A Markov model of glycosylation elucidates isozyme specificity and glycosyltransferase interactions for glycoengineering

Liang, Chenguang; Chiang, Austin W. T.; Hansen, Anders H.; Arnsdorf, Johnny; Schoffelen, Sanne; Sorrentino, James T.; Kellman, Benjamin P.; Bao, Bokan; Voldborg, Bjørn G.; Lewis, Nathan E.

Published in:
Current Research in Biotechnology

Link to article, DOI:
10.1016/j.crbiot.2020.01.001

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
A Markov model of glycosylation elucidates isozyme specificity and glycosyltransferase interactions for glycoengineering

Chenguang Liang a,b,1, Austin W.T. Chiang a,c,1, Anders H. Hansen d, Johnny Arnsdorfd, Sanne Schoffelen d, James T. Sorrentinoa,c,e, Benjamin P. Kellmana,e, Bokan Bao a,c,e, Bjørn G. Voldborg d, Nathan E. Lewis a,b,c,e,f

*Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, USA
*Department of Bioengineering, University of California, La Jolla, CA 92093, USA
fThe Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark

Abstract

Glycosylated biopharmaceuticals are important in the global pharmaceutical market. Despite the importance of their glycan structures, our limited knowledge of the glycosylation machinery still hinders controllability of this critical quality attribute. To facilitate discovery of glycosyltransferase specificity and predict glycoengineering efforts, here we extend the approach to model N-linked protein glycosylation as a Markov process. Our model leverages putative glycosyltransferase (GT) specificity to define the biosynthetic pathways for all measured glycans, and the Markov chain modeling is used to learn glycosyltransferase isoform activities and predict glycosylation following glycosyltransferase knock-in/knockout. We apply our methodology to four different glycoengineered therapeutics (i.e., Rituximab, erythropoietin, Enbrel, and alpha-1 antitrypsin) produced in CHO cells. Our model accurately predicted N-linked glycosylation following glycoengineering and further quantified the impact of glycosyltransferase mutations on reactions catalyzed by other glycosyltransferases. By applying these learned GT-GT interaction rules identified from single glycosyltransferase mutants, our model further predicts the outcome of multi-gene glycosyltransferase mutations on the diverse biotherapeutics. Thus, this modeling approach enables rational glycoengineering and the elucidation of relationships between glycosyltransferases, thereby facilitating biopharmaceutical research and aiding the broader study of glycosylation to elucidate the genetic basis of complex changes in glycosylation.

© 2020 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Glycans are major post-translational modifications, and their structures can directly impact protein characteristics such as binding kinetics, stability, and bioavailability (Pinho and Reis, 2015; Varki et al., 2009; Amann et al., 2019b). Therefore, an understanding of their associated biosynthetic pathways is essential for efforts to modify or engineer glycosylation (Reilly et al., 2019; Moremen et al., 2012; Luo et al., 2017). However, since glycan synthesis is highly stochastic and compartmentalized, real-time observation of the glycosylation process is extremely difficult and further complicated by the dynamic structures of the endoplasmic reticulum and Golgi apparatus (Schwarz and Blower, 2016; Bankaitis et al., 2012). Thus it has been challenging to fully understand the dynamic process of glycan synthesis (Hossler, 2012). Given our incomplete understanding of the glycosylation machinery and the costly and laborious glycomics procedures, predictive computational glycosylation models can be invaluable for capturing the features of the complex glycosylation machinery and to understand how the glycosylation machinery responds to external or internal signals and perturbations.

Over the past two decades, several computational models have been built to quantify and model glycan synthesis (Umaña and Bailey, 1997; Krambeck and Betenbaugh, 2005; Liu et al., 2008; Liu and Neelamegham, 2014; McDonald et al., 2016; Kremskov and Lee, 2018). Recently, a Markov chain method (Spahn et al., 2016; Spahn et al., 2017) was developed for modeling N-linked glycosylation. This approach has the advantage of being a low-parameter framework that does not require kinetic characterization a priori. The Markov chain process effectively captures the sequential and stochastic nature of glycan modification. In the model, each node represents a glycan and the state transitions are the reactions that add a single sugar to the glycan. Thus, the edge weight is a transition probability, which represents the ratio of total flux making a single glycan from a single precursor glycan, divided by the total flux to make all glycans from that...
same precursor. The stationary distribution of a Markov model represents the distribution of all fluxes used to make all measured glycans. One can learn the transition probabilities for each reaction by fitting the model to a single glycoprofile, and subsequently predict changes in glycosylation following glycoengineering. Initial studies have laid the groundwork for this approach, but further work is needed to develop models that are broadly applicable and practical to predict the glycosylation outcome of complex glycoengineering for diverse protein products.

One challenge in model-based glycoengineering is how to account for complex regulatory mechanisms of the glycosylation machinery and accurately define enzyme and isozyme specificity for different glycan substrates. Indeed, glycosyltransferase (GT) isozyme specificity and interactions between glycosyltransferases remain unclear and therefore difficult to model. Recently, studies have confirmed functional interactions among several GT isozymes, wherein one GT impacts the function of another. Examples include interactions between β-1,4-galactosyltransferase (B4gal) and Mannosyl-glycoprotein N-acetylglucosaminyltransferases (Mgat), B4gal and β-1,3-N-acetylglucosaminyltransferase (B3gnt), Mgat and B3gnt, and B4gal and beta-galactosidase alpha-2,3-sialyltransferase (Sialgal) (Bydlnski et al., 2018; Hassinen et al., 2019; Ujita et al., 2000; Mkhikian et al., 2016). Evidence of these interactions has been based on an observed dependency of glycoprofiles or omics data of GT-knockout cell lines (e.g. ST3GAL1 and B4GALT1 interaction (Hassinen et al., 2019)). While these findings suggest GT isozymes interact with each other through direct protein-protein interactions or transcriptional regulation, the specific mechanisms of these interactions and the extent of such interactions have not been extensively studied.

Another significant hurdle for predictive modeling for glycoengineering is our incomplete understanding of GT catalytic specificity. Some glycosyltransferase isozymes, such as those from the B4galt and St3gal families, have more specific catalytic activity on different branches of N glycans (Bydlnski et al., 2018; El-Battari et al., 2003; Rohrfrisch et al., 2006; Mondal et al., 2015; Yang et al., 2015). However, the complex GT-GT interactions, unknown glycan substrate specificities, and the difficulty in obtaining comprehensive omics and enzyme kinetic data, have all presented great challenges to rational model-driven glycoengineering. Therefore, while considerable efforts have been made for predicting glycosylation patterns of recombinant proteins upon the glycoengineered CHO cells (Spahn et al., 2016; Spahn et al., 2017; Krambeck et al., 2017; Chuang et al., 2012), model-based prediction of a glycoengineered glycoprofile from the wildtype glycoprofile is still challenging.

