted G1F at node $N_{G1F}$, is given by $P(N_{G1F} = 0 \mid N_{G0F} = 1)$, and the conditional probability of evolving from G1F to G2F is $P(N_{G1F} = 1 \mid N_{G0F} = 1)$. Based on the Bayes’ rule, the joint probability of secreted G1F, that is, the percentage of G1F measured in the glycoprofile, equals:

$$P(N_{G0F} = 1, N_{G1F} = 0) = P(N_{G1F} = 0 \mid N_{GOF} = 1) * P(N_{GOF} = 1)$$  \hspace{1cm} (2)

and the joint probability of G2F formation, that is, the percentage of G2F measured in the glycoprofile, equals:

$$P(N_{G0F} = 1, N_{G1F} = 1) = P(N_{G1F} = 1 \mid N_{GOF} = 1) * P(N_{GOF} = 1)$$  \hspace{1cm} (3)

As previously mentioned, several studies have shown the impact of several cultivation parameters on the glycoprofile, for example, pH, concentration of carbon sources. We assume that the changes of these parameters affect the probability of the glycan evolution at each node. Therefore, these parameters are used as inputs of the Bayesian model. The model outputs are the binary states towards fate of secretion or evolution. Due to the discrete nature of the model output, a sigmoid function can be used to calculate the probability of this binary choice:

$$P(N = i \mid X) = \frac{1}{1 + e^{-(wX + b)}},$$  \hspace{1cm} (4)

where $N = i$ is a switch towards either secretion ($i = 0$) or evolution ($i = 1$), $X$ is a vector representing the culture parameters, for example, sugar concentration, cell specific IgG production rate, pH, temperature. $w$ is a weighting row vector of the influences of the elements of $X$. $b$ is a bias tuned to adjust the model so that the estimated values are fitting the experimental values with an error as little as possible.

Let us illustrate the modelling approach with the example of node of G0F. G0F can be issued from $f11$ or $f12$ and can either be secreted through flux $f15$ or continue to evolve to G1F (flux $f16$).
FIGURE 1  (a) Simplified glycosylation network in the Golgi. The glycans in the Golgi are the critical nodes of the network. At each critical node, the glycan either evolves to the next glycoform (represented by solid line arrows) or stops their maturation and are secreted out of the Golgi (represented by dash line arrows). “f” stands for the flux of each pathway. In (b), parts of the G0F and G1F stop its maturation and are secreted from the critical nodes, with a distribution between the evolving glycans and the secreted glycans assumed to be a sigmoid function. The fluxes of secreted glycans are observed, while the fluxes of the evolving glycans are unobserved. (c) When a critical node has more than two output pathways, we split a critical node into two so that each critical node has only two output pathways to use a binary distribution. (d) Glycosylation network where the node of G0-N and G0 are split according to (c) [Color figure can be viewed at wileyonlinelibrary.com]
TABLE 1  Feeding conditions of the pseudo perfusion—Case study 1

| Mannose (mM) | Galactose (mM) |
|--------------|---------------|
| 1            | 8             |
| 2            | 15            |
| 3            | 25            |
| 4            | 8             |
| 5            | 15            |
| 6            | 25            |
| 7            | 8             |
| 8            | 15            |
| 9            | 25            |
| 10           | 5             |
| 11           | 15            |
| 12           | 5             |
| 13           | 10            |

By definition, the flux of a secreted glycan \( i \), \( f_{i,j} \), is given by:

\[
f_{i,j} = q_{iG} \times F_{G,F} \tag{5}
\]

where \( q_{iG} \) is the specific production of IgG and \( F_{G,F} \) is the fraction of glycan \( i \) among all the secreted glycans. In the present example of excreted GOF, \( f_{15,j} \) is \( f_{15} \) and it equals \( q_{iG} \) multiplied by the percentage of measured GOF in the sample.

\[
f_{15} = q_{iG} \times F_{GOF} \text{ or } F_{GOF} = \frac{f_{15}}{q_{iG}} \tag{6}
\]

Using a sigmoidal function (Equation 4), we can express the conditional probability that GOF becomes secreted as follows:

\[
P(N_{GOF} = 0 | X_{1,GOF}) = \frac{1}{1 + e^{-(X_{1,GOF} + b)}} \tag{7}
\]

where \( X_{1,GOF} \) is the input to the model at this node, which is \([C_{m ann}, C_{gal}, f_{in,GOF}]^T\).

