Title

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

Authors

Liang, Chenguang

Chiang, Austin

Hansen, Anders

et al.

Publication Date

2020-11-01

DOI

10.1016/j.crbiot.2020.01.001

HHS Public Access

Author manuscript

Curr Res Biotechnol. Author manuscript; available in PMC 2020 November 01.

Published in final edited form as:

Curr Res Biotechnol. 2020 November ; 2: 22–36. doi:10.1016/j.crbiot.2020.01.001.

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

Chenguang Liang[1,2,+], **Austin W.T. Chiang**[1,3,+], **Anders H. Hansen**[4], **Johnny Arnsdorf**[4], **Sanne Schoffelen**[4], **James T. Sorrentino**[1,3,5], **Benjamin P. Kellman**[1,2,5], **Bokan Bao**[1,3,5], **Bjørn G. Voldborg**[4], **Nathan E. Lewis**[1,2,3,*]

[1]Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, USA

[2]Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA

[3]The Novo Nordisk Foundation Center for Biosustainability at the University of California, San Diego, La Jolla, CA 92093, USA

[4]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark

[5]Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, CA 92093, USA

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 modelling 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.

[*]Corresponding author: Nathan E. Lewis, 9500 Gilman Drive MC 0760, La Jolla, CA 92093, nlewisres@ucsd.edu.
[+]These authors contributed equally to this work

Keywords

Glycosylation model; glycomics; systems glycobiology; glycoengineering; isozyme specificity; glycosyltransferase interactions

1. Introduction

Glycans are major post-translational modifications, and their structures can directly impact protein characteristics such as binding kinetics, stability, and bioavailability [1, 2, 62]. Therefore, an understanding of their associated biosynthetic pathways is essential for efforts to modify or engineer glycosylation [3–5]. 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 [6, 7]. Thus it has been challenging to fully understand the dynamic process of glycan synthesis [8]. 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 [9–14]. Recently, a Markov chain method [15, 16] was developed for modelling 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 (B4galt) and Mannosyl-glycoprotein N-acetylglucosaminyltransferases (Mgat), B4galt and β−1,3-N-acetylglucosaminyltransferase (B3gnt), Mgat and B3gnt, and B4galt and beta-galactoside alpha-2,3-sialyltransferase

(St3gal) [17–20]. 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 [18]). While these findings suggested 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 [17, 21–24]. 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 [15, 16, 25, 26], 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 [69], in which glycosylation play essential roles for its bioactivity [69] and serum half-life [73]. 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, we show our model effectively predicts 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 [68]; Enbrel is a fusion protein of tumor necrosis factor receptor and the Fc part of IgG1, used primarily for treating autoimmune diseases [67]; and alpha1 antitrypsin is a protein whose deficiency leads to liver and kidney damage [66]. Studies have shown that glycosylation is extremely important for their functionalities, inflammatory trigger, and other pharmacokinetic/pharmacogenomic properties [66–69]. 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 [15, 16] 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 rather than a 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 [41]; 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 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 (Figure 1) on glycoprofiles of erythropoietin (EPO) produced in a panel of glycoengineered Chinese hamster ovary (CHO) cell lines [24], 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 squared 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 (Figure 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-,

B4galt-, and St3gal-family glycosyltransferases were knocked out. For the most improved glycoprofile (i.e., B3gnt2 and Mgat4a/4b/5 multiple knockouts; Figure 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; Figure 2C). We note that the accuracy of knock-out prediction in our model depends on the accuracy and comprehensiveness 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), and 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. However, this observation would require 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 Mgat-family knockout samples; however, this is because the Mgat-family glycosyltransferases (Mgat2, Mgat4a, Mgat4b, and Mgat5) 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 isozymes differentially contributed to galactosylation and sialylation on different branches [17, 22, 27]. 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 (Figure 3A), which suggested that the modeling framework could effectively account for the experimental glycoprofile.

Four important findings from the model TPs (Figure 3B) are as follow. 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 [17]. 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 [9, 10]. Lastly, glycosyltransferase reactions showed, in general, much larger (tenfold) 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 [28]. 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 isozymes 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 [24]. 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 (Figure 4; Table F1, 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 – GnTII, GnTIV and GnTV, respectively (see the highlighted red lines in Figure 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-β1,4-Man-α1,3 branching [24]. Besides their own specifically catalyzed reactions, the model captured the GT interactions between Mgat and other GT isozymes (the black lines in Figure 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 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 [29, 30]. 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 [20]. 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 [31].

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 Figure 4B (i)): B3gnt1, B3gnt2 and B3gnt8 single knockout models all carried significantly reduced flux through poly-LacNAc extensions on branch 4 (Figure D4, Appendix D). The result was consistent with the fact that they all contribute to poly-LacNAc formation in N-linked glycosylation [20, 32, 33]. 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* [34, 35], 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 (Figure 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 St3gal isozymes. However, while the reduction of fluxes through iGnT B4 reactions was determined to be statistically significant (Figure 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 F1 (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 isoforms) 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 combinatory 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 (Figures G1 and G2, 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 $\log_2$(RMSEs) < −5.5, mean $\log_2$(RMSE) = −6.1, $\log_2$(RMSE) St. Dev. = 1.1), comparable to (two-tailed t test, p-val = 0.23) the fitting performance in general ($\log_2$(RMSE) RMSE = −6.6, $\log_2$(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 B4galt/St3gal isozyme knockouts (B4galt1/3 and St3gal3/4; Figures 5B and 5C) reduced sialylation even further than B4galt1 knockout alone (Figure 4B and Figure E2C, Appendix E), validating the active roles of B4galt2 and B4galt3 in galactosylation despite their lack of impact when they were individually knocked out [17]. 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 B4galt2 and B4galt3 only applied small modifications to TPs beyond

B4galt1'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 Figure 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 isozymes 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 isozymes.

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 (Figures 6B–C and F1A–B). We found that the predicted KO glycoprofiles demonstrated outstanding performance (all $\log_2(\text{RMSE}) < -4$) for both slightly impacted (Rituximab; Figure F1B) and severely impacted (alpha-1 antitrypsin; Figures 6C) glycoprofiles, in addition to the highly complex Enbrel glycoprofiles (Figures 6B and F1A). Successful prediction of perturbed glycoprofiles of Enbrel and AAT is especially encouraging as their extremely complex WT glycoprofiles. For the glycoengineered Enbrel glycoprofile prediction (Figure 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 isozymes. 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 (Figure 4B), we didn't observe the activated sialylation effect. This observation is not surprising since the knockout of multiple St3gal isozymes already eliminated the sialylation, and there was no LacNAc from the WT glycoprofile. For the glycoengineered AAT glycoprofile, our model showed that knocking out Mgat4a/4b/5 and B4galt1–5 upregulated the only glycan (Glycan #1) but decreased most of the other glycans (Figure 6C), which is in accordance with the experimentally measured glycoprofile (RMSE=4.89e-02). Indeed, the Mgat-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 [8, 10, 25, 36, 37]. Here, we extended our low-parameter Markov model framework [15] and demonstrated its ability to predict GT substrate specificity and the outcome of multiplex 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 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 (Mgat4b and Mgat5) identified significantly increased poly-LacNAc extension fluxes, which is consistent with known competition between the Mgat isozymes and B3gnt isozymes for the same GlcNAc monosaccharides ([29, 30], 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 hours for a model with 8,435 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 [70]. 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, erythropoietin, Enbrel, 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 (Figure 5BC, 6BC; 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 multiplex 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 isozyme. 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 isozymes