To overcome these challenges, we present a more extensive Markov modeling framework for glycosylation. Specifically, this modeling framework can learn glycosyltransferase activities, including substrate specificities of individual GT isozymes. The methodology was tested on four glycoproteins, including erythropoietin (EPO), Rituximab, Enbrel, and alpha-1-antitrypsin. EPO is a hormone protein widely used for anemia treatment by increasing red blood cell count (Yang et al., 2017b), in which glycosylation play essential roles for its bioactivity (Yang et al., 2017b) and serum half-life (Solá and Griebenow, 2010). We first present models that predicted the N-linked glycosylation of EPO produced by glycoengineered Chinese hamster ovary (CHO) cells with multiple glycosyltransferase isozyme knockouts. The EPO models demonstrated the benefits of introducing substrate specificity. Then, we demonstrated that our EPO models can estimate the isozyme specificity, and we further employed the model to predict the glycoprofiles of multiple glycosyltransferase knockouts. Finally, the EPO model performed similarly to glycoengineered glycoprofiles for three diverse recombinant proteins based solely on the wildtype glycoprofiles for three protein drugs (Rituximab, Enbrel, and alpha-1 antitrypsin) produced by CHO cells. Rituximab is a chimeric monoclonal antibody and specifically binds to CD20 for B-cell lymphoma (Pierpont et al., 2018); Enbrel is a fusion protein of tumor necrosis factor receptor and the Fc part of IgG1, used primarily for treating autoimmune diseases (Montacir et al., 2018); and alpha1 antitrypsin is a protein whose deficiency leads to liver and kidney damage (Blanchard et al., 2011). Studies have shown that glycosylation is extremely important for their functionalities, inflammatory trigger, and other pharmacokinetic/pharmacogenomic properties (Blanchard et al., 2011; Darling et al., 2002; Montacir et al., 2018; Pierpont et al., 2018; Yang et al., 2017b). These results demonstrate that our updated modeling framework provides a valuable approach for rational glycoengineering and for elucidating the relationships among glycosyltransferases, wherein one can discover the genetic basis of complex glycosylation regulatory mechanisms.

2. Results

2.1. A branch-specific N-glycosylation Markov model effectively predicts glycosylation of glycoengineered CHO cells

Here, we present four major changes to the N-glycosylation Markov model (Spahn et al., 2016; Spahn et al., 2017) to overcome the aforementioned challenges (see details in Materials and Methods, Section 5.1). These changes are summarized here: 1) we used a complete glycosyltransferase reaction network, tailored one to fit the EPO glycoprofiles, which enables a more accurate prediction of transition probabilities (TPs); 2) we have deployed the more efficient Pattern Search algorithm for obtaining the best TP vector, instead depending on the COBRA toolbox (Heirendt et al., 2019) 3) instead of optimizing hundreds of transition probabilities for individual reactions in the transition probability matrix (TPX), we optimized only the twenty TPs (see details of the twenty different reaction types in the Materials and Methods, Section 5.1 and Table 1); and 4) instead of a general TP to all branches, we distinguished the TPs for different branches of sialylation, galactosylation, and poly-LacNac elongation (Table 1). Furthermore, these modifications allow us to incorporate unannotated glycan signals and efficiently fit a large network of all theoretically synthesizable glycans to a given glycoprofile.

To test the changes in the modeling framework, we defined two different types of models: a branch-specific model and a branch-general model. The branch-specific model introduced the possibility of branch-specific substrate specificity for each isozyme catalyzing sialylation, galactosylation, and poly-LacNac elongation reactions (see details in Materials and Methods, Section 5.1). Meanwhile, the branch-general model does not distinguish the glycan substrate branches. We tested this updated framework (Fig. 1) on glycoprofiles of erythropoietin (EPO) produced in a panel of glycoengineered Chinese hamster ovary (CHO) cell lines (Yang et al., 2015), compared to the reference WT glycoprofile (i.e., from the EPO-producing non-glycoengineered cell line). For each model-predicted glycoprofile, we evaluated the performance of our framework by two criteria (see details in Materials and Methods): 1) the root mean square error (RMSE) assesses goodness of fit between the model predicted glycan abundance and the experimentally measured glycan abundance; and 2) the coverage, quantifies how many of the experimentally measured glycans were accurately included in our model predictions.

Our newly modified framework demonstrated notable improvements in RMSE and coverage (Fig. 2), due to the inclusion of the possibility for enzymes to exhibit specificity to individual branches in a complex N-glycan. While the branch-specific and branch-general models can fit experimental glycoprofiles well (high density interval (HDI) = 95%), the branch-specific models provided more accurate results. All model-predicted glycoprofiles have significantly reduced RMSEs (mean = 1.1e-2, Std Dev = 3.0e-3) in comparison to those produced by random models (i.e., branch-specific Markov models assigned with random transition probability (TP) vectors, mean = 7.2e-2, Std Dev = 7.2e-3). In addition, they have high coverage (~90% on average) of experimentally measured glycans. Furthermore, introducing branch specificity significantly enhanced the performance of most model predictions of EPO glycoprofiles from the glycoengineered CHO cells, wherein the B3gnt, B4galnt, and St3 gal-family glycosyltransferases were knocked out. For the most improved glycoprofile (i.e., B3gnt2 and Mgat4/4b/5 multiple knockouts; Fig. 2B), the branch-specific model produced significantly enhanced performance (RMSE = 3.8e-3 and coverage = 100%) compared to the branch-general model (RMSE = 1.7e-2 and coverage = 100%). The least improved glycoprofile
by the branch-specific model (RMSE = 1.4e-2 and coverage = 82%) resulted in a significantly decreased performance (two-sample t-test, p < 0.05) compared to the branch-general model (RMSE = 9.7e-3 and coverage = 91%) (B4galT1 knockout; Fig. 2C). We note that the accuracy of knock-out prediction in our model depends on the accuracy and completeness of the knock-out glycoprofile annotation. In this case, the annotation of the B4galT1 was missing annotation of multiple peaks, and there were some peaks that seemed to contradict each other (e.g., triantennary peaks that differed in the branches, Appendix G). These issues could impact the prediction performance of the branch-specific model. However, our method suggests the identities of unannotated peaks and corrections of one annotated peak. This observation requires further validation, and we aim to pursue this systematically in a future study.

Another interesting observation is that model predictions did not significantly improve with the branch-specific models in the Mgt-family knockout samples; however, this is because the Mgt-family glycosyltransferases (Mgt2, Mgt4a, Mgt4b, and Mgt5) are intrinsically branch-specific in that they are responsible for initiating different branches of N-linked glycans. The improved accuracy after introducing branch specificity was consistent with previous reports wherein individual B4galT and St3gal isoforms differentially contributed to galactosylation and sialylation on different branches (Bydlinski et al., 2018; Rohfrisch et al., 2006; Ito et al., 2007). All these results illustrate that the proposed branch-specific framework can more effectively simulate glycosylation of the glycoengineered CHO cells.

2.2. Substrate specificity of glycosyltransferases can be predicted by model transition probabilities

To gain insights into effective glycosylation prediction using the branch-specific models, we closely examined the optimized transition probabilities (TPs) of these models. Each transition probability (TP) is regarded as the probability of transition from one state (substrate) to another (product) for a specific reaction type. The wild-type (WT) model is the basis used to compare with the other glycoengineered models. Therefore, we used the wild-type model to explore if substrate specificity of glycosyltransferases could be described by the TPs. The overall WT model showed a good fit (RMSE = 7.72e-03) and complete (100%) coverage (Fig. 3A), which suggested that the modeling framework could effectively account for the experimental glycoprofile.