\( f_{in,GOF} \) in \( X_{1,GOF} \) is the total flux entering node GOF, that is, the sum of fluxes \( f_{11} \) and \( f_{12} \) (Figure 1a), so

\[
f_{in,GOF} = f_{11} + f_{12}. \tag{8}
\]

With Bayes’ theorem, we can derive the joint probability of the glycans stopping evolving at the node of GOF, that is, the percentage of secreted GOF among all glycoforms, as follows:

\[
P(N_{GOF} = 0, N_{GOF−N} = 1, N_{GOF} = 1 | X_{1,GOF}) = \frac{1}{1 + e^{-(X_{1,GOF} + b)}} \times \left( \frac{f_{11}}{q_{iG}} + \frac{f_{12}}{q_{iG}} \right). \tag{9}
\]

where \( N_{GOF−N} = 1 \) refers to the proportion of GOF-N becoming to GOF, that is, \( f_{11} \), and \( N_{GOF} = 1 \) refers to the proportion of GOF evolving to GOF, that is, \( f_{12} \). GOF secretion has been expressed by two ways (8) and (11), and consequently

\[
FrGOF = \frac{f_{15}}{q_{iG}} = \frac{1}{1 + e^{-(X_{1,GOF} + b)}} \times \left( f_{11} + f_{12} \right), \tag{10}
\]

so flux \( f_{15} \) equals

\[
f_{15} = \frac{1}{1 + e^{-(X_{1,GOF} + b)}} \times \left( f_{11} + f_{12} \right). \tag{11}
\]

The fluxes \( f_{11} \) and \( f_{12} \) can be derived in a similar approach.

The same principle as exposed here for GOF was applied for the whole network 1.

A further complication for the glycosylation model is that some nodes can have multiple evolution pathways. For instance, for node G0, either a galactose can be added \( f_{13} \) generating G1, or a fucose can be added \( f_{12} \) resulting in GOF, or G0 can be secreted \( f_{14} \), see Figure 1a.

In this case a simple switch is inappropriate. Instead, for such cases, the node is split into two nodes or as many nodes as needed if the branching is more complex. In the present example, node G0 is duplicated into two G0 nodes connected by a new artificial reaction as shown in Figure 1c; from one G0 node, secretion can take place while from the other G0 node evolution towards either GOF or G1 can occur. In final, the network is composed of 10 nodes and 22 fluxes (including the fluxes \( f_{8} \) and \( f_{13} \) generated from the split described above) as shown in Figure 1d.

The unknown parameters \( w \) and \( b \) in the model can be estimated by minimizing the difference between the joint probability of the secreted glycans and their experimental values as follows:

\[
w, b = \arg \min_{w,b} \sum_{j=1}^{n} \left( \left| P_j - R_j \right| \right)^2. \tag{12}
\]

where \( P_j \) denotes the joint probabilities of glycoform \( j \) \((j = 1, ..., n)\), and \( R_j \) is the measured percentage of glycoform \( j \). The parameter values are identified by the function \( f \text{minunc} \) in MATLAB.

4 | RESULTS—CASE STUDIES

4.1 | Case study 1: Pseudo perfusion cultures fed with different combinations of carbon sources

The pseudo perfusion cultures can be found in Zhang et al. (2020). Briefly, CHO cells were cultured in spintubes with 11 ml working volume. The culture medium was daily renewed by centrifugation and the cell pellets were resuspended back to 2 MVC/ml to mimic steady state. Different combinations of mannose and galactose at various concentrations were fed as carbon sources to investigate their impact on the IgG glycosylation patterns, see Table 1. All the cultures were duplicated.

The concentrations of mannose and galactose were inputs of the BN model. The cell specific IgG productivity, \( q_{iG} \), was also considered among the inputs of the BN model as detailed below. The selection of \( q_{iG} \) was motivated by its influence on the glycosylation (Fan, Jimenez Del Val, Muller, Lund, et al., 2015; Jimenez del Val et al., 2016; Trummer et al., 2006).

To model a node, the inputs are the carbon sources and the flux of glycan \( i \), \( f_{in,i} \), coming to this node. Thus, input \( X_{1,i} \) for the node of glycan \( i \) is
where \( C_{\text{man}} \) and \( C_{\text{gal}} \) stand for the average concentrations of mannose and galactose respectively during 1 day.

\[
f_{i,i} \\
\begin{align*}
P(N_{\text{GOF}} = 0|X_{1:GOF}) &= \frac{1}{1 + e^{-n_{i}X_{1:GOF} + b}} \\
X_{1:GOF} \\
f_{i,GOF} \\
f_{N:GOF}
\end{align*}
\]

\[
P(N_{\text{GOF}} = 0, N_{\text{GOF} - N} = 1, N_{\text{GOF} = 1}|X_{1:GOF}) \\
= \frac{1}{1 + e^{-n_{i}X_{1:GOF} + b}} \left( f_{11} \frac{q_{i}\text{G}}{q_{i}\text{G}} + f_{12} \frac{q_{i}\text{G}}{q_{i}\text{G}} \right) \\
f_{15} \\
f_{15} = \frac{1}{1 + e^{-n_{i}X_{1:GOF} + b}} \left( f_{11} + f_{12} \right)
\]

Modelling was performed by segregating the data into a training data set and a testing data set. The model was trained using all the data except the ones corresponding to three feeding conditions, 15 mM mannose, 15 mM galactose, and combination of 15 mM mannose and 15 mM galactose, which were used to test and validate the model. After training, the model gave a satisfactory prediction of the testing data set with an absolute error lower than 1% for each glycoform (Figure 2a,b).