We demonstrated here that model-based analyses can discover or reinforce our understanding of the unique functions of different GT isozymes. We found that there are major isozymes whose knockouts impacted more reactions. Several studies have demonstrated the diversity of GT isozymes. For example, in different mammalian cells, Mgat4b is more responsible for the GlcNAc-β1,4-Man-α1,3 branching [24], B4galt1 for galactosylation [24, 38], St3gal4 for sialylation [39], and B3gnt2 for poly-GlcNAc formation [20, 32, 33]. Our glycosylation modelling 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 [39]. This knowledge is particularly important and could be applied to improve product quality through glycoengineering by being able to partially dial 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 [40]. On the other hand, limiting sialylation on monoclonal antibodies (mAb) could enhance antibody-dependent cell-mediated cytotoxicity (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 isozyme 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 glycoprofiles from single-KO models. With the learned fold changes of transition probabilities from EPO, we achieved *de novo* prediction of GT-KO glycoprofiles directly from their WT glycoprofiles for new protein drugs 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 [15], with a few adaptations described here to improve the fitting to glycoprofiles subsequent model predictions (Figure 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). Second, instead of using the COBRA toolbox [41], 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 [64], 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 shrunken 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 [64], which solves a series of unconstrained problem with penalty and a Lagrange function instead of the constrained problem [64]. The Lagrange function allows the approximation of unknown function gradients from the linear combination of the constraint gradients at stationary points satisfying the constraints [45, 46]. 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-dimension solution space [42–44]. Third, instead of optimizing hundreds of transition probabilities in the transition probability matrix (TPX) by using the COBRA framework [15, 47], 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 [10, 12, 15, 25, 48, 49]. Notably, unlike all previously published models, the reaction constraint for a6FucT was removed from its reaction rule as new studies have confirmed the feasibility and presence of fucosylation without the presence of α−1,3-branch (Branch 1/3) GlcNAc moiety [50–52].

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 glycoprofiles were fitted by minimizing the RMSE of TP vectors between the model prediction glycoprofile and the experimental glycoprofile. The RMSE was calculated by equation 1, where N represents the number of all glycan compositions (m/z values or retention time points) in an experimental glycoprofile. $y_{pre,i}$ ($y_{exp,i}$) 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.

$$RMSE = \sqrt{\sum_i^N \frac{(y_{pre,i} - y_{exp,i})^2}{N}} \quad (1)$$

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 isozyme 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. 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 ($|\log2(\text{mean flux fold change})| \geq 0.58$), 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 isozymes of the same family are knocked out, the TP $\log_2$ fold changes of the associated direct reaction(s) will be reduced to at most −10 (eliminating fluxes of direct reactions). Fourth, $\log_2$(flux fold change) and $\log_2$(TP fold change) must have the same sign for the KO model of glycosyltransferase *k*. These four criteria were applied in equations 2–3for deriving the final combined TP vectors:

$$log_2\left(FC\left(TP_{Ci,k}\right)\right) = \sum_k log_2\left(FC\left(TP_{Fi,k}\right)\right) + \frac{1}{A_i}\sum_k log_2\left(FC\left(TP_{Si,k}\right)\right) \quad (2)$$

$$FC\left(TP_{Fi,k}\right) and\, FC\left(TP_{Si,k}\right)) = 0,\ if\ any\ of\ the\ four\ criteria\ are\ not\ met\ . \quad (3)$$

Briefly, the fold change of the transition probability values, $FC(TP_{Fi,k})$, is defined as the TP fold change of reaction *i*, which is the reaction (denoted as 'F') directly catalyzed by GT-isozyme *k*, whereas $FC(TP_{Si,k})$ 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 GTs (i.e. B4galt and Mgat4). A_i is the number of non-zero $FC(TP_{Si,k})$, and $FC(TP_{Ci,k})$ 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 (Figures 1B and 1C). 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

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 [29]. 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 [53]. 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 ug 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 minutes. The resultant 1.5 mL plasmid/lipid mixture was added dropwise to the cells. Supernatants containing model protein were harvested after 72h by centrifugation of cell culture at 1,000g for 10 minutes and stored at −80°C until purification and N-glycan analysis.

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 EDQVDPRLIDGK), was purified over a 1-mL column of anti-protein C affinity matrix according to the manufacturer's protocol (Roche, cat. no. 11815024001). 1 mM $CaCl_2$ was added to the supernatants, equilibration 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).

N-glycan analysis.—For Rituximab, Enbrel, and alpha-1-antitripsin, 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 [53]. 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 5.1. Meanwhile, we quantified the TP fold changes for each single GT knockout by fitting their experimental measured EPO glycoprofiles. Then, to assess the impact of a desired combination of GT knockouts on drug X's glycoprofile, we quantified the total impact of these knockouts as TP fold changes estimated by the algorithm described in Materials and Methods 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 glycoprofile of drug X with the given GT knockouts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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: A.L., A.W.T.C., A.H.H., B.G.V.) and NIGMS (R35 GM119850: N.E.L.)

References

[1]. Pinho SS, Reis CA. Glycosylation in cancer: mechanisms and clinical implications. Nat Rev Cancer. 2015;15:540–55. doi:10.1038/nrc3982. [PubMed: 26289314]

[2]. Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Bertozzi CR, et al., editors. Essentials of Glycobiology. 2nd edition Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2009 http://www.ncbi.nlm.nih.gov/books/NBK1908/. Accessed 29 Jun 2019.

[3]. Reily C, Stewart TJ, Renfrow MB, Novak J. Glycosylation in health and disease. Nat Rev Nephrol. 2019;:1. doi:10.1038/s41581-019-0129-4.

[4]. Moremen KW, Tiemeyer M, Nairn AV. Vertebrate protein glycosylation: diversity, synthesis and function. Nat Rev Mol Cell Biol. 2012;13:448–62. doi:10.1038/nrm3383. [PubMed: 22722607]

[5]. Luo C, Chen S, Xu N, Wang C, Sai W bo, Zhao W, et al. Glycoengineering of pertuzumab and its impact on the pharmacokinetic/pharmacodynamic properties. Sci Rep. 2017;7:46347. doi:10.1038/srep46347. [PubMed: 28397880]

[6]. Schwarz DS, Blower MD. The endoplasmic reticulum: structure, function and response to cellular signaling. Cell Mol Life Sci. 2016;73:79–94. doi:10.1007/s00018-015-2052-6. [PubMed: 26433683]

[7]. Bankaitis VA, Garcia-Mata R, Mousley CJ. Golgi Membrane Dynamics and Lipid Metabolism. Curr Biol. 2012;22:R414–24. doi:10.1016/j.cub.2012.03.004. [PubMed: 22625862]

[8]. Hossler P Protein Glycosylation Control in Mammalian Cell Culture: Past Precedents and Contemporary Prospects In: Hu WS, Zeng A-P, editors. Genomics and Systems Biology of Mammalian Cell Culture. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012 p. 187–219. doi:10.1007/10_2011_113.

[9]. Umaña P, Bailey JE. A mathematical model of N-linked glycoform biosynthesis. Biotechnol Bioeng. 1997;55:890–908. [PubMed: 18636599]

[10]. Krambeck FJ, Betenbaugh MJ. A mathematical model of N-linked glycosylation. Biotechnol Bioeng. 2005;92:711–28. doi:10.1002/bit.20645. [PubMed: 16247773]

[11]. Liu G, Marathe DD, Matta KL, Neelamegham S. Systems-level modeling of cellular glycosylation reaction networks: O-linked glycan formation on natural selectin ligands. Bioinformatics. 2008;24:2740–7. doi:10.1093/bioinformatics/btn515. [PubMed: 18842604]

[12]. Liu G, Neelamegham S. A Computational Framework for the Automated Construction of Glycosylation Reaction Networks. PLOS ONE. 2014;9:e100939. doi:10.1371/journal.pone.0100939. [PubMed: 24978019]

[13]. McDonald AG, Tipton KF, Davey GP. A Knowledge-Based System for Display and Prediction of O-Glycosylation Network Behaviour in Response to Enzyme Knockouts. PLOS Comput Biol. 2016;12:e1004844. doi:10.1371/journal.pcbi.1004844. [PubMed: 27054587]

[14]. Kremkow BG, Lee KH. Glyco-Mapper: A Chinese hamster ovary (CHO) genome-specific glycosylation prediction tool. Metab Eng. 2018;47:134–42. doi:10.1016/j.ymben.2018.03.002. [PubMed: 29522825]

[15]. Spahn PN, Hansen AH, Hansen HG, Arnsdorf J, Kildegaard HF, Lewis NE. A Markov chain model for N-linked protein glycosylation – towards a low-parameter tool for model-driven glycoengineering. Metab Eng. 2016;33:52–66. doi:10.1016/j.ymben.2015.10.007. [PubMed: 26537759]

[16]. Spahn PN, Hansen AH, Kol S, Voldborg BG, Lewis NE. Predictive glycoengineering of biosimilars using a Markov chain glycosylation model. Biotechnol J. 2017;12. doi:10.1002/biot.201600489.