Four important findings from the model TPs (Fig. 3B) are as follows. First, the TPs of sialylation on branch 3 and 4 (a3SiaT Branch 3–4) were significantly higher than those on branches 1 and 2 (a3SiaT Branch 1–2), which is consistent with the predominant signals of sialylation on branches 3 and 4 from the experimental glycoprofile. This preferential sialylation on branches 3 and 4 compared to branches 1 and 2 has been previously reported (Bydlinski et al., 2018). Second, the TPs of branch elongation reactions on branches 3 and 4 (iGnT Branch 3–4) are significantly lower than the TPs of sialylation on branches 1–4 (a3SiaT Branch 1–4). This finding was consistent across all KO profiles. Third, the TPs of GnTII branching were considerably higher than those on GnTIV branching, which was consistent with their differentiated enzyme kinetics (Umaña and Bailey, 1997; Krambeck and Betenbaugh, 2005). Lastly, glycosyltransferase reactions showed, in general, much larger (ten fold) TPs than intercompartmental transportation TPs in trans Golgi and secretion, with the exception of LacNac addition. The small TP for LacNac addition is consistent with its small portion of glycans containing poly-LacNac in the experimental profile, and previous reports of poly-LacNac motifs being uncommon in normal mammalian cells (Stanley et al., 2009). The fitted WT model and the consistency between the TPs and the documented glycosyltransferase activities suggested that the optimized TPs quantitatively describe the substrate preferences collectively contributed by all glycosyltransferase isoforms and shed light on the competition between different glycosyltransferase reactions.

2.3. The branch-specific Markov model reveals glycosyltransferase isozyme specificity and co-dependence

Perturbation experiments are widely used to identify potential regulators (e.g., transcriptional regulator), their gene targets, and their regulatory relationships. Here, we employed the same rationale to study how glycosyltransferases regulate N-linked glycan synthesis, using a comprehensive compilation of GT-perturbed glycoprofiles (Yang et al., 2015). Specifically, we systematically quantified the contribution of each GT isozyme to different GT reactions by investigating the impact of a single knockout GT on all other reactions. This was done by computing the fold change of TP vectors between the WT model and the GT-knockout models. A significant interaction between a GT and a reaction is detected if the GT knockout significantly altered both the transition probability (TP) and the reaction flux of the GT-knockout model in comparison with those of the WT model (Materials and Methods, section 5.3).
Our results show the total effects of glycosyltransferases on N-linked glycosylation, as identified by the branch-specific models (Fig. 4; Table D1, Appendix D). Specifically, the loss of function of a glycosyltransferase impacts not only the GT’s primary enzymatic function in glycan synthesis, but also the activities of other GTs beyond their own catalytic function. For example, the Mgat-family glycosyltransferases are the key enzymes responsible for the branching of N-linked glycans. We observed that single gene knockout lines for Mgat2, Mgat4b, or Mgat5 gene significantly impacted their own canonical catalyzed reactions – GnTI, GnTIV and GnTV, respectively (see the highlighted red lines in Fig. 4A (i)). Moreover, for the isozymes of Mgat4a and Mgat4b, our model identified Mgat4b as the major isozyme in catalyzing GlcNAc branching. This is consistent with previous observations wherein Mgat4a showed low gene expression levels in CHO cells, and knocking out Mgat4b led to near complete loss of GlcNAc-α1,4-Man-α1,3 branching (Yang et al., 2015). Besides their own specifically catalyzed reactions, the model captured the GT interactions between Mgat and other GT isozymes (the black lines in Fig. 4A (i)). We found that Mgat4b or Mgat5 significantly increased the poly-LacNAc extension fluxes, in which the Mgat isozymes seem to compete for the same monosaccharides. Specifically, the Mgat4b KO increases iGnT activity (Branch 4) and

---

**Fig. 1.** Glyc prof eres are fit to the Markov model using global optimization with the Pattern Search algorithm. (Start) A list of all possible reactions (including compartment transportation of glycans) involved in the reaction network is generated based on reaction rules from Table 1. The network complexity is restricted by the number of steps required to generate the most complex glycoform in the WT profile. Transition probabilities (TPs) for each enzyme are assigned to each relevant reaction. (Step 1) Given the assigned TPs, an adjacency matrix of transition probabilities (TPX) is constructed to represent the Markov chain process. (Step 2) Given the TPX and a starting flux feeding into the root node (representing the initial glycan Man9GlcNAc2), the predicted glycoprofile is calculated by running the Markov chain model until reaching a stationary flux distribution. (Step 3) The Pattern Search algorithm is used to identify the optimal TP vector by minimizing the RMSE between the predicted glycoprofile and the experimentally measured glycoprofile. The blue dot represents the current TP vector (i.e., polling center), which produced the minimal RMSE = 1.3 from all previous rounds of optimization. The newly selected TP vector (red dot) was identified as the optimal solution (the minimal RMSE = 0.3) for the next round of optimization. (Step 4) The optimization process will be iterated from (Step 1) to (Step 4) until less than 1e-6 RMSE reduction is achieved for 50 consecutive iterations (defined as “convergence”). If the optimization process fails to reach convergence within 1000 iterations or exceed two hours, the current round of optimization will be terminated, and the currently optimized TP vector will be excluded from any further analysis. The resulting optimized TP vectors will be used for further analysis. RMSE: root mean squared error. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
the Mgat5 KO increases iGnT (Branch 3). Indeed, following Mgat gene knockouts, the Golgi can generate glycans of equivalent mass (or monosaccharide composition) to compensate for the loss of GlcNAc branching by extending the poly-LacNAc (Čaval et al., 2018; Goh et al., 2014). Meanwhile, the lack of GlcNAc branching makes existing branches more accessible to subsequent monosaccharide additions. Another possible explanation could be the redistribution of excessive UDP-GlcNAc from med to trans via inter-cisternal tubules (Mkhikian et al., 2016). In addition, the increased sialylation on branch 1 after the Mgat5 knockout was also captured by the model, as reflected by increased free sialyltransferase available to branch 1 following removal of preferentially sialylated branch 4 (Gupta et al., 2016).