The glycosylation fluxes for each culture condition were simulated by the model and presented in Figure 2c with the values given in Supporting Information Material 1. The simulation indicated that in absence of mannose, the maturation into G1F occurred according to the pathway f1-f3-f6-f11-f16-f19 with fluxes increasing with the galactose concentration while the pathway f1-f4-f8-f13-f17-f19 was favored in presence of mannose, in combination or not with galactose. In other words, in absence of mannose, a higher galactose concentration pushed the pathway towards early fucosylation (f3), while fucosylation took place much later (f17) in the second pathway. This could be associated to the higher consumption rate of mannose than galactose, which probably enhanced the synthesis of UDP-Gal when both carbon sources were used. The Golgi consists of four compartments, cis, medial, trans, and TGN, where the glycosylation takes sequentially place thanks to the glycosylation enzymes differently distributed. The glycosylation model developed by the groups of Kontoravdi assume that FucT is located in medial and trans, and GaT is located in trans and TGN, while the model developed by the group of Lewis assume that FucT is located in medial and GaT is located in trans, which means that the proteins are first fucosylated and then galactosylated (Jimenez del Val et al., 2016; Spahn et al., 2016). In a study to localize these enzymes in HeLa cells, GnTII, and FucT were not included (Rabouille et al., 1995). Fucosylation in plant (Fitchette-Lainé et al., 1994) was localized in trans however the relevance of this information for animal cells is uncertain, and the exact localization of FucT remains to be identified. Notice as well that the localizations of the glycosylation enzymes are not restricted to one compartment but have an uneven distribution in several compartments. The different pathways for the fucosylation taking place early or not obtained in the present model using different substrates are thus plausible.

### 4.2 Case study 2: Perfusion cultures with different cell densities and perfusion rates

The modelling of perfusion cultures was performed using the data of the work by Hutter and collaborators (Hutter et al., 2018).

In this study, the authors studied different cell densities and perfusion rates, listed in Table 2, to investigate how the cultivation settings influenced the IgG glycosylation. Nine culture parameters, including the VCD, perfusion rate (D), CSPR, concentrations of glucose (C\text{glc}), lactate (C\text{lact}), ammonium (C\text{ammon}), and IgG (C\text{IgG}), cell growth rate (\( \mu \)) and specific IgG productivity (\( q_{\text{IgG}} \)) were measured or calculated. All these parameters were potential inputs of the present BN model. An objective was however to avoid a number of inputs too large since the larger the number of parameters, the larger the number of unknown parameters, possibly resulting in overfitting.

Different methods were applied and evaluated for the parameter selection To obtain an objective way to reduce the number of inputs, concentration correlation analysis, feature importance by decision tree and principal components analysis (PCA). In the perfusion culture, each culture condition was maintained during several consecutive days. For each condition, the first 3-4 days were not included in the data set to guarantee the steady state and exclude the data of transient state.

#### 4.2.1 Input parameter selection

**Pearson correlation analysis and feature importance by decision tree**

The Pearson’s correlation coefficient reflects the linear relationship between different variables (Sedgwick, 2012). The feature importance of the decision tree can provide nonlinear relationships between parameters (Saeys et al., 2007). Here we used both methods to select the input parameters the mostly related to the output, that is, the glycoforms. From Figure 3a,b, one can see that \( q_{\text{IgG}} \), C\text{glc}, VCD, and CSPR had the largest impact on the glycosylation with both methods. From a bioprocessing and biological point of view, it can be assumed that the impact of VCD is in fact related to the CSPR, so VCD was excluded. CSPR and C\text{glc} were selected as inputs, as well as the vector f of the coming fluxes at each node. In other word, input \( X_{2J} \) here, was

\[
X_{2J} = [\text{CSPR}, \text{Cglc}, f_{m}]^T
\]
The modelling gave an acceptable curve fitting for most of the culture days (Supporting Information Material 2). To study if this model was predictive, this latter was trained with all the data of the steady state except the condition of VCD 30 MVC/ml and perfusion rate, D, 1.5 reactor volumes per day (RV/day) taken as validation set. As presented in Figure 4, the model gave a prediction with an absolute small error of about 2% for G0F and G1F and about 0.3% for G2F.