[17]. Bydlinski N, Maresch D, Schmieder V, Klanert G, Strasser R, Borth N. The contributions of individual galactosyltransferases to protein specific N-glycan processing in Chinese Hamster Ovary cells. J Biotechnol. 2018;282:101–10. doi:10.1016/j.jbiotec.2018.07.015. [PubMed: 30017654]

[18]. Hassinen A, Khoder-Agha F, Khosrowabadi E, Mennerich D, Harrus D, Noel M, et al. A Golgi-associated redox switch regulates catalytic activation and cooperative functioning of ST6Gal-I with B4GalT-I. Redox Biol. 2019;24:101182. [PubMed: 30959459]

[19]. Ujita M, Misra AK, McAuliffe J, Hindsgaul O, Fukuda M. Poly-N-acetyllactosamine Extension inN-Glycans and Core 2- and Core 4-branchedO-Glycans Is Differentially Controlled by i-Extension Enzyme and Different Members of the β1,4-Galactosyltransferase Gene Family. J Biol Chem. 2000;275:15868–75. doi:10.1074/jbc.M001034200. [PubMed: 10747980]

[20]. Mkhikian H, Mortales C-L, Zhou RW, Khachikyan K, Wu G, Haslam SM, et al. Golgi self-correction generates bioequivalent glycans to preserve cellular homeostasis. eLife. 2016;5. doi:10.7554/eLife.14814.

[21]. El-Battari A, Prorok M, Angata K, Mathieu S, Zerfaoui M, Ong E, et al. Different glycosyltransferases are differentially processed for secretion, dimerization, and autoglycosylation. Glycobiology. 2003;13:941–53. [PubMed: 14514709]

[22]. Rohfritsch PF, Joosten JAF, Krzewinski-Recchi M-A, Harduin-Lepers A, Laporte B, Juliant S, et al. Probing the substrate specificity of four different sialyltransferases using synthetic β-d-Galp-(1→4)-β-d-GlcpNAc-(1→2)-α-d-Manp-(1→O) (CH2)7CH3 analogues: General activating effect of replacing N-acetylglucosamine by N-propionylglucosamine. Biochim Biophys Acta BBA - Gen Subj. 2006;1760:685–92. doi:10.1016/j.bbagen.2005.12.012.

[23]. Mondal N, Buffone A, Stolfa G, Antonopoulos A, Lau JTY, Haslam SM, et al. ST3Gal-4 is the primary sialyltransferase regulating the synthesis of E-, P-, and L-selectin ligands on human myeloid leukocytes. Blood. 2015;125:687–96. [PubMed: 25498912]

[24]. Yang Z, Wang S, Halim A, Schulz MA, Frodin M, Rahman SH, et al. Engineered CHO cells for production of diverse, homogeneous glycoproteins. Nat Biotechnol. 2015;33:842–4. [PubMed: 26192319]

[25]. Krambeck FJ, Bennun SV, Andersen MR, Betenbaugh MJ. Model-based analysis of N-glycosylation in Chinese hamster ovary cells. PLOS ONE. 2017;12:e0175376. doi:10.1371/journal.pone.0175376. [PubMed: 28486471]

[26]. Chuang G-Y, Boyington JC, Joyce MG, Zhu J, Nabel GJ, Kwong PD, et al. Computational prediction of N-linked glycosylation incorporating structural properties and patterns. Bioinformatics. 2012;28:2249–55. doi:10.1093/bioinformatics/bts426. [PubMed: 22782545]

[27]. Ito H, Kameyama A, Sato T, Sukegawa M, Ishida H-K, Narimatsu H. Strategy for the fine characterization of glycosyltransferase specificity using isotopomer assembly. Nat Methods. 2007;4:577–82. doi:10.1038/nmeth1050. [PubMed: 17529980]

[28]. Stanley P, Schachter H, Taniguchi N. N-Glycans. In: Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Bertozzi CR, et al., editors. Essentials of Glycobiology. 2nd edition Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2009 http://www.ncbi.nlm.nih.gov/books/NBK1917/. Accessed 29 Jun 2019.

[29]. Čaval T, Tian W, Yang Z, Clausen H, Heck AJR. Direct quality control of glycoengineered erythropoietin variants. Nat Commun. 2018;9. doi:10.1038/s41467-018-05536-3. [PubMed: 29339724]

[30]. Goh JSY, Liu Y, Chan KF, Wan C, Teo G, Zhang P, et al. Producing recombinant therapeutic glycoproteins with enhanced sialylation using CHO-gmt4 glycosylation mutant cells. Bioengineered. 2014;5:269–73. [PubMed: 24911584]

[31]. Gupta R, Matta KL, Neelamegham S. A systematic analysis of acceptor specificity and reaction kinetics of five human α(2,3)sialyltransferases: Product inhibition studies illustrates reaction mechanism for ST3Gal-I. Biochem Biophys Res Commun. 2016;469:606–12. doi:10.1016/j.bbrc.2015.11.130. [PubMed: 26692484]

[32]. Taniguchi T, Woodward AM, Magnelli P, McColgan NM, Lehoux S, Jacobo SMP, et al. N-Glycosylation affects the stability and barrier function of the MUC16 mucin. J Biol Chem. 2017;292:11079–90. [PubMed: 28487369]

[33]. Nielsen MI, Stegmayr J, Grant OC, Yang Z, Nilsson UJ, Boos I, et al. Galectin binding to cells and glycoproteins with genetically modified glycosylation reveals galectin-glycan specificities in a natural context. J Biol Chem. 2018;293:20249–62. [PubMed: 30385505]

[34]. Lee PL, Kohler JJ, Pfeffer SR. Association of β−1,3-N-acetylglucosaminyltransferase 1 and β −1,4-galactosyltransferase 1, trans-Golgi enzymes involved in coupled poly-N-acetyllactosamine synthesis. Glycobiology. 2009;19:655–64. doi:10.1093/glycob/cwp035. [PubMed: 19261593]

[35]. Praissman JL, Live DH, Wang S, Ramiah A, Chinoy ZS, Boons G-J, et al. B4GAT1 is the priming enzyme for the LARGE-dependent functional glycosylation of α-dystroglycan. eLife. 2014;3:e03943. doi:10.7554/eLife.03943.

[36]. Kawano S, Hashimoto K, Miyama T, Goto S, Kanehisa M. Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. Bioinforma Oxf Engl. 2005;21:3976–82.

[37]. Puri A, Neelamegham S. Understanding Glycomechanics using Mathematical Modeling: A review of current approaches to simulate cellular glycosylation reaction networks. Ann Biomed Eng. 2012;40:816–27. doi:10.1007/s10439-011-0464-5. [PubMed: 22090146]

[38]. Lee J, Sundaram S, Shaper NL, Raju TS, Stanley P. Chinese Hamster Ovary (CHO) Cells May Express Six β4-Galactosyltransferases (β4GalTs) CONSEQUENCES OF THE LOSS OF FUNCTIONAL β4GalT-1, β4GalT-6, OR BOTH IN CHO GLYCOSYLATION MUTANTS. J Biol Chem. 2001;276:13924–34. doi:10.1074/jbc.M010046200. [PubMed: 11278604]

[39]. Chung C-Y, Yin B, Wang Q, Chuang K-Y, Chu JH, Betenbaugh MJ. Assessment of the coordinated role of ST3GAL3, ST3GAL4 and ST3GAL6 on the α2,3 sialylation linkage of mammalian glycoproteins. Biochem Biophys Res Commun. 2015;463:211–5. doi:10.1016/j.bbrc.2015.05.023. [PubMed: 25998389]

[40]. Chiang AWT, Li S, Spahn PN, Richelle A, Kuo C-C, Samoudi M, et al. Modulating carbohydrate-protein interactions through glycoengineering of monoclonal antibodies to impact cancer physiology. Curr Opin Struct Biol. 2016;40:104–11. doi:10.1016/j.sbi.2016.08.008. [PubMed: 27639240]

[41]. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat Protoc. 2019;14:639. doi:10.1038/s41596-018-0098-2. [PubMed: 30787451]

[42]. Conn A, Gould N, Toint P. A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds. SIAM J Numer Anal. 1991;28:545–72. doi:10.1137/0728030.