The B3gnt-family glycosyltransferases add GlcNAc to the galactose of the N-linked glycans (poly-LacNAc extension). We observed their differentiated catalytic capabilities on LacNAc extension (red lines in Fig. 4B (i)): B3gnt1, B3gnt2 and B3gnt8 single knockout models all carried significantly reduced flux through poly-LacNAc extensions on branch 4 (Fig. D4, Appendix D). The result was consistent with the fact that they all contribute to poly-LacNAc formation in N-linked glycosylation (Mkhikian et al., 2016; Taniguchi et al., 2017; Nielsen et al., 2018). Beyond its direct impact on
the poly-LacNAc extension, a B3gnt1 knockout also significantly resulted in changes in the reactions of branching (GnTIV/V), galactosylation (b4GalT Branch 2/4), and sialylation (a3SiaT Branch 1/2). The discovery is consistent with the finding that the gene products of B4galt1 and B3gnt1 co-localize and physically associate in vivo (Lee et al., 2009; Praissman et al., 2014), and knocking out B3gnt1 will impact B4galt1 activity and all other interacting glycosyltransferases. B3gnt1 knockout further impaired Mgat4 and Mgat5 branching in addition to sialylation on most branches as shown by the modeling result (Fig. D4, Appendix D). Finally, while knocking out both B3gnt1 or B3gnt8 impacted poly-LacNAc elongation, only knocking out B3gnt2 significantly impacted total poly-LacNAc extension flux, resulting in significantly increased sialylation on branch 1 due to diminished competition for St 3 gal isozymes. However, while the reduction of fluxes through iGnT B4 reactions was determined to be statistically
significant (Fig. 4B), the impact of B3gnt2 and B3gnt8 on branch-4 LacNAc extension requires further validation because the sum of fluxes through iGnT B4 reactions is smaller than 2.1% of the total flux in both cases. Similarly, for B3gnt1, the fact that knocking out B3gnt1 impacted reactions beyond poly-LacNAc extension could stem from its interactions with other glycosyltransferases, clonal variation, or phenotypic impacts from the changes in glycosylation.

Other salient findings of interactions for the B4galt and St3gal glycosyltransferases are summarized in Table D1 (Appendix D). Intriguingly, despite that glycosylation has been known as a non-templated glycan...
synthesis process, all these results suggest glycosylation to be a robust cellular process with the mechanism in response to GT knockout. While interactions between different isozymes in the same family and other GTs are complicated, our model TPs and flux variation were highly consistent with the GTs’ known interactive mechanisms or enzyme kinetics. While further experimental validation is required, our
model captured glycosyltransferase isozyme specificity and suggested how glycosyltransferases influence the activities with each other. However, while the experimental annotations are highly consistent with the model-predicted major glycoforms (with the highest model secretion flux among all isomers) predicted at the m/z values (92.9 ± 12.5% accuracy for all fitted glycoprofiles, total flux < 5% for mismatched major glycoforms), glycosidic linkages cannot be assigned by MS in the current setups (Rapiflour LC-MS and Maldi-MS). Although biological knowledge about the glycans of these model proteins can allow experts to manually assign some linkages, the specific positions of galactoses, sialic acids, and LacNAc moieties remain largely uncertain. While future analysis is necessary, we are hopeful that our model can assist in annotating accurate glycosidic linkages to overcome this current characterization limitation of the MS technologies measured glycan composition (m/z). These insights may shed light on the regulation of N-linked glycosylation.

2.4. Glycoprofiles for complex GT mutants can be predicted from single GT knockout models

Genetic interactions complicate the prediction of multi-gene knockout phenotypes, especially when the genes are involved in the same pathway. However, since our modeling framework captures the pathway architecture in N-linked glycosylation, we examined if our models trained on single GT mutants could predict glycoprofiles for mutants with more complex genotypes. Specifically, after obtaining the fitted models of single GT knockouts, we extracted transition probability (TP) vectors from these models and combined them to create new TP vectors, which predicted the GTs' collective influence on the N-glycosylation synthesis for the combinatorial knockout experiments. We developed an algorithm that enabled us to assess the significance of TP fold change vector elements for a multiplex glycoengineered Markov model (Materials and Methods, Section 5.3). Briefly, our algorithm identifies the fitted single-knockout TPs that define the changes in reaction flux following the knockout of an isozyme. It subsequently merges these TPs for all gene knockouts in the more complex mutant to establish a new multi-gene knockout TP vector for glycoprofile prediction.

The predicted glycoprofiles produced by our models showed high consistency with the experimental profiles for the multi-gene knockouts (Figs. E1 and E2, Appendix E). Specifically, glycoprofiles were accurately predicted for eight erythropoietin (EPO) samples, each produced in different glycoengineered CHO cells with different combinations of glycosyltransferases knocked out. The multi-gene knockout models predict glycoprofiles with excellent performance (all log2(RMSEs) < − 5.5, mean log2(RMSE) = − 6.1, log2(RMSE) St. Dev. = 1.1), comparable to (two-tailed t-test, p-value = 0.23) the fitting performance in general (log2(RMSE) RMSE = − 6.6, log2(RMSE) St. Dev. = 0.5). Furthermore, the model reliably predicted glycoprofiles involving major St3gal or B4gal isozyme knockouts, which had remained challenging due to their complicated interactions with the functions of other glycosyltransferases and difficulty in correlating specific isozyme manipulation with model parameters. For example, the double B4gal1/St3gal isozyme knockouts (B4gal1/3 and St3gal3/4; Fig. 5B and C) reduced sialylation even further than B4gal1 knockout alone (Fig. 4B and Fig. E2C, Appendix E), validating the active roles of B4gal1 and B4gal3 in glycosylation despite their lack of impact when they were individually knocked out (Bydlnski et al., 2018). The robust prediction performance further validated the quantification of isozymes' catalytic capabilities by TP vectors and alluded to the model's potential for de novo prediction of biologically accurate glycoprofiles for glycoengineered CHO cell lines. Indeed, by comparing the fitted TPs to the predicted TPs, for each isozyme we identified the fluxes they impacted and quantified their influence on those fluxes. Intriguingly, while B4gal1/2 and B4gal3 only applied small modifications to TPs beyond B4gal1's impact, the predicted glycoprofiles were distinctive from each other and consistent with the fitted results. Therefore, our modeling framework can be used to predict glycoprofiles of multiple glycosyltransferase knockouts using single GT knockout models.

2.5. Glycoprofiles can be predicted for additional glycoengineered drugs de novo, based solely on TP fold changes learned from EPO

Various factors impact the glycoprofile of each unique protein, including protein sequence, structure, post-translational modifications, etc. Thus, it is unclear if glycosyltransferase preferences for one glycoprotein substrate will translate to other protein substrates. Thus, we tested if the EPO-trained models could be generalized to predict the glycoprofiles of other glycoengineered protein drugs (see details in Materials and Methods, Section 5.5) directly from their corresponding wildtype models (see Fig. 6A for procedure). To do this, the modeling framework learns TPs for the wildtype glycoprofiles of a new protein. We hypothesized that the TP fold changes captured by the EPO models are strongly associated with the isozymes' intrinsic catalytic capabilities and are therefore applicable to other protein drugs produced by CHO cells. In particular, N-linked glycosylation for EPO uses a wide variety of glycosyltransferase isoforms from all four families (Mgat- , B4gat- , St3gal-, and B3gnt-family) and produces complex glycoprofiles. This allowed us to extract rich and more complete information regarding the isozyme activities and preferences. Thus, this information could enable the prediction of equally or less complex glycoprofiles of other protein drugs, which may only utilize a subset of glycosyltransferase isoforms.

