Article:
West, Ben, Wood, A Jamie and Ungar, Daniel (2021) Computational Modeling of Glycan Processing in the Golgi for Investigating Changes in the Arrangements of Biosynthetic Enzymes. Methods in Molecular Biology. pp. 209-222. ISSN: 1064-3745

https://doi.org/10.1007/978-1-0716-1685-7_10

Computational modelling of glycan processing in the Golgi for investigating changes in the arrangements of biosynthetic enzymes

Ben West[1], A. Jamie Wood[1,2], Daniel Ungar[1]

Departments of [1]Biology and [2]Mathematics, University of York, York, UK

Abstract

Modelling glycan biosynthesis is becoming increasingly important due to the far-reaching implications that glycosylation can exhibit, from pathologies to biopharmaceutical manufacturing. Here we describe a stochastic simulation approach, to overcome the deterministic nature of previous models, that aims to simulate the action of glycan modifying enzymes to produce a glycan profile. This is then coupled with an approximate Bayesian computation methodology to systematically fit to empirical data in order to determine which set of parameters adequately describes the organisation of enzymes within the Golgi. The model is described in detail along with a proof of concept and therapeutic applications.

Key words: glycosylation, stochastic simulation, approximate Bayesian computation, modelling

1.0 Introduction

Protein glycosylation is a complex and flexible post-translational modification that has been associated with a diverse set of biological processes and pathologies *(1–5)*. The high level of complexity arises from the vast variation in the structures of glycans that can be produced. Glycan structures are altered by two enzyme families: glycosidases and glycosyltransferases. Glycosidases hydrolyse glycosidic bonds, cleaving part of a glycan, whereas glycosyltransferases catalyse the formation of a glycosidic bond, thereby initiating, extending, and branching glycans.

Glycans are polymers consisting of several different monosaccharide units that can be added to each other in different orders and into different positions. A large set of competing enzymes, of the aforementioned glycosidase and glycosyltransferase families, are used to generate the polymers in the absence of a template. In comparison to a polysaccharide polymer, such as cellulose, where a more limited number of enzymes act in a concerted manner, the competing reactions when making a glycan result in a highly heterogenous mix of glycans. Yet this heterogeneity is never completely random. For example, different cell types in the human body show distinct and reproducible glycan profiles. The glycan profiles are, in part, influenced by cell-line specific expression of the biosynthetic enzymes, however, this is not sufficient to explain the differences arising in the profiles of different cell types *(6)*. Hence, the importance of sub-compartmentalisation of enzymes across Golgi cisternae has been highlighted as a key feature in glycan synthesis *(7)*. Understanding how non-uniform enzyme distributions across the Golgi *(8)* are maintained and in what way these control reproducible but distinct glycan profiles across cell types has become a key area of investigation for glycan biosynthesis.

Given the pervasiveness and far-reaching applications of glycosylation, a good understanding of how glycan heterogeneity is controlled is very important. Due to the complexities of this biosynthetic process, which involves a large number of competing enzymes working in concert, systems biology approaches are needed. Several

computational models describing the synthesis of glycans have been produced in an effort to understand the biosynthetic requirements for generating different glycan patterns *(9–12)*. Of particular interest for our method is the computational representation of the way different enzyme arrangements guide glycan processing.

To understand the implications of models assessing Golgi enzyme arrangements, it is first important to understand how this organelle works with regards to the glycosylation of proteins. The cisternae of the Golgi should be thought of as dynamic reaction chambers that contain the glycan modifying enzymes. The cisternae are arranged from the *cis* side of the Golgi to the *trans* side *(8)*, and glycoprotein substrates remain in the same cisterna throughout their residence in the organelle *(13)*. In order to maintain a sequence of glycan processing reactions, the enzymes must be retrogradely (i.e. in the opposite direction to secretion) trafficked in vesicles, so glycoproteins meet an ever changing subset of the enzymes *(14, 15)*. This process of Golgi trafficking is called the cisternal maturation model *(13, 16)* (Figure 1).

N-glycosylation is initiated on the cytoplasmic face of the endoplasmic reticulum (ER) with the creation of a glycan precursor consisting of two N-acetylglucosamine (GlcNAc) residues and five mannose residues ($Man_5GlcNAc_2$), which is flipped to the ER lumen. There, four more mannose residues are added as well as three glucoses, resulting in $Glc_3Man_9GlcNAc_2$ that is transferred *en bloc* to an asparagine residue of a newly synthesised protein. The three glucoses are used as part of a protein quality control step, and are removed sequentially, leaving $Man_9GlcNAc_2$ that can undergo further trimming prior to trafficking of the glycoprotein to the Golgi. However, some glycans with a single glucose can still enter the Golgi, meaning that a glycan enters the Golgi in one of three states: $Man_8GlcNAc_2$, $Man_9GlcNAc_2$ or $Glc_1Man_9GlcNAc_2$. Within the Golgi, alterations to a glycan result from the action of glycosidases and glycosyltransferases (Figure 2), until the glycan leaves the Golgi. This processing generates *N*-glycans belonging to one of three classes: oligomannose, hybrid, or complex. All *N*-glycans share a common core sequence: two GlcNAc residues

extended by three mannose residues ($Man_3GlcNAc_2$) onto which other monomer units are attached. Oligomannose glycans are those in which only mannose residues are attached onto the core and these are subject to mannose trimming enzymes such as mannosidase I and II. Complex glycans contain antennae initiated by GlcNAc residues added to the core. This occurs through the actions of N-acetylglucosamine transferases, such as MGAT1 – MGAT5. Hybrid glycans are characterised as containing both mannose and GlcNAc initiated antennae *(17)*.