[43]. Lewis R, Shepherd A, Torczon V. Implementing Generating Set Search Methods for Linearly Constrained Minimization. SIAM J Sci Comput. 2007;29:2507–30. doi:10.1137/050635432.

[44]. García-Ródenas R, Linares LJ, López-Gómez JA. A Memetic Chaotic Gravitational Search Algorithm for unconstrained global optimization problems. Appl Soft Comput. 2019;79:14–29. doi:10.1016/j.asoc.2019.03.011.

[45]. Kolda TG, Lewis RM, Torczon V. A generating set direct search augmented Lagrangian algorithm for optimization with a combination of general and linear constraints. 2006.

[46]. Deb K, Srivastava S. A genetic algorithm based augmented Lagrangian method for constrained optimization. Comput Optim Appl. 2012;53:869–902. doi:10.1007/s10589-012-9468-9.

[47]. Megchelenbrink W, Huynen M, Marchiori E. optGpSampler: An Improved Tool for Uniformly Sampling the Solution-Space of Genome-Scale Metabolic Networks. PLOS ONE. 2014;9:e86587. doi:10.1371/journal.pone.0086587. [PubMed: 24551039]

[48]. Hou W, Qiu Y, Hashimoto N, Ching W-K, Aoki-Kinoshita KF. A systematic framework to derive N-glycan biosynthesis process and the automated construction of glycosylation networks. BMC Bioinformatics. 2016;17:240. doi:10.1186/s12859-016-1094-6. [PubMed: 27454116]

[49]. Krambeck FJ, Bennun SV, Narang S, Choi S, Yarema KJ, Betenbaugh MJ. A mathematical model to derive N-glycan structures and cellular enzyme activities from mass spectrometric data. Glycobiology. 2009;19:1163–75. doi:10.1093/glycob/cwp081. [PubMed: 19506293]

[50]. Ardèvol A, Rovira C. Reaction Mechanisms in Carbohydrate-Active Enzymes: Glycoside Hydrolases and Glycosyltransferases. Insights from ab Initio Quantum Mechanics/Molecular Mechanics Dynamic Simulations. J Am Chem Soc. 2015;137:7528–47. doi:10.1021/jacs.5b01156. [PubMed: 25970019]

[51]. Yang Q, Zhang R, Cai H, Wang L-. Revisiting the substrate specificity of mammalian α1,6-fucosyltransferase reveals that it catalyzes core fucosylation of N-glycans lacking α1,3-arm GlcNAc. J Biol Chem. 2017;292:14796–803. [PubMed: 28729420]

[52]. Castilho A, Gruber C, Thader A, Oostenbrink C, Pechlaner M, Steinkellner H, et al. Processing of complex N-glycans in IgG Fc-region is affected by core fucosylation. mAbs. 2015;7:863–70. doi:10.1080/19420862.2015.1053683. [PubMed: 26067753]

[53]. Amann T, Hansen AH, Kol S, Hansen HG, Arnsdorf J, Nallapareddy S, et al. Glyco-engineered CHO cell lines producing alpha-1-antitrypsin and C1 esterase inhibitor with fully humanized N-glycosylation profiles. Metab Eng. 2019;52:143–52. [PubMed: 30513349]

[54]. Kruschke JK. Bayesian estimation supersedes the t test. J Exp Psychol Gen. 2013;142:573–603. [PubMed: 22774788]

[55]. Kruschke JK, Liddell TM. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychon Bull Rev. 2018;25:178–206. doi:10.3758/s13423-016-1221-4. [PubMed: 28176294]

[56]. Winter N Matlab Toolbox for Bayesian Estimation. Contribute to NilsWinter/matlab-bayesian-estimation development by creating an account on GitHub. Matlab. 2019 https://github.com/NilsWinter/matlab-bayesian-estimation. Accessed 1 Apr 2019.

[57]. Depaoli S, Clifton JP, Cobb PR. Just Another Gibbs Sampler (JAGS): Flexible Software for MCMC Implementation. J Educ Behav Stat. 2016;41:628–49. doi:10.3102/1076998616664876.

[58]. Muller P, Parmigiani G, Rice K. FDR and Bayesian Multiple Comparisons Rules. Johns Hopkins Univ Dept Biostat Work Pap 2006 https://biostats.bepress.com/jhubiostat/paper115.

[59]. Beerli P Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics. 2006;22:341–5. doi:10.1093/bioinformatics/bti803. [PubMed: 16317072]

[60]. Guo H-B, Nairn A, Harris K, Randolph M, Alvarez-Manilla G, Moremen K, et al. Loss of expression of N-acetylglucosaminyltransferase Va results in altered gene expression of glycosyltransferases and galectins. FEBS Lett. 2008;582:527–35. doi:10.1016/j.febslet.2008.01.015. [PubMed: 18230362]

[61]. Jeong YT, Choi O, Lim HR, Son YD, Kim HJ, Kim JH. Enhanced sialylation of recombinant erythropoietin in CHO cells by human glycosyltransferase expression. J Microbiol Biotechnol. 2008;18:1945–52. [PubMed: 19131698]

[62]. Amann T, Schmieder V, Kildegaard HF, Borth N, Andersen MR. Genetic engineering approaches to improve posttranslational modification of biopharmaceuticals in different production platforms. Biotechnology and Bioengineering. 2019;116:2778–96. doi:10.1002/bit.27101. [PubMed: 31237682]

[63]. Rios LM, Sahinidis NV, 2013 Derivative-free optimization: a review of algorithms and comparison of software implementations. J Glob Optim 56, 1247–1293. 10.1007/s10898-012-9951-y

[64]. Lageveen-Kammeijer GSM, Haan N. de, Mohaupt P, Wagt S, Filius M, Nouta J, Falck D, Wuhrer M, 2019 Highly sensitive CE-ESI-MS analysis of N -glycans from complex biological samples. Nat Commun 10, 1–8. 10.1038/s41467-019-09910-7 [PubMed: 30602773]

[65]. Riley NM, Hebert AS, Westphall MS, Coon JJ, 2019 Capturing site-specific heterogeneity with large-scale N-glycoproteome analysis. Nat Commun 10, 1–13. 10.1038/s41467-019-09222-w [PubMed: 30602773]

[66]. Blanchard V, Liu X, Eigel S, Kaup M, Rieck S, Janciauskiene S, Sandig V, Marx U, Walden P, Tauber R, Berger M, 2011 N-glycosylation and biological activity of recombinant human alpha1-antitrypsin expressed in a novel human neuronal cell line. Biotechnol. Bioeng 108, 2118–2128. 10.1002/bit.23158 [PubMed: 21495009]

[67]. Montacir O, Montacir H, Springer A, Hinderlich S, Mahboudi F, Saadati A, Parr MK, 2018 Physicochemical Characterization, Glycosylation Pattern and Biosimilarity Assessment of the

Fusion Protein Etanercept. Protein J. 37, 164–179. 10.1007/s10930-018-9757-y [PubMed: 29411222]

[68]. Pierpont TM, Limper CB, Richards KL, 2018 Past, Present, and Future of Rituximab-The World's First Oncology Monoclonal Antibody Therapy. Front Oncol 8, 163 10.3389/fonc.2018.00163 [PubMed: 29915719]

[69]. Yang Q, An Y, Zhu S, Zhang R, Loke CM, Cipollo JF, Wang L-X, 2017 Glycan Remodeling of Human Erythropoietin (EPO) Through Combined Mammalian Cell Engineering and Chemoenzymatic Transglycosylation. ACS Chem Biol 12, 1665–1673. 10.1021/acschembio.7b00282 [PubMed: 28452462]

[70]. Salciccioli JD, Crutain Y, Komorowski M, Marshall DC. Sensitivity Analysis and Model Validation. In: MIT Critical Data, editor. Secondary Analysis of Electronic Health Records. Cham: Springer International Publishing; 2016 p. 263–71. doi:10.1007/978-3-319-43742-2_17.