Testing our hypothesis, we predicted glycoprofiles for three different drugs (Rituximab, alpha-1 antitrypsin, and Enbrel) produced by CHO cell lines with both single and multiplex GT knockouts covering all the four GT families (Fig. 6B–C and FIAB, Appendix F). We found that the predicted KO glycoprofiles demonstrated outstanding performance (all log2(RMSE) < − 4) for both slightly impacted (Rituximab; Fig. 1B, Appendix F) and severely impacted (alpha-1 antitrypsin; Fig. 6C) glycoprofiles, in addition to the highly complex Enbrel glycoprofiles (Fig. 6B and FIAB, Appendix F). Successful prediction of perturbed glycoprofiles of Enbrel and AAT is especially encouraging as their extremely complex WT glycoprofiles. For the glycoengineered Enbrel glycoprofile prediction (Fig. 6B), our model showed that knocking out B3gnt2 and St3gal3/4/6 severely impacts sialylation, which agrees well with the experimental measured glycoprofile (RMSE = 5.85e-02). This result was expected due to the major roles of these St3gal isoforms. Moreover, although we learned from the previous models that B3gnt2 single knockout could decreases LacNAc elongation on branch 4 and activate sialylation on branch one (Fig. 4B), we didn't observe the activated sialylation effect. This observation is not surprising since the knockout of multiple St 3 gal isozymes already eliminated the sialylation, and there was no LacNAc from the WT glycoprofile. For the
Multiple GT knockout glycoprofiles can be predicted de novo for diverse drugs. (A) We established a workflow for de novo model prediction of glycoengineered glycoprofile for drugs, wherein TPs learned from glycoengineered EPO (Fig. 5A) are used to inform changes from WT TPs for any engineered glycoprotein. The multiple GT knockout glycoprofiles for (B) Enbrel and (C) alpha-1 antitrypsin were predicted directly from their corresponding wildtype models by adjusting the TP vector fold changes (isozyme impact) inferred from the EPO models. For Enbrel and Rituximab, the glycoprofiles with Sppl3 single knockout were treated as the wildtype glycoprofile, as it was the base genotype used prior to GT knockouts. For each glycoprofile, at least 90% of the total flux was accounted by present signals. The error bars were calculated as the standard deviations of the glycan intensities produced from 48 iterative runs of the model prediction.
glycoengineered AAT glycoprofile, our model showed that knocking out Mga4a/4b/5 and B4gal1–5 upregulated the only glycan (Glycan #1) but decreased most of the other glycans (Fig. 6C), which is in accordance with the experimentally measured glycoprofile (RMSE = 4.89e-02). Indeed, the Mga4-family glycosyltransferases are responsible for the N-glycan branching, and the B4galt-family glycosyltransferases are responsible for the galactosylation of N-glycans. The model therefore demonstrated that we are able to capture the dominance of this exact glycan (Glycan #1) in this knockout profile. All these results suggest that, with little a priori knowledge, the TP fold changes learned from EPO models could be employed to predict the glycoprofiles of other protein drugs.

3. Discussion

3.1. The low-parameter Markov framework is further simplified for more efficient modeling of glycosylation

Over the past two decades, several mathematical models have provided insights into the complex glycosylation machinery (Hossler, 2012; Krambeck and Betenbaugh, 2005; Krambeck et al., 2017; Kawano et al., 2005; Puri and Neelamegham, 2012). Here, we extended our low-parameter Markov model framework (Spahn et al., 2016) and demonstrated its ability to predict GT substrate specificity and the outcome of multiple glycosyltransferase mutations. This low parameter approach does not require the input of kinetic or concentration information, and we further simplified it by updating the transition probability (TP) formulation only describe the activity of the 20 different glycosyltransferases and glycosidases (the previous formulation considered all transitions at each branch point in the biosynthetic network independently). Note that, the details of these 20 different glycosyltransferases and glycosidases are described in the Materials and Methods Section 5.1 and Table 1. In essence, the updated framework makes strong ties between transition probabilities (TPs) and the enzymes’ catalytic capabilities, which is especially effective for modeling glycoengineered glycoprofiles.

By closely examining the fluxes of glycosylation models, our results demonstrated that the new method comprehensively captures the active parts of the glycosylation network following glycoengineering. For example, our single knockout models (Mga4b and Mga5s) identified significantly increased poly-LacNac extension fluxes, which is consistent with known competition between the Mga isozymes and B3gnt isozymes for the same GlcNac monosaccharides ([Čával et al., 2018; Goh et al., 2014], see Results, Section 2.3). Furthermore, we replaced the original flux variability analysis (FVA) with the efficient global optimization algorithm—Pattern Search. At present, we are able to model a glycoprofile within 2 h for a model with 8435 glycans and 19,719 reactions, which took a few days to complete by using the original FVA optimization algorithm. Both the reduced number of TPs and the new algorithm make the computational time of fitting a large reaction network more practical.

Another common issue in modeling is the overfitting problem. Overfitting is seen when a model fits the training data well but generalizes poorly to new data (Saliccioli et al., 2016). In this study, we addressed the overfitting issue by examining the generalizability of our model in the below two scenarios: 1) predicting multiplex mutants from single knockout models. Specifically, the model parameters (TPs) were trained on the single-knockout EPO glycoprofiles, and they were used to predict the unseen data of the multiple-knockout glycoprofiles for EPO (Results 2.4 and Appendix E); and, 2) predicting the glycoprofiles of different glycoengineered drugs (Rituximab, etuximab, Eshre, and alpha-1 antitrypsin) produced in a different parental CHO cell host (EPO was produced in an adherent CHO-K1 derivative, while the rest were produced in suspension grown CHO-S derivative lines), based solely on TP fold changes trained from EPO (Results 2.5 and Appendix F). Despite the variety of GT knockout combinations and drugs, the previously trained models showed generalizability in predicting the unseen datasets with excellent performance (Figs. 5 B–C, 6 B–C; Appendix E, F), further diminishing the concern for overfitting.

3.2. Computational analyses can unravel multi-glycosyltransferase interactions impacting activities beyond their simple enzyme rules

A critical challenge in developing a predictive glycosylation model lies in the difficulties of quantifying the genetic interactions beyond each GT’s simple enzyme rules. Recently, large amounts of glycoprofiling data were generated from GT knockouts. These data allow us to capture how each perturbed GT impacts the expected activities of other GTs, providing new insights into the genetic interactions between different glycosyltransferases. We presented here a comprehensive documentation of genetic interactions between glycosyltransferases. Importantly, while GTs are expected to be specific toward their own catalytic functions, we show here that knocking out a glycosyltransferase could impact the function of other GTs. For instance, the Mgat2 knockout decreased its own GnTII reaction but promoted the b4GalT–Branch2 reaction (galactosylation). The above findings raise at least two important issues for biotherapeutic glycoengineering applications. The first issue concerns the extent to which potential unintended GT changes (off-target effects) may arise from a specific GT perturbation, and rational glycoengineering of a specific glycoform could be more non-intuitive than we thought. However, as multiple GT mutants are constructed and profiled, computational approaches as presented here can identify and account for genetic interactions, thus helping improve rational glycoengineering of biotherapeutics. Furthermore, such computational analyses can be leveraged to guide research into the underlying molecular mechanisms (e.g., transcription, epigenetic, and feedback loops) regulating GT-GT interactions. Despite that the surrounding literature on these GT-GT interactions (Results 2.3 and Appendix D) appears to be generally compatible with our model predictions obtained in the present study, we should be cautious about the potential clonal variation among the differentially glycoengineered cell lines when interpreting the model-assessed GT-GT interactions, as knocking out these GTs can potentially trigger more profound and diverse changes in cellular phenotypes. Future research is therefore necessary to determine with certainty the exact effect at which a GT-GT interaction in the glycoengineering of CHO cells.