Modelling the action, abundance, and localisation of these glycan modifying enzymes will help us to understand how the synthesis of glycans is controlled. Despite lagging behind other biological systems, due to inherent structural complexity, the computational modelling of glycosylation has gained traction in recent years. The first model of glycosylation, developed by Umana and Bailey, was used to generate 33 *N*-glycan reactions *in silico*, up to the point of galactosylation in the *N*-glycosylation pathway. Each Golgi cisterna was modelled as a reaction chamber that follows Michaelis-Menten kinetics with literature-derived parameters. By solving a set of ordinary differential equations (ODEs) the solution gave glycan structures that correctly simulated the experimental glycan profile typical for secreted recombinant proteins produced in Chinese hamster ovary (CHO) cells *(9)*. With the development of better technology this work has been greatly expanded on, going from 33 possible reactions to a possible 22,871 in a model by Krambeck and Betenbaugh, which not only includes galactosylation, but fucosylation and sialylation as well *(10)*. Another model utilising ODEs, which further extends on the previous modelling, utilised structure-specific turnover rates to provide a kinetic description of *N*-glycan processing along the entire secretory pathway *(18)*. Additionally, a study utilised the above described models to identify that changes in GalT activity unexpectedly affect branching during *N*-glycan processing *(19)*. A different aspect of modelling glycosylation is to model how altered culture conditions rather than altered enzyme arrangements/activities change glycan profiles. This has been extensively investigated for the influence of temperature *(20)*, culture feed *(21)*, and sugar

nucleotide donor abundance *(22)*. This work, also based on ODE methods, demonstrates the complexity of trying to model glycosylation whilst encapsulating different parameters that may exhibit an effect on glycosylation. Whilst these ODE models have refined our knowledge of glycosylation, they assume that the dynamics of glycosylation are captured thoroughly through a deterministic approach. Furthermore, ODE models cannot readily be used for fitting to experimental data. This is because the data needed to fit such models is not present at the appropriate level of detail to be sure that the models are constrained to the desired subspace of state space. Due to the low concentrations of enzymes and the high level of competition in the Golgi apparatus, stochastic models that incorporate biological noise are more appropriate when modelling glycosylation *(11, 12)*.

One such stochastic model by Spahn *et al.* does not rely on kinetic information, but rather uses methods from Monte-Carlo Markov chain (MCMC) theory. In this model each glycan is regarded as a state within a network that transitions to other states with certain reaction probabilities, independent of the past. This coupled with flux based analysis and a genetic algorithm approach for optimisation, was used to model glycosylation.

The stochastic model that is the focus of the remainder of this chapter utilises MCMC and the Gillespie algorithm to simulate biological noise in conjunction with an approximate Bayesian computation (ABC) fitting methodology *(12)*. This method allows us to link the organisation of Golgi enzymes to generated glycan profiles and thereby provides a tool for problems such as probing glycan engineering strategies, answering cell biological questions on intra-Golgi protein sorting, and pinpointing strategies for alleviating human diseases caused by defective glycan processing. A genetic algorithm is generally better at finding an accurate solution quickly, only if the solution is found, as the parameter perturbations used are random to counter the size of the biosynthetic flux system, which is too large for a systematic search. Each fit is independent of the last which leads to a loss of information regarding the trajectory of the fitting, making the found solutions less reliable, and preventing the direct assessment of relative shifts in the parameter space. In contrast, our use of

Bayesian computation, enabled by the more streamlined flux map, is a statistical approach to fit the parameters of the biosynthetic machinery to a state that produces the expected glycan pattern. This allows a systematic approach for parameter fitting, delivering high quality relative information on parameter shifts, thereby providing important cell biological information on the changes to the glycosylation machinery between the assessed cellular states.

The computational model of glycan biosynthesis was created using custom written Java code. The model aims to simulate the action of glycan modifying enzymes to produce a glycan profile that is then compared to an experimental glycan profile to determine which set of parameters adequately describes enzyme organisation in the Golgi. The modelling method is divided into two separate bodies of code: stochastic simulation and model fitting. Both will be explored in depth below. Broadly, the stochastic simulation is designed to create a glycan profile based largely on parameters termed the “effective” enzymatic rate (EFER). The EFER is an amalgamation of the enzyme’s amount, its turnover rate, and the sugar-nucleotide substrate concentration where appropriate. By subsuming these parameters under one value we decrease the parameter space, making the modelling computationally efficient. The EFER is used to describe the rate constant of a particular reaction experienced by a focal glycan. Using an ABC fitting algorithm that relies on some (often limited) prior knowledge of the parameters, randomly selected parameter values from a prior distribution are accepted or rejected based on similarity between simulated and experimentally obtained data. We are using data obtained using MALDI mass spectrometry of permethylated glycans, as this has been shown to provide reliable quantitative glycan profiles *(23)*. Results from the fitting process tell the researcher how a parameter set needs to change from a starting state to generate the altered glycan profile, providing insights into altered enzyme arrangements within the Golgi. Crucially this means that the key information is not the final parameter derived, but rather the changes needed to improve the fit.

2.0 Stochastic Simulation Algorithm (SSA)

Underpinning the stochasticity of the model is the Gillespie algorithm, which uses the EFER as a reaction probability per unit time *(24)*. By treating the actions of independent enzymes in each of the cisternae as probabilities, we can generate heterogeneity similar to that seen in experimental glycan profiles.

$Man_8GlcNAc_2$, $Glc_1Man_9GlcNAc_2$ and $Man_9GlcNAc_2$ are the three possible input glycans used as the starting point of processing. Which of these three gets used is determined probabilistically weighted by using two input parameters (Table 1 'E'), the $Man_8GlcNAc_2$ and $Glc_1Man_9GlcNAc_2$ fractions. Each enzymatic processing step that is chosen via the stochastic process then progressively alters the