### 4.2.2 PCA for the parameter reduction

To carry out the parameter reduction while losing as little information as possible, a third approach was used: PCA. The two eigenvectors, denoted \( E_{V1} \) and \( E_{V2} \), with the highest eigenvalues were used as inputs together with the vector \( f \) of the coming fluxes at each node. The input \( X \), called \( X_{3,i} \) here, was

\[
X_{3,i} = [E_{V1}, E_{V2}, f_{in}]^T. \tag{15}
\]
This approach provided a very good fit of the estimates given by the model for the experimental data. The fit of the model, obtained with all the data, was of higher quality in comparison with modelling using the feature selection as input dimension reduction (Supporting Information Materials 2 and 3). The model was validated in the same way as described above. As shown in Figure 4, the validation test showed an error significantly smaller for the prediction of G0F, G1F, and G2F using $X_3$, with input reduction based on PCA, compared to $X_2$, with input reduction based on feature selection. The fluxes in the glycosylation network simulated by the model for the case of the validation data set with input reduction based on PCA, are shown in Supporting Information Material 4. The simulation of the fluxes showed that high CSPR or low $q_IgG$ could enhance the galactosylation fluxes $f_{13}$, $f_{16}$, and $f_{20}$. For instance Condition 3 with a CSPR of 100 pl/cell/day showed the strongest galactosylation fluxes. A higher CSPR provides more nutrients to the cells and a higher by-products removal, which could potentially improve the synthesis of UDP-Gal and transcription of galactosyltransferase. On the other hand a low $q_IgG$ could lower the burden of galactosyltransfer and indirectly increase the proportion of galactosylated glycans. Finally, the present culture conditions did not have any effects on the fucosylation fluxes ($f_3$, $f_7$, $f_{12}$, $f_{17}$) since fucosylation was saturated in all the conditions.

| Conditions for the perfusion culture—Case study 2 |
|-----------------------------------------------|
| VCD (MVC/ml) | CSPR (pl/cell/day) | $q_IgG$ (pmol/cell/day) |
|---|---|---|
| 1 | 20 | 50 | 9.3 |
| 2 | 20 | 35 | 7.5 |
| 3 | 20 | 100 | 7.2 |
| 4 | 40 | 50 | 8.4 |
| 5 | 30 | 50 | 6.2 |
| 6 | 30 | 33 | 8.1 |
| 7 | 10 | 100 | 6.3 |

4.3 | Case study 3: The fed batch culture with different carbon sources

4.3.1 | Effect of sugars in fed-batch cultures

To investigate the performance of the model in dynamic cultures, two fed batch cultures were carried out in duplicates with identical sugar-free feed medium, supplemented either with glucose (Glc) or glucose and galactose (Glc + Gal), as described in Section 2. Figure 5a-d and Supporting Information Material 5 show the evolution of the main parameters with time in the cultures. The cell density (Figure 5a) reached 8.7 MVC/ml in Glc with glucose supplementation and a slightly higher titer of 10 MVC/ml in Glc + Gal with glucose and galactose supplementations. This was accompanied by IgG accumulation of 410 mg/L in Glc and a slightly higher titer of 445 mg/L in Glc + Gal (Figure 5d).
FIGURE 5  (See caption on next page)
The glucose concentration decreased faster in Glc compared to Glc + Gal, indicating a higher consumption in Glc (Figure 5b). As can be seen in Figure 5c, the cells consumed galactose in Glc + Gal. The concentration of lactate (Figure 5e), up to 12 mM in both cases, does not seem to account for the difference of the maximal cell densities since the highest level of this toxic by-product was observed for the highest cell concentration in Glc + Gal. This difference cannot have been caused by the ammonium either (Figure 5f), since its concentration was not higher than 2.7 mM in both cases. As it can be expected for CHO cells expressing GS, the ammonium level decreased most of the time indicating uptake. However, to the contrary of Glc + Gal culture, in Glc this metabolite switched to production at Day 8. This occurred simultaneously with a switch from lactate production to consumption concurrently with glucose depletion.

In Glc + Gal, supplemented with galactose on Day 5, ammonium production did not occur on Day 8 and the shift towards lactate consumption occurred at Day 9, caused by the very low level of glucose at Day 8 down to 1.5 mM. High supplementation of galactose at Day 8 did not stop the lactate consumption, which was reversed to production (observed at Day 9) only after supplementation of glucose to 5 mM at Day 8. In Glc, supplementation of 5 mM glucose provoked a reduction of the lactate consumption on Day 9 and the much larger supplementation of 32 mM of this substrate caused a switch back to lactate production the day after. Thus, in both feeding regimes, a very low level of glucose on the anterior day provoked a lactate shift independently of the galactose availability.

This indicates as well that the galactose uptake supported the cells with the necessary energy in Glc + Gal culture while the cells in Glc switched their metabolism towards TCA cycle and increased amino acid catabolism as indicated by the ammonia production in Glc and the earlier shift towards lactate consumption. The slightly lower cell densities and IgG accumulation (Figure 4a,b) observed from Day 8 in Glc compared to Glc + Gal were probably caused by the lower energy access in Glc compared to Glc + Gal.