3.3. Predicting glycosylation with minimal a priori knowledge

Another major goal of developing glycosylation models is to provide valuable guidance for glycoengineering therapeutic proteins. The present findings of this research contribute to the field’s understanding of the underlying rules acting on single GT knockout models resulting in a complex GT mutated model, which enables us to predict glycoprofiles of multi-gene mutations. The excellent performance for our model indicates that TP fold changes capture the specificity of each isoyme. These TP values that were learned and quantified from glycoengineered EPO profiles could be combined to predict the glycoprofiles from multi-gene mutants producing distinct glycoproteins, as long as one has the WT glycoprofile for the new protein of interest (Results 2.5). These results lend credence to the hypothesis that the GT interactions are generally encoded in the glycosylation machinery, which could be captured by our glycosylation model. It is apparent that the effect of complex GT knockout strategies impact different biologics in a similar manner. The satisfying accuracy of prediction results and the generalizability of the model pave the way to prospective research for consolidating the study of glycosyltransferase interactions and for rational glycoengineering for better biopharmaceuticals.

3.4. Disentangling the functions of different isoymes

We demonstrated here that model-based analyses can discover or reinforce our understanding of the unique functions of different GT isoymes. We found that there are major isoymes whose knockouts impacted more reactions. Several studies have demonstrated the diversity of GT isoymes. For example, in different mammalian cells, Mga4b is more responsible for the galactosylation of N-glycans. The model therefore demonstrated that knocking out a glycosyltransferase could impact the function of other GTs. For instance, the Mgat2 knockout decreased its own GnTII reaction but promoted the b4GalT–Branch2 reaction (galactosylation). The above findings raise at least two important issues for biotherapeutic glycoengineering applications. The first issue concerns the extent to which potential unintended GT changes (off-target effects) may arise from a specific GT perturbation, and rational glycoengineering of a specific glycoform could be more non-intuitive than we thought. However, as multiple GT mutants are constructed and profiled, computational approaches as presented here can identify and account for genetic interactions, thus helping improve rational glycoengineering of biotherapeutics. Furthermore, such computational analyses can be leveraged to guide research into the underlying molecular mechanisms (e.g., transcription, epigenetic, and feedback loops) regulating GT-GT interactions. Despite that the surrounding literature on these GT-GT interactions (Results 2.3 and Appendix D) appears to be generally compatible with our model predictions obtained in the present study, we should be cautious about the potential clonal variation among the differentially glycoengineered cell lines when interpreting the model-assessed GT-GT interactions, as knocking out these GTs can potentially trigger more profound and diverse changes in cellular phenotypes. Future research is therefore necessary to determine with certainty the exact effect at which a GT-GT interaction in the glycoengineering of CHO cells.

3.3. Predicting glycosylation with minimal a priori knowledge

Another major goal of developing glycosylation models is to provide valuable guidance for glycoengineering therapeutic proteins. The present findings of this research contribute to the field's understanding of the underlying rules acting on single GT knockout models resulting in a complex GT mutated model, which enables us to predict glycoprofiles of multi-gene mutations. The excellent performance for our model indicates that TP fold changes capture the specificity of each isoyme. These TP values that were learned and quantified from glycoengineered EPO profiles could be combined to predict the glycoprofiles from multi-gene mutants producing distinct glycoproteins, as long as one has the WT glycoprofile for the new protein of interest (Results 2.5). These results lend credence to the hypothesis that the GT interactions are generally encoded in the glycosylation machinery, which could be captured by our glycosylation model. It is apparent that the effect of complex GT knockout strategies impact different biologics in a similar manner. The satisfying accuracy of prediction results and the generalizability of the model pave the way to prospective research for consolidating the study of glycosyltransferase interactions and for rational glycoengineering for better biopharmaceuticals.

3.4. Disentangling the functions of different isoymes

We demonstrated here that model-based analyses can discover or reinforce our understanding of the unique functions of different GT isoymes. We found that there are major isoymes whose knockouts impacted more reactions. Several studies have demonstrated the diversity of GT isoymes. For example, in different mammalian cells, Mga4b is more responsible for the GlcNac–β1,4-Man–α1,3 branching (Yang et al., 2015), B4gal1t for galactosylation (Yang et al., 2015; Lee et al., 2001), St 3 gal4 for sialylation (Chung et al., 2015), and B3gnt2 for poly-GlcNac formation (Mikhikian
et al., 2016; Taniguchi et al., 2017; Nielsen et al., 2018). Our glycosylation modeling framework confirmed putative GT specificity but reinforced the dominant role of these major GT isozymes in CHO cells. Furthermore, our results also suggest that different GT isozymes have differences in their functions. For instance, our model suggests that knocking out St3gal6 or St3gal4 had the most severe impact on sialylation (decreased sialylation fluxes by >85%), but knocking out St3gal3 had little influence. These results are in accordance with its primary role for sialylation (Chiang et al., 2015; Jeong et al., 2008). This knowledge is particularly important and could be applied to improve product quality through glycoengineering by being able to partially dialyze down some glycan epitopes. Indeed, sialylation is a key factor in most glycoengineering, since it can improve the serum half-life and activity of these drugs (Chiang et al., 2016). On the other hand, limiting sialylation on monoclonal antibodies (mAb) could enhance antibody-dependent cell-mediated cytoxicity (ADCC) and complement-dependent cytotoxicity (CDC). In these cases, we could consider knocking out a few sialyltransferases (St3gal3, St3gal4, or St3gal6) for better control of the sialylation on mAbs. The proposed model framework thus provides a toolbox that could help identify the best combination of different GT isozymes for desired glycoforms. The more we are able to disentangle the functions of different isozymes, the better we can ultimately control the glycosylation machinery, which should be an important steppingstone toward rational glycoengineering.

4. Conclusions

Here we present a substantial improvement to the Markov chain modeling framework for glycosylation, which accounts for branch-specificity and isoform preference. These refined models effectively simulated the N-glycosylation process of recombinant proteins produced by various glycoengineered CHO cell lines. The essence of our model is transition probabilities, which capture the catalytic capabilities of glycosyltransferase isozymes and quantify the changes in glycosylation after knocking out various isozymes. Exploiting the new modeling framework, we systematically examined the potential interactions between different families of glycosyltransferases and their substrate/branch specificities, which provides insights into the roles of GT isozymes in specific contexts. Our results here further demonstrated that we can predict complex glycoengineered glycoprofles from single-KO models. With the learned fold changes of transition probabilities from EPO, we achieved de novo prediction of GT-KO glycoprofiles of recombinant proteins produced by CHO cells. Therefore, as this framework facilitates rational glycoengineering of various glycosylated protein drugs, it will accelerate the development of effective, safe, and affordable glycosylated biopharmaceuticals.