[71]. Darling RJ, Kuchibhotla U, Glaesner W, Micanovic R, Witcher DR, Beals JM, 2002 Glycosylation of erythropoietin affects receptor binding kinetics: role of electrostatic interactions. Biochemistry 41, 14524–14531. 10.1021/bi0265022 [PubMed: 12463751]

[72]. Singh A. Thakur N. Sharma A. A review of supervised machine learning algorithms. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom); 2016. 1310–5. 1.

[73]. Solá RJ, Griebenow K, 2010 Glycosylation of Therapeutic Proteins: An Effective Strategy to Optimize Efficacy. BioDrugs 24, 9–21. [PubMed: 20055529]

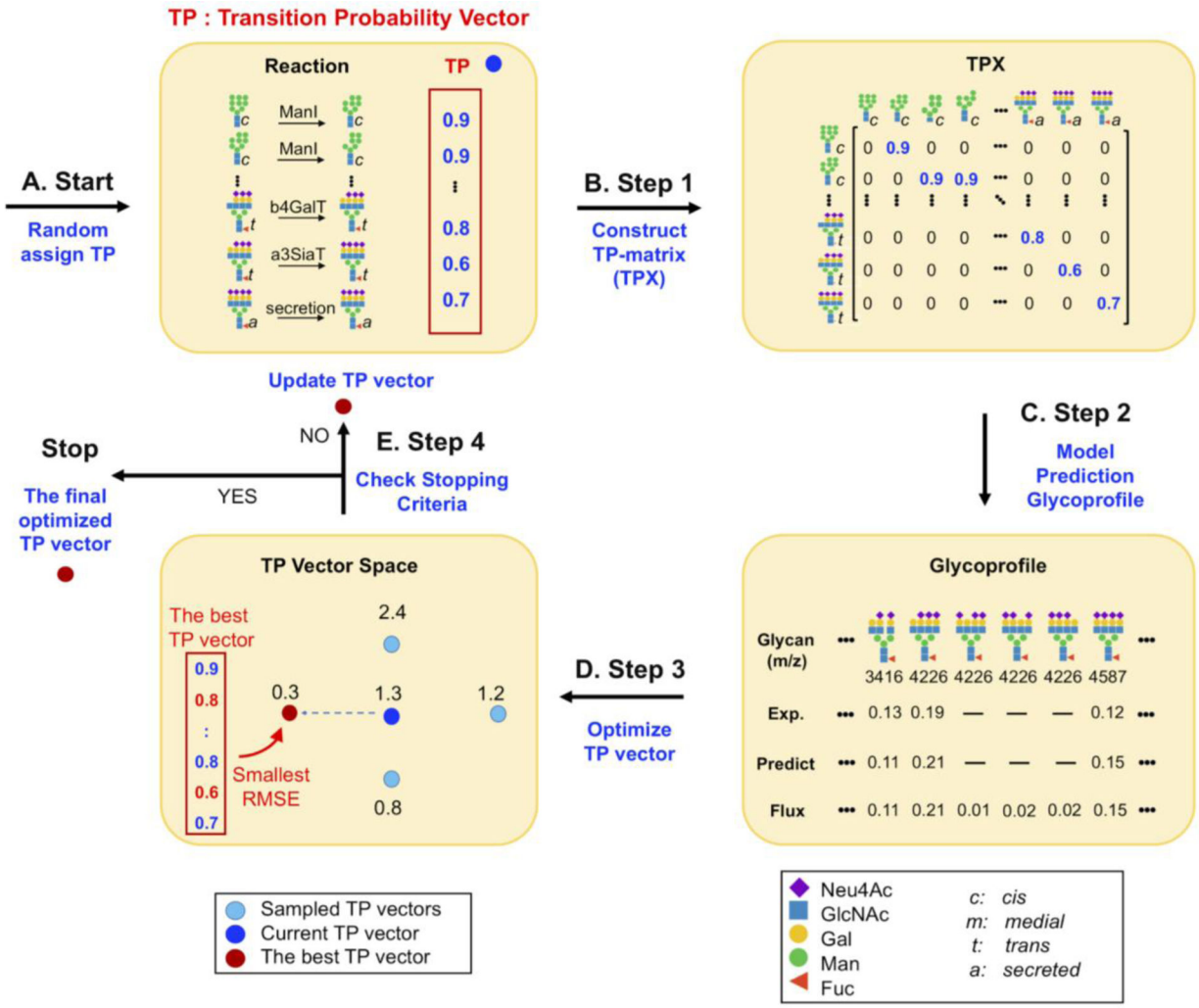

Figure 1. Glycoprofiles are fit to the Markov model using global optimization with the Pattern Search algorithm.

(**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 (see Appendix A). 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 $Man_9GlcNAc_2$), 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.