The galactosylation fluxes slightly increased with time probably because of the continuous decrease of cell specific IgG productivity. Furthermore, the galactosylation fluxes were higher in culture Glc + Gal compared to Glc culture. Several studies have reported that the depletion of glucose could affect the galactosylation (Fan, Jimenez Del Val, Muller, Wagberg Sen, et al., 2015; Liu et al., 2014; Villarces et al., 2015), the reduced galactosylation fluxes could be attributed to the glucose shortage on Day 8 in Glc culture. This trend was completely restored to higher levels of G1F and G2F and lower G0F levels on Day 10 thanks to the glucose addition on Day 9. A slight tendency for higher secretion of G0F glycoforms was observed in culture Glc + Gal at Day 9, occurring at the same time as lactate consumption subsequent to glucose depletion. This tendency had completely disappeared at Day 10. The variation of the glycosylation fluxes with time is presented in Figure 5i and Supporting Information Material 6. The fluxes f3-f6-f11-f16 of an early fucosylation were comparable in both cultures while the fluxes f8-f13-f17 (with later fucosylation) diverged to a larger extent between the cultures. For these latter fluxes, these differences correlated with the galactosylation profile with time.

4.3.2 | BN modelling

On the contrary to steady state cultures, the different parameters, that is, cell density, growth rate, IgG productivity, concentrations of carbon sources, lactate, NH4 were changing with time in these dynamic cultures. To deal with this issue, a pseudo steady state assumption was used (Hutter et al., 2017; Zupke & Stephanopoulos, 1994) thus the culture conditions and metabolic flux rates were considered constant during 1 day. In other words, the culture was divided into pseudo steady states, as one state per day, and the average values of the culture parameters over 1 day were used for each state. Here, the pH, concentrations of glucose, galactose, lactate, NH4, cell growth rate, cell density, cell specific consumptions rates of glucose and galactose, and the cell specific productivities of IgG and lactate were used as inputs for modelling. A reduction of the input dimension by PCA had been the most efficient method for Case study 2, Section 4.2.2. For this reason, a similar approach was adopted here. The first two principle components as well as the fluxes of the critical nodes, were used as inputs X of the model.

The BN modelling approach was first applied to all the data of the glycosylation profiles obtained in the fed batch cultures. The simulation gave an accurate fitting for both cultures (Figure 5g-h), and captured well the variations with time including the reverse trend observed at Day 8 and 10 in Glc culture.

To finally validate the modelling approach by cross-validation, the data of the duplicates were considered in two separate data sets, 1 and 2. The data set of duplicate 2 was used to train the model, which was then tested using duplicate 1 set; and duplicate 2 set was used as testing set of the model trained with duplicate 1 set. The experimental data of glycosylation and the corresponding data predicted by the model were then compared (see Figure 6). It can be seen that the model gave a quite...
The glycosylation plays an important role in the quality of therapeutic glycoproteins, and matching the glycoprofile of an originator medicine is essential for development of biosimilars (Kirchhoff et al., 2017). Substantial studies have shown that the cultivation process parameters can affect the glycosylation patterns. A number of glycosylation models have been published in the past two decades. These kinds of tools could contribute to a better control of the glycosylation profile in bioproduction. In some kinetic models, the Golgi apparatus is regarded as a plug flow reactor, and the reaction rates are based on the kinetics of the glycosylation and nucleotide sugar synthesis related enzymes (Jedrzejewski et al., 2014; Jimenez del Val et al., 2011; Karst et al., 2017). The partial differential equations and the large amount of unknown kinetic parameters involved in these models limit their applications for real world scenarios (Spahn & Lewis, 2014). To lower the number of unknown parameters, several non-kinetic models have been recently developed (Hutter et al., 2017; Spahn et al., 2016, 2017). However, most of these models were designed to reflect how the enzymes of the glycosylation affected the glycoprofile. Very few models have focused on the impact of the cultivation parameters on the glycosylation profile, although some of the models could have the potential capacity. As non-kinetic model, the Bayesian model gives posterior probability of a node condition on its parents in the network. In the present work, the percentage of each measured glycoform was treated as the joint probability of the glycoprotein secretion, influenced by the cultivation process parameters. This approach can be a very useful tool for process optimization. Although this model does not explicitly explain the mechanisms of the impact of culture conditions on the glycosylation, it can still give a quantification of the effect of the culture parameters on the glycan patterns by estimating the weight parameters of the model.

Unlike pure black-box models such as ANNs, the present approach enables a probabilistic interpretation of the glycosylation evolution process, and includes the glycosylation maturation mechanism at non-detailed level. As illustrated in Supporting Information Material 1, 4, 6, the model provides the fluxes of the glycosylation network. Here, to simplify the network structure and minimize the burden of parameter estimation, the model only contained the common glycosylation patterns, representing more than 90% of the N-linked glycoforms of IgG produced by CHO cells. The glycosylation network could actually be extended if more complex glycoforms were needed, for example, sialylated glycans in EPO. However, similarly to glycosylation model using the Markov chain (Spahn et al., 2016), as nonparametric inference model, the number of unknown parameters will grow with the network complexity.