5. Materials and methods

5.1. Framework of Markov chain model for the N-linked glycosylation

The Markov model of glycosylation is implemented as previously published (Spahn et al., 2016), with a few adaptations described here to improve the fitting to glycoprofles subsequent model predictions (Fig. 1). In essence, this updated Markov model framework can be used for modeling the N-glycosylation process by accounting for all measured and quantified glycans. The new proposed model also provides additional capabilities, such as the means to address glycosyltransferase isozyme specificity and interactions for model-based rational glycoengineering. Here, we highlight four major changes in the newly proposed framework to overcome the aforementioned challenges. First, our updated framework enables the use of a complete glycosyltransferase reaction network rather than a tailored one (i.e., we do not trim out unannotated glycans), which enables us to account for all measured glycans and to fit the model with more accurate transition probabilities (TPs) (see details in the Discussions Section 3). Second, instead of using the COBRA toolbox (Heirrendt et al., 2019), we have deployed the Pattern Search algorithm (MATLAB 2018b, Global Optimization Toolbox) for obtaining the best TP vector. Briefly, the algorithm employed a GPS-like searching strategies (Lageveen-Kammeijer et al., 2019; Rios and Sahinidis, 2013), which iteratively samples the solution space with increasingly higher resolution. Specifically, it first creates a coarse-grained grid of points (a matrix of sampled solutions–TP vectors) centered at the current TP vector and observing whether the objective function value improves or worsens at each of the grid points. The best solution will serve as the new center in the next iteration and the algorithm samples new solutions. If no new solutions are better than the current center-point solution, the sampling grid will be shrunk to a fine-grained grid of points by decreasing the Euclidean distances (among the fixed number of grid points) and look for new solution points. Such process is repeated until the convergence criteria (RMSE change < 1e-6 for consecutive 50 iterations) is met. Furthermore, the algorithm constrains the optimization problem by the augmented Lagrangian method (Lageveen-Kammeijer et al., 2019), which solves a series of unconstrained problem with penalty and a Lagrange function instead of the constrained problem (Lageveen-Kammeijer et al., 2019). The Lagrange function allows the approximation of unknown function gradients from the linear combination of the constraint gradients at stationary points satisfying the constraints (Kolda et al., 2006; Deb and Srivastava, 2012). This well-established, derivative-free global optimization algorithm has been known for its excellent optimization performance in efficient convergence and effective identification of global extrema in a high-dimensional solution space (Conn et al., 1991; Lewis et al., 2007; García-Ródenas et al., 2019). Third, instead of optimizing hundreds of transition probabilities in the transition probability matrix (TPX) by using the COBRA framework (Spahn et al., 2016; Megchelenbrink et al., 2014), only the twenty TPs are defined, corresponding to the twenty different reaction types (17 glycosidases and glycosyltransferase reactions listed in Table 1 and three Golgi intercompartmental transport reactions), which were optimized by the Pattern Search algorithm. Fourth, the TPs for sialylation, galactosylation, and poly-LacNac elongation were further distinguished by the branch on which the corresponding monosaccharides were added (Table 1). The reaction rules were compiled and curated for consistency based on previous publications on Markov or kinetic-based models (Krambeck and Betenbaugh, 2005; Liu and Neelamegham, 2014; Spahn et al., 2016; Krambeck et al., 2017; Hou et al., 2016; Krambeck et al., 2009). Notably, unlike all previously published models, the reaction constraint for α6FucT was removed from its reaction rule as new studies have confirmed the feasibility and presence of fucosylation without the presence of α-1,3-branched (Branch 1/3) GloCNAc moiety (Arédovol and Rovira, 2015; Yang et al., 2017a; Castillo et al., 2015). For branch-general models, substrate branches were not distinguished for B4GalT (BX), a3SiaT (BX), and iGnT (BX) (10 reaction types), resulting in only B4GalT, a3SiaT, and iGnT reaction types (3 reaction types). ’X’ denotes branches numbered 1, 2, 3, or 4, which represent GNb2|Ma3, GNb4|Ma3, GNb2|Ma6, and GNb6|Ma6 respectively.

5.2. Model evaluation metrics – RMSE and coverage

Two model evaluation metrics were used for evaluating the performance of our models. The first one is the root mean squared error (RMSE) for assessing the goodness of fit between the model-predicted glycan intensities and the experimentally measured glycan intensities. The experimental glycoprofles were fitted by minimizing the RMSE of TP vectors between the model prediction glycoprofile and the experimental glycoprofile. The RMSE was calculated by Eq. (1), where N represents the number of all glycan compositions (m/z values or retention time points) in an experimental glycoprofile. \( y_{pred}(t_{exp}) \) represents the predicted (experimentally measured) signal intensity measured at the ith m/z value or at the ith retention time for the LC data. Note that, the glycans predicted but without experimental signals were also considered for RMSE calculation by setting their experimental signals to be 0.

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_{pred}(t_{exp}) - y_{exp})^2}
\]

Statistical significance was further assessed using the highest density interval (HDI), wherein the statistical meaning of HDI = 95% is that the two
groups of tested models are significantly different with a 95% confidence interval (for details see Appendix A).

Another model evaluation metric is ‘coverage’ for assessing how many of the experimentally measured glycans were accurately included among the glycans predicted by our framework. For an experimental glycoprofile, the m/z values corresponding to glycans with the top signal intensities and collectively representing at least 90% of the total signal intensity were selected as experimentally detected glycans. The coverage was defined as the ratio of these glycan compositions that can be captured by the glycoprofiles predicted by the Markov models (branch-specific and branch-general models).

5.3. Predicting multiple GT knockouts from single GT knockout models

The TP vector for a given multiple knockout glycoprofile was derived from the TP vectors of the relevant fitted single-knockout glycoprofiles. Four criteria were used to define the significance of TP vector elements for a multiplex glycoengineered Markov model. Specifically, the fitted single-knockout TPs are required for substantiating the impact of knocking out an isoform on the reactions listed in Table 2. First, the TP fold change of reaction i after knocking out glycosyltransferase k must be statistically different from 0 (i.e., the 95% highest density interval (HDI) does not include 0 from the BEST analysis, as described in Appendix A). Assessment of the statistical credibility of flux and TP using Bayesian estimation is described in Appendix A (Beerli, 2006; Kruschke, 2013; Kruschke and Liddell, 2018; Muller et al., 2006; Winter, 2019). Second, the mean flux fold change of reaction i, after knocking out glycosyltransferase k, must have a scaling factor of at least 1.5 fold (\(\log_2(\text{mean flux fold change}) \geq 0.588\)), and the mean flux fold change ± one standard deviation does not include 1. Then, two additional criteria were established for predicting a new TP for a glycoprofile with combinatorial glycosyltransferase knockouts. Third, if all isoforms of the same family are knocked out, the TP fold changes of the associated direct reaction(s) will be reduced to at most −10 (eliminating fluxes of direct reactions). Fourth, \(\log_2(\text{flux fold change})\) and \(\log_2(\text{TP fold change})\) must have the same sign for the KO model of glycosyltransferase k. These four criteria were applied in Eqs. (2)-(3) for deriving the final combined TP vectors:

\[
\log_2(FC(TP_{k,i})) = \sum_k \log_2(FC(TP_{k,l})) + \frac{1}{A_k} \sum_i \log_2(FC(TP_{k,i}))
\]