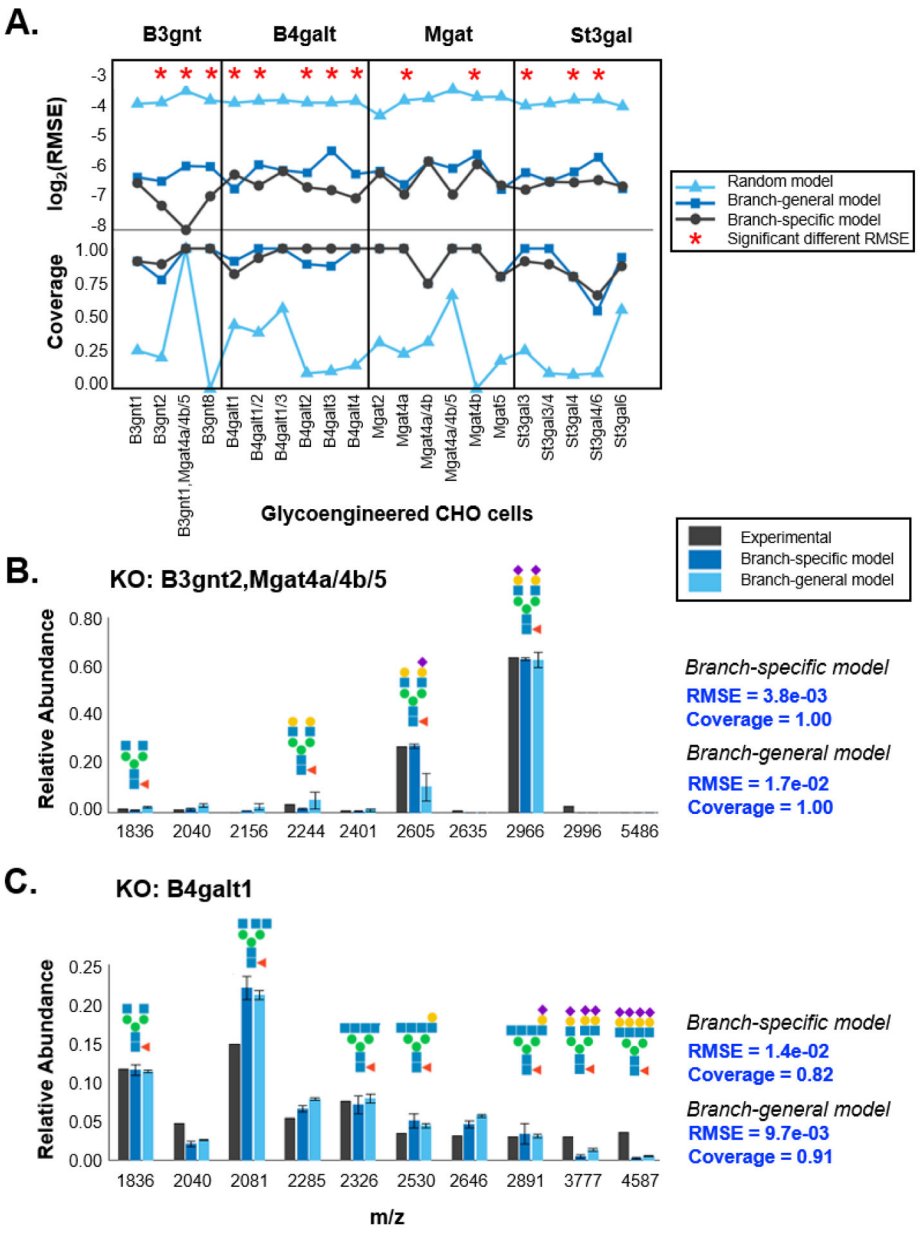

Figure 2. Prediction performance of the N-glycosylation Markov model.

(A) The RMSE and coverage were quantified for the model predictions of EPO produced in glycoengineered CHO cells. We tested three different categories of models: the branch-specific models, the branch-general models, and the random models (i.e., the branch-specific models assigned with random TP vectors). A red star indicates a significant difference of RMSE between the branch-specific and the branch-general models (highest density interval (HDI) = 95%). RMSEs of all models (branch-specific models and branch-general models) were significantly different from those of the random models. **(B)** The glycoprofile of EPO from the CHO cell line wherein B3gnt2 and Mgat4a/4b/5 were all knocked out has the greatest improvement in prediction (RMSE decreased) after introducing branch specificity reactions. **(C)** The glycoprofile from the B4galt1 knockout showed the least improvement in

prediction (RMSE slightly increased and coverage slightly decreased) after introducing branch-specific reactions. The error bars were calculated as the standard deviations of the glycan intensities produced by multiple optimized models. Note that the originally annotated glycoform at m/z = 3777 was potentially a misannotation and corrected in this figure. Please refer to Appendix G for the original annotated glycoprofile.

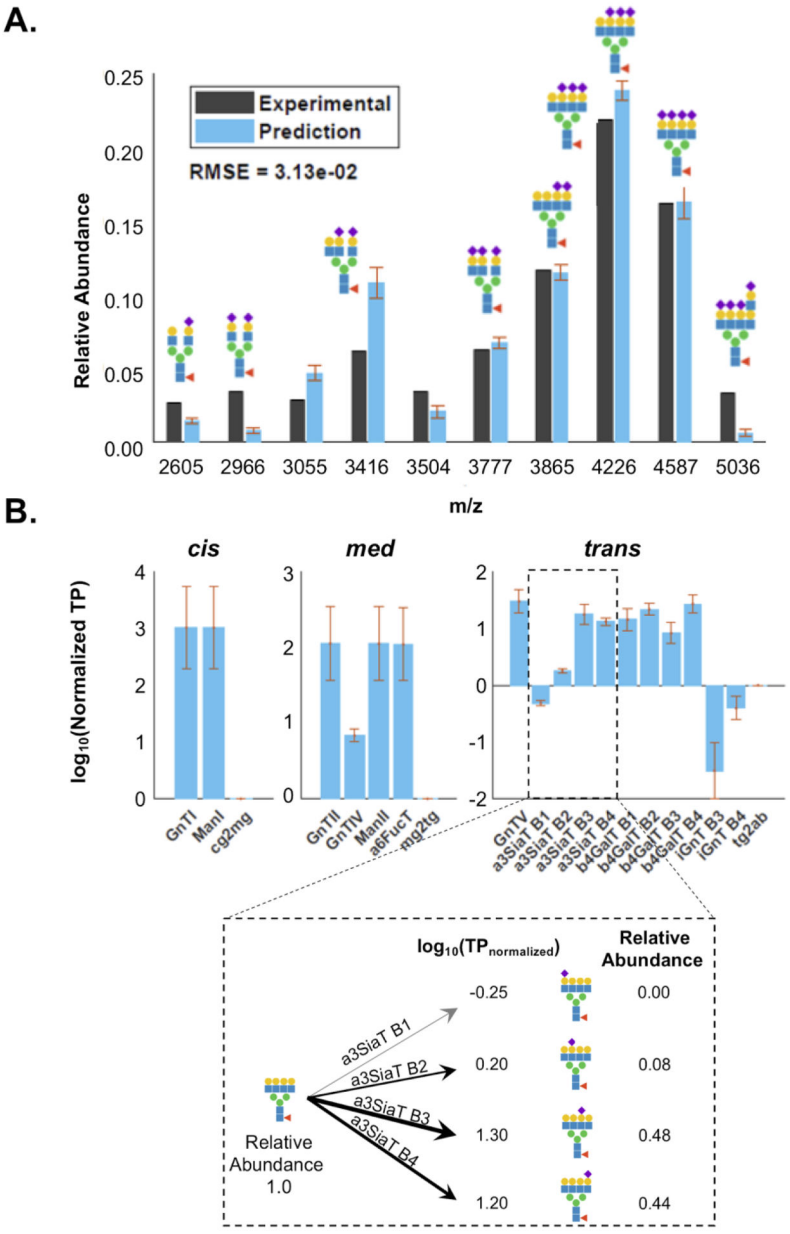

Figure 3. The branch-specific model effectively predicted glycoprofile and optimized transition probabilities for EPO produced in the wild-type CHO cell line.

(A) The model-predicted and experimental glycoprofiles of EPO produced in wild type CHO cells. Note that the top ten glycans presented here account for >85% of the total detected glycan abundance in the experimental glycoprofile. (**B)** The optimized transition probabilities (TPs) by reaction types, in which the TPs were normalized by transport TPs to the next compartments. For example, TPs of reactions localized in *cis* (Golgi apparatus) were normalized by the TP of glycan transport from *cis* to *medial*. The reaction types were separated into three subplots by compartments: *cis* (*cis to medial* transport, cg2mg), *medial* (*medial* to *trans* transport, mg2tg), and *trans* (secretion, tg2ab). The error bars were calculated as the standard deviations of the glycan intensities produced by multiple optimized models.