The predictive capacity of the model approach was here validated using three different cultivation modes, different CHO cells and different data sets. When a potentially larger number of parameters can be used as input, as Case studies 2 and 3, it is necessary to reduce the input dimension, since the complexity and variance of the model geometrically increase with the dimension of the input, potentially resulting in over-fitting. Three methods were compared for the input selection; either linear correlation by Pearson correlation coefficient, nonlinear relationship by feature selection, or the input eigenvalues obtained by PCA. The PCA-based model showed the best performance in the validation test due to the fact that the eigenvalues included a larger amount of information while the input dimension was maintained low. In other words, with the PCA the reduction of the number of parameters, or dimension reduction, did not imply the same loss of information as the finite selection of the Pearson correlation analysis and feature importance by decision tree. A drawback of the PCA compared to the feature selection or Pearson correlation is that the influence of the culture parameters on the glycosylation flux is not explicitly expressed.

6 | CONCLUSION

In this study, a mathematical model based on a BN approach was created for the prediction of the glycosylation of glycoproteins produced by mammalian cells. The model expressed the percentage of each glycoform secreted out of the cells or evolving to a different glycoform as a
probability affected by the culture conditions. This allowed the model to directly link the culture conditions such as the concentration of the carbon sources and the CSPR, to the glycosylation profile. Among different approaches, PCA used to reduce the number of inputs was the most efficient. The model was characterized by a small number of unknown parameters and a simplified structure, reducing the risk of overfitting. The applications of pseudo perfusion, perfusion and fed batch cultures demonstrated that dynamics could be caught by the present approach. The very good predictive capacity demonstrated in the cross-validation test could be used to support medium development and process optimization.

The BN model could be used to virtually explore the design space. An approach could involve alternating the generation of experimental information with process development performed in silico, which would then be confirmed by experimental work, after which refinement of the BN model and thus of the process would be carried out, and so on.

Another continuation of the present work could be to apply the BN modelling approach to complement a process characterization study. This kind of study typically characterizes a process using statistical tools to determine the range of operation failure of a process. Applying the BN model would generate a mechanistic probabilistic view on the data, which would provide a better understanding of the process and thus potentially a better safety. It could enable a reduction of the experimental efforts, using the model to partially cover the design space, that is, replace experiments by virtual ones.

ACKNOWLEDGEMENTS

This study was supported by the Chinese Scholarship Council, by the Competence Centre for Advanced BioProduction by Continuous Processing, AdBIOPRO, funded by the Sweden’s Innovation Agency VINNOVA, Cytiva, Sweden (diaries nr. 2016-05181), by Digital Futures - KTH Royal Institute of Technology, and by Cytiva.

AUTHOR CONTRIBUTIONS

Liang Zhang: Conceptualization, methodology, investigation, software, writing—original draft, first submission and review. Mingliang Wang: Methodology. Andreas Castan: Resources. Håkan Hjalmarsson: Supervision. Veronique Chotteau: Methodology, supervision, writing—first submission and review.

DATA AVAILABILITY STATEMENT

The code of the Bayesian network programmed with MATLAB is available on https://github.com/liang900530/Bayesian_glycosylation.

ORCID

Andreas Castan https://orcid.org/0000-0002-5210-6558
Håkan Hjalmarsson https://orcid.org/0000-0002-9368-3079
Veronique Chotteau https://orcid.org/0000-0002-5370-4621

REFERENCES

Aghamohseni, H., Ohadi, K., Spearman, M., Krahn, N., Moo-Young, M., Scharer, J. M., Butler, M., & Budman, H. M. (2014). Effects of nutrient levels and average culture pH on the glycosylation pattern of camelid-humanized monoclonal antibody. Journal of Biotechnology, 186, 98–109.

Ahn, W. S., Jeon, J. J., Jeong, Y. R., Lee, S. J., & Yoon, S. K. (2008). Effect of culture temperature on erythropoietin production and glycosylation in a perfusion culture of recombinant CHO cells. Biotechnology and Bioengineering, 101, 1234–1244.

Airolidi, E. M. (2007). Getting started in probabilistic graphical models. PLOS Computational Biology, 3, 3.

Blondeel, E. J. M., & Aucoin, M. G. (2018). Supplementing glycosylation: A review of applying nucleotide-sugar precursors to growth medium to affect therapeutic recombinant protein glycoform distributions. Biotechnology Advances, 36, 1505–1523.

Brunner, M., Fricke, J., Kroll, P., & Herwig, C. (2017). Investigation of the interactions of critical scale-up parameters (pH, pO2 and pCO2) on CHO batch performance and critical quality attributes. Bioprocess and Biosystems Engineering, 40, 251–263.

Chotteau, V. (2015). Perfusion processes, Animal Cell Culture (pp. 407–443). Springer.

Darwiche, A. (2010). Bayesian networks. Communications of the ACM, 53(12), 80–90.

En Chai, L., Saberi Mohamad, M., Deris, S., Khim Chong, C., Wen Choon, Y., & Omatu, S. (2014). Current development and review of dynamic Bayesian network-based methods for inferring gene regulatory networks from gene expression data. Current Bioinformatics, 9, 531–539.