(2)

\[
FC(TP_{FKsex}) = 0, \text{ if any of the four criteria are not met.}
\]

(3)

Briefly, the fold change of the transition probability values, \(FC(TP_{k,i})\), is defined as the TP fold change of reaction i, which is the reaction (denoted as ‘F’) directly catalyzed by GT-isozyme k, whereas \(FC(TP_{k,l})\) is another reaction (denoted as ‘S’) potentially impacted by GT-isozyme k knockout. Table 2 listed the reactions directly catalyzed by a given enzyme based on their known reaction rules. The potentially impacted reactions are all the other reactions not directly influenced by the GT-isozyme k knockout, which can be indirectly influenced by either kinetic or genetic interactions of these GT’s (i.e. B4galt and Mgatal). \(A_k\) is the number of non-zero \(FC(TP_{k,l})\) and \(FC(TP_{k,i})\) is the TP fold change of reaction i for the predicted multiple glycosyltransferase knockout glycoprofile. FC (Fold change) is defined as the TP of reaction i for the fitted WT divided by the predicted multiple GT-KO glycoprofiles. The derived (predicted) TP vector for a combined GT-KO Markov model was then assigned to the initial TPX, which was used in models to predict the multiple knockout glycoprofile (Fig. 1 B and C). Here, nonparametric cosine similarity is used to measure how similar between two vectors (predicted and fitted) for fluxes and TPs. Specifically, it measures the cosine of the angle between two vectors, and a smaller angle means higher similarity.

5.4. Protein purification and glycan analysis for additional glycoengineered drugs

5.4.1. GT-knockout cell line generation and model protein expression

Glyco gene knockout cells used for the expression of EPO were derived from CHO-K1 cell line with glutamine synthetase knocked out, and glycoprofiled in a previous study (Čaval et al., 2019). Here we conducted further glycosyltransferase knockouts for the cells expressing Rituximab, alpha-1-antitrypsin, and Enbrel. These lines were derived from the CHO-S cell line (Gibco Cat. # A11557-01), and they were generated and verified according to the procedures described previously (Amann et al., 2019a). Cells were cultured in CD CHO medium (Gibco 10743-029) supplemented with 8 mM l-glutamine (Lonza BE17-605F) and 2 mL/L of anti-clumping agent (Gibco 0010057AE) according to the Gibco guidelines. The day prior to transfection, cells were washed and cultured in exponential phase in medium not supplemented with anti-clumping agent. At the day of transfection, viable cell density was adjusted to 800,000 cells/mL in 125 mL shake flasks (Corning 431143) containing 30 mL medium only supplemented with 8 mM l-glutamine. Plasmids encoding for Rituximab, Enbrel, and alpha-1-antitrypsin, respectively, were used for transient transfections. For each transfection, 30 mg plasmid was diluted in OptiPro SFM (Gibco 12309019) to a final volume of 750 uL. Separately, 90 uL FuGene HD reagent (Promega E2311) was diluted in 660 uL OptiPro SFM. The plasmid/OptiPro SFM mixture was added to the FuGENE HD/OptiPro SFM mixture and incubated at room temperature for 5 min. The resultant 1.5 mL plasmid/lipid mixture was added dropwise to the cells. Supernatants containing model protein were harvested after 72 h by centrifugation of cell culture at 1000g for 10 min and stored at −80 °C until purification and N-glycan analysis.

5.4.2. Protein purification and N-glycan labeling

Rituximab and Enbrel were purified by protein A affinity chromatography. A 5-mL MABSelect column (GE Healthcare) was equilibrated with 5 column volumes (CV) of 20 mM sodium phosphate, 0.15 M NaCl, pH 7.2. Following column equilibration, the supernatant was loaded, the column was washed with 8 CV of 20 mM sodium phosphate, 0.15 M NaCl, pH 7.2, and the protein was eluted using 0.1 M citrate, pH 3.0. Elution fractions (0.5 mL) were collected in deep-well plates containing 60 μL of 1 M Tris, pH 9 per well. Alpha-1-antitrypsin, C-terminally tagged with the HPC4 tag (amino acids EQDVPRLIDNG), was purified by over a 1-mL-column of anti-protein C affinity matrix according to the manufacturer’s protocol (Roche, cat. no. 11815024001). 1 mM CaCl2 was added to the supernatants, equilibrium buffer and wash buffer. The protein was eluted in 0.5 mL fractions using 5 mM EDTA in the elution buffer. For all three proteins, elution fractions containing the highest concentration of protein were concentrated ten-fold using Amicon Ultra 0.5-mL centrifugal filter units (MWCO 10 kDa).

5.4.3. N-glycan analysis

For Rituximab, Enbrel, and alpha-1-antitrypsin, 12 μL of concentrated protein solutions (concentrations varying between 0.1 and 1 mg/mL) were subjected to N-glycan labeling using the GlycoWorks RapiFluor-MS N-Glycan Kit (Waters). Labeled N-glycans were analyzed by LC-MS as described previously (Amann et al., 2019a). Initial conditions 25% 50 mM ammonium formate buffer 75% Acetonitrile, separation gradient from 30% to 43% buffer. MS were run in positive mode, no source fragmentation. The normalized, relative amount of the N-glycans is calculated from the area under the peak with Thermo Xcalibur software (Thermo Fisher Scientific).

5.5. Framework of de novo prediction of glycoengineered glycoprofiles for diverse glycoengineered drugs (Enbrel, rituximab, and alpha-1-antitrypsin) from their corresponding wildtype glycoprofiles

The wildtype glycoprofile of new drug X (produced by wildtype CHO-S cells) was first obtained by fitting the model to its experimental glycoprofile
as described in Materials and Methods Section 5.1. Meanwhile, we quantified the TP fold changes for each single GT knockout by fitting their experimentally measured EPO glycosylations. Then, to assess the impact of a desired combination of GT knockouts on drug X's glycosylation, we quantified the total impact of these knockouts as TP fold changes estimated by the algorithm described in Materials and Methods Section 5.3. Finally, the predicted TP fold changes were applied to the TPs of drug X's wildtype models, resulting in predictive models for the glycosylation of drug X with the given GT knockouts.

### Author contribution statement

**Chenguang Liang:** conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft, writing – review editing, visualization

**Austin W.T. Chiang:** conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft, writing – review editing, visualization

**Anders H. Hansen:** writing – review editing, validation, investigation, data curation

**Johnny Arnsdorf:** investigation, validation, data curation, writing - original draft

**Sanne Schoffelen:** investigation, validation, data curation, writing - original draft

**James T. Sorrentino:** investigation

**Benjamin P. Killman:** data curation

**Bokan Bao:** validation, data curation

**Bjorn G. Voldborg:** resources, supervision, project administration

**Nathan E. Lewis:** conceptualization, methodology, resources, writing - review & editing, supervision, project administration, funding acquisition

### Acknowledgements

This work was conducted with support from the Novo Nordisk Foundation provided to the Center for BioSustainability at the Technical University of Denmark (NNF10CC1016517: C.L., A.W.T.C., A.H.H., B.G.V.) and NIGMS (R35 GM19850: N.E.L.)

### Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

### Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crbiot.2020.01.001.

### References