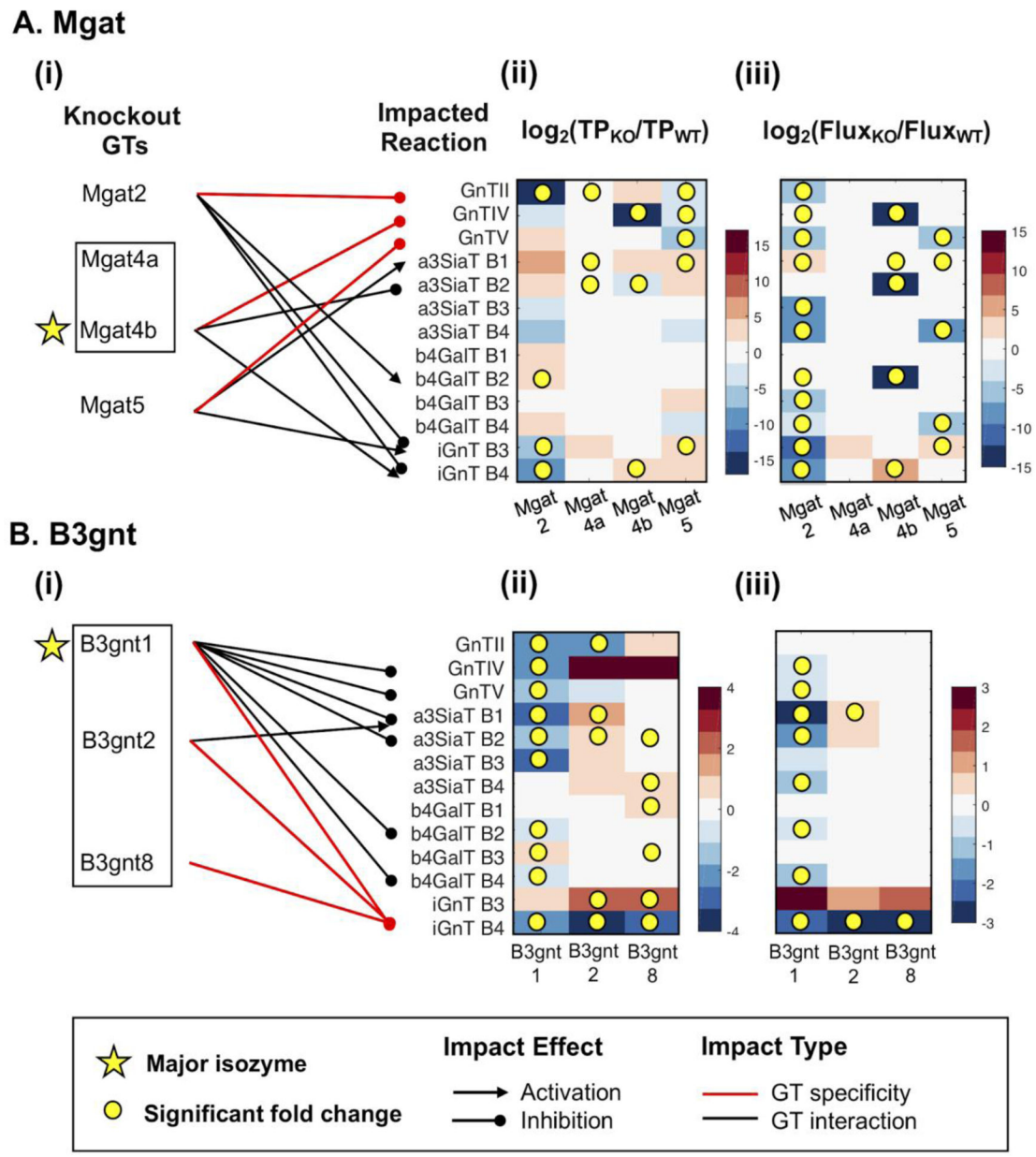

Figure 4. The branch-specific model identifies indirect effects one GT may have on the activity of other GTs.

(A) The Mgat family GTs. **(B)** The B3gnt family GTs. In each enzyme family, there are three subplots. **(i)** When a GT isozyme is knocked out (left), we detect changes in the flux of reactions (right), where expected reaction changes are in red, and genetic interactions with other GTs are shown in black. **(ii)** The heatmap shows the $\log_2$ fold change transition probability (TP) in for the GT knockout models, compared to the WT models. **(iii)** The heatmap shows the $\log_2$ fold change in flux for the GT knockout models in comparison with the WT models. The yellow dots indicate significant non-zero fold changes of the corresponding TP (HDI = 95%). The yellow stars indicate the major isozymes whose knockouts most significantly and severely impacted their enzymatically catalyzed reactions. The color for the solid line represents the type of reaction impacted by the GT knockout:

‘red’ (GT specific impact) and ‘black’ (GT-GT interaction). The terminal symbol for a line represents the interaction type impacted by the GT knockout: ‘arrow’ denotes activation and ‘filled circle’ is inhibition.

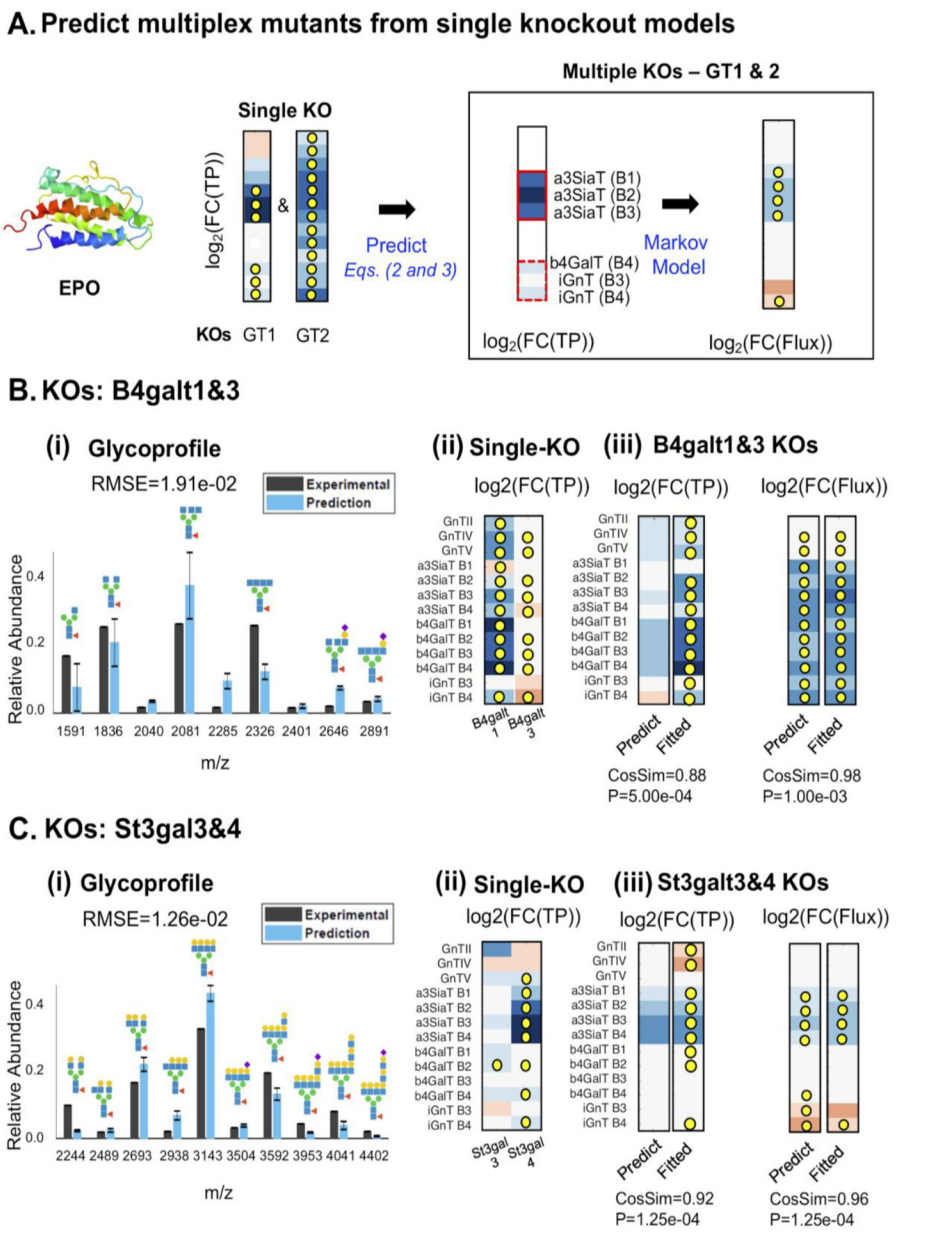

Figure 5. Models predict glycoprofiles of multi-GT knockouts.

(A) The multiple GT knockout models were built by combining TP vectors from single GT knockout models. We simulated the glycoprofiles for several multi-GT KOs involving the B4galt- and St3gal- family GTs: **(B)** B4galt1/3 and **(C)** St3gal3/4 (see Supplementary Figure E1 and E2 in Appendix E for more multi-GT KOs). The relative intensity (m/z) of glycans shown in each barplot correspond to the most abundant 7–10 glycans detected in the corresponding experimental glycoprofiles. For each profile, these m/z values collectively capture >85% of the total experimental signal intensity. Three different heatmaps (from left to right) show the fold change (FC) for TP values: the FC(TP) for single GT KOs, the FC(TP) for multiple GT KOs, and the FC(Flux) for multiple GT KOs. In the multiple GT KOs (FC(TP) and FC(Flux)). The yellow dots indicate significant non-zero fold changes of

the corresponding TP or Flux (HDI = 95%). Note, EPO is a Human erythropoietin NMR structure from PDB database (PDB ID code: 1buy). Cosine similarity (CosSim) is a nonparametric method used to measure the similarity of the two vectors (predicted and fitted). Specifically, it measures the cosine of the angle between two vectors, and the smaller angle means the higher similarity. The error bars were calculated as the standard deviations of the glycan intensities produced by multiple predicting models.

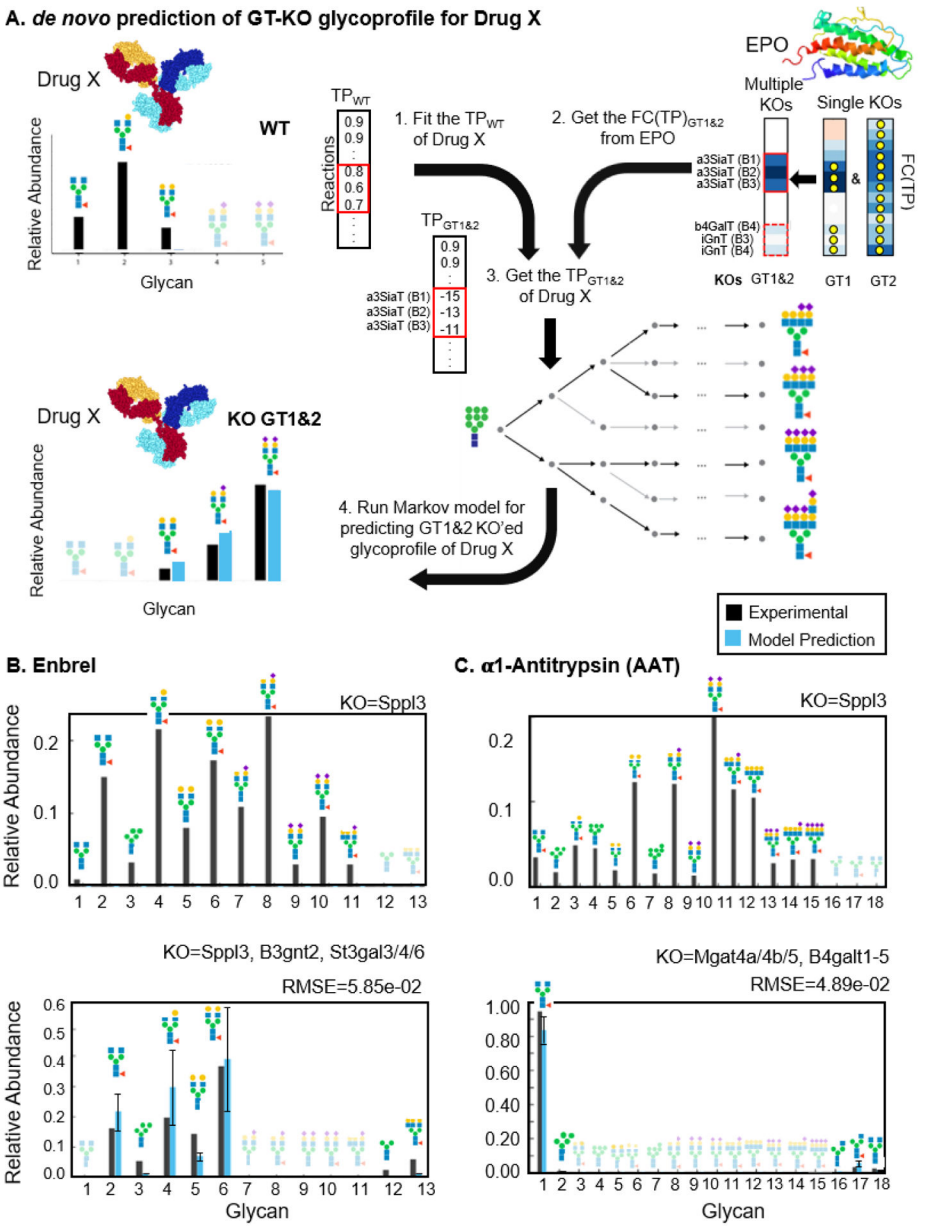

Figure 6. Multiple GT knockout glycoprofiles can be predicted *de novo* for diverse drugs.

(**A**) We established a workflow for *de novo* model prediction of glycoengineered glycoprofile for drugs, wherein TPs learned from glycoengineered EPO (Figure 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.

Table 1.

Branch-specific reaction rules for the N-linked glycosylation model

Reaction[*c]	Substrate[*a]	Product[*a]	Constraint[*a,b]	Localization
ManI	(Ma2Ma	Ma	-	cis
ManII	(Ma3(Ma6)Ma6	(Ma6Ma6	(GNb2|Ma3	medial
ManII	(Ma6)Ma6	(Ma6	(GNb2|Ma3	medial
GnTI	(Ma3(Ma3(Ma6)Ma6)Ma4	(GNb2Ma3(Ma3(Ma6)Ma6)Ma4	-	cis
GnTII	(GNb2|Ma3(Ma6)Mb4	(GNb2|Ma3(GNb2Ma6)Mb4	-	medial
GnTIV	(GNb2Ma3	(GNb2(GNb4)Ma3	-	medial
GnTV	(GNb2Ma6	(GNb2(GNb6)Ma6	-	trans
a6FuT	GNb4GN	(GNb4(Fa6)GN	~*…Ma2	medial
b4GalT (B1)	(GN	(Ab4GN	*…GNb2|Ma3	trans
b4GalT (B2)	(GN	(Ab4GN	*…GNb4|Ma3	trans
b4GalT (B3)	(GN	(Ab4GN	*…GNb2|Ma6	trans
b4GalT (B4)	(GN	(Ab4GN	*…GNb6|Ma6	trans
a3SiaT (B1)	(Ab4GN	(NNa3Ab4GN	*…GNb2|Ma3	trans
a3SiaT (B2)	(Ab4GN	(NNa3Ab4GN	*…GNb4|Ma3	trans
a3SiaT (B3)	(Ab4GN	(NNa3Ab4GN	*…GNb2|Ma6	trans
a3SiaT (B4)	(Ab4GN	(NNa3Ab4GN	*…GNb6|Ma6	trans
iGnT (B3)	(Ab4GN	(GNb3Ab4GN	*…GNb2|Ma6	trans
iGnT (B4)	(Ab4GN	(GNb3Ab4GN	*…GNb6|Ma6	trans

*a ‘A’, ‘F’, ‘GN’, ‘M’, and ‘NN’ represent galactose, fucose, GlcNAc, mannose, and NAcNAc respectively, whereas ‘aX’ or ‘bX’ (where ‘X’ is a number) represents an alpha or beta glycosidic bond connecting the two adjacent sugars (e.g. a3 represents alpha 1,3 glycosidic bond).

b ‘’ indicates the position of added moiety and associated bond strings, ‘…’ a string of any length with all brackets matched, and ‘|’ a branching point.

*c ‘B1–4’ indicate the four possible branches of a glycan as described by the Constraint

Note that, we specified the glycans as linear code strings with complete linkage and composition information for easy computation in the model.

Table 2.

Reactions potentially influenced by the knockout of a given enzyme.

Gene Knockout	Direct Reactions*
Mgat2	GnTII
Mgat4a/4b	GnTIV
Mgat5	GnTV
St3gal3/4/6	a3SiaT (Branch 1–4)
B4galt1/2/3/4	b4GalT (Branch 1–4)
B3gnt1/2/8	iGnT (Branch 3/4)

*Direct reaction included reactions directly catalyzed by the given enzyme encoded by the knocked-out gene(s), whereas the potentially impacted (dependent) reactions are those whose TPs may be influenced by the knockout of a given enzyme.