Fan, Y., Jimenez Del Val, I., Muller, C., Lund, A. M., Sen, J. W., Rasmussen, S. K., Kontoravdi, C., Baycin-Hizil, D., Betenbaugh, M. J., Wellungy, D., & Andersen, M. R. (2015b). A multi-pronged investigation into the effect of glucose starvation and culture duration on fed-batch CHO cell culture. Biotechnology and Bioengineering, 112, 2172–2184.

Fan, Y., Jimenez Del Val, I., Muller, C., Wagberg Sen, J., Rasmussen, S. K., Kontoravdi, C., Wellungy, D., & Andersen, M. R. (2015). Amino acid and glucose metabolism in fed-batch CHO cell culture affects antibody production and glycosylation. Biotechnology and Bioengineering, 112, 521–535.

Fitchette-Laine, A. C., Gomord, V., Chekkafi, A., & Faye, L. (1994). Distribution of xylosylation and fucosylation in the plant Golgi apparatus. The Plant Journal, 5(5), 673–682.

Goetze, A. M., Liu, Y. D., Zhang, Z., Shah, B., Lee, E., Bondarenko, P. V., & Flynn, G. C. (2011). High-mannose glycans on the Fc region of therapeutic IgG antibodies increase serum clearance in humans. Glycobiology, 21, 949–959.

Hossler, P., Goh, L. T., Lee, M. M., & Hu, W. S. (2006). GlycoVis: Visualizing glycan distribution in the protein N-glycosylation pathway in mammalian cells. Biotechnology and Bioengineering, 95, 946–960.

Hutter, S., Villiger, T. K., Brühlmann, D., Stettler, M., Broly, H., Soos, M., & Gunawan, R. (2017). Glycosylation flux analysis reveals dynamic changes of intracellular glycosylation flux distribution in Chinese hamster ovary fed-batch cultures. Metabolic Engineering, 43, 2–9.

Hutter, S., Wolf, M., Papili Gao, N., Lepori, D., Schweiger, T., Morbidelli, M., & Gunawan, R. (2018). Glycosylation flux analysis of immunoglobulin G in chinese hamster ovary perfusion cell culture. Processes, 6(10), 176.

Jedrzejewski, P. M., del Val, I. J., Constantinou, A., Dell, A., Haslam, S. M., Polizzi, K. M., & Kontoravdi, C. (2014). Towards controlling the glycoform: A model framework linking extracellular metabolites to antibody glycosylation. International Journal of Molecular Sciences, 15, 4492–4522.

Jeffers, R. (2009). Glycosylation as a strategy to improve antibody-based therapeutics. Nature Reviews Drug Discovery, 8, 226–234.

Jimenez del Val, I., Fan, Y., & Wellungy, D. (2016). Dynamics of immature mAb glycoform secretion during CHO cell culture: An integrated modelling framework. Biotechnology Journal, 11, 610–623.

Jimenez del Val, I., Nagy, J. M., & Kontoravdi, C. (2011). A dynamic mathematical model for monoclonal antibody N-linked glycosylation and nucleotide sugar donor transport within a maturing Golgi apparatus. Biotechnology Progress, 27, 1730–1743.
Kars, D. J., Scibona, E., Serra, E., BIELSER, J. M., SOUQUET, J., STETTLER, M., Broly, H., Soos, M., Morbidelli, M., & Villiger, T. K. (2017). Modulation and modeling of monoclonal antibody N-linked glycosylation in mammalian cell perfusion reactors. Biotechnology and Bioengineering, 114, 1978–1990.

Kirchhoff, C. F., Wang, X. M., Conlon, H. D., Anderson, S., Ryan, A. M., & Bose, A. (2017). Biosimilars: Key regulatory considerations and similarity assessment tools. Biotechnology and Bioengineering, 114, 2696–2705.

Krambeck, F. J., & Betenbaugh, M. J. (2005). A mathematical model of N-linked glycosylation. Biotechnology and Bioengineering, 92, 711–728.

Lalonde, M. E., & Durocher, Y. (2017). Therapeutic glycoprotein production in mammalian cells. Journal of Biotechnology, 251, 128–140.

Liu, B., Spearman, M., Doering, J., Lattova, E., Perreault, H., & Butler, M. (2014). The availability of glucose to CHO cells affects the intracellular lipid-linked oligosaccharide distribution, site occupancy and the N-glycosylation profile of a monoclonal antibody. Journal of Biotechnology, 170, 17–27.

Marcot, B. G., & Penman, T. D. (2019). Advances in Bayesian network modelling: Integration of modelling technologies. Environmental & Modelling Software, 111, 386–393.

Pacié, E., Yu, M., Autsen, J., Bayer, R., & Li, F. (2011). Effects of cell culture conditions on antibody N-linked glycosylation—What affects high mannose 5 glycoform. Biotechnology and Bioengineering, 108, 2348–2358.

Planinc, A., Bones, J., Dejaeger, B., Van Antwerpen, P., & Delporte, C. (2016). Glycan characterization of biopharmaceuticals: Updates and perspectives. Analytica Chimica Acta, 921, 13–27.

Rabouille, C., Hui, N., Hunte, F., Kieckbusch, R., Berger, E. G., Warren, G., & Nilsson, T. (1995). Mapping the distribution of Golgi enzymes involved in the construction of complex oligosaccharides. Journal of Cell Science, 108(4), 1617–1627.

Raju, T. S. (2008). Terminal sugars of Fc glycans influence antibody effector functions of IgGs. Current Opinion in Immunology, 20, 471–478.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23, 2507–2517.

Sedgewick, P. (2012). Pearson’s correlation coefficient. BMJ, 345, e4483.

Senger, R. S., & Karim, M. N. (2008). Prediction of N-linked glycan branching patterns using artificial neural networks. Mathematical Biosciences, 211, 89–104.

Shields, R. L., Lai, J., Keck, R., O’Connell, L. Y., Hong, K., Meng, Y. G., Weikert, S. H., & Presta, L. G. (2002). Lack of fucose on human IgG1 N-linked oligosaccharide improves binding to human Fcgamma RIi and antibody-dependent cellular toxicity. Journal of Biological Chemistry, 277, 26739–26740.

Sou, S. N., Sellick, C., Lee, K., Mason, A., Kyriakopoulos, S., Polizzi, K. M., & Kontaravi, C. (2015). How does mild hyperthermia affect monoclonal antibody glycosylation? Biotechnology and Bioengineering, 112, 1165–1176.

Spahn, P. N., Hansen, A. H., Hansen, H. G., Arnsdorf, J., Kildegaard, H. F., & Lewis, N. E. (2014). Systems glycobiology for glycoengineering. Current Opinion in Biotechnology, 30, 218–224.

Surve, T., & Gadgil, M. (2015). Manganese increases high mannose glycoform on monoclonal antibody expressed in CHO when glucose is absent or limiting: Implications for use of alternate sugars. Biotechnology Progress, 31, 460–467.

Trummer, E., Fauland, K., Seidinger, S., Schriebl, K., Lattenmayer, C., Kunert, R., Vorauer-Uhl, K., Weik, R., Borth, N., Katinger, H., & Muller, D. (2006). Process parameter shifting: Part I. Effect of DOT, pH, and temperature on the performance of Epo-Fc expressing CHO cells cultivated in controlled batch bioreactors. Biotechnology and Bioengineering, 94, 1033–1044.

Villacres, C., Tayi, V. S., Lattova, E., Perreault, H., & Butler, M. (2015). Low glucose depletes glycan precursors, reduces site occupancy and galactosylation of a monoclonal antibody in CHO cell culture. Biotechnology Journal, 10, 1051–1066.

Walsh, G. (2018). Biopharmaceutical benchmarks 2018. Nature Biotechnology, 36, 1136–1145.

Wilkinson, D. J. (2007). Bayesian methods in bioinformatics and computational systems biology. Briefings in Bioinformatics, 8, 109–116.

Yu, D., Kim, M., Xiao, G., & Hwang, T. H. (2013). Review of biological network data and its applications. Genomics & Informatics, 11, 200–210.

Zhang, A., Tsang, V. L., Markely, L. R., Kurt, L., Huang, Y. M., Prajapati, S., & Kshirsagar, R. (2016). Identifying the differences in mechanisms of mycophenolic acid controlling fucose content of glycoproteins expressed in different CHO cell lines. Biotechnology and Bioengineering, 113, 2367–2376.

Zhang, L., Castan, A., Stevenson, J., Chatzissavidou, N., Vilaplana, F., & Chotteau, V. (2019). Combined effects of glycosylation precursors and lactate on the glycoprofile of IgG produced by CHO cells. Journal of Biotechnology, 289, 71–79.

Zhang, L., Wang, M., Castan, A., Stevenson, J., Chatzissavidou, N., Hjalmarsson, H., Vilaplana, F., & Chotteau, V. (2020). Glycan Residues Balance Analysis–GReBA: A novel model for the N-linked glycosylation of IgG produced by CHO cells. Metabolic Engineering, 57, 118–128.

Zupke, C., & Stephanopoulos, G. (1994). Modeling of isotope distributions and intracellular fluxes in metabolic networks using atom mapping matrices. Biotechnology Progress, 10, 489–498.

---

**SUPPORTING INFORMATION**

Additional Supporting Information may be found online in the supporting information tab for this article.

---

**How to cite this article:** Zhang, L., Wang, M., Castan, A., Hjalmarsson, H., & Chotteau, V. (2021). Probabilistic model by Bayesian network for the prediction of antibody glycosylation in perfusion and fed-batch cell cultures. *Biotechnology and Bioengineering*, 1–13. [https://doi.org/10.1002/bit.27769](https://doi.org/10.1002/bit.27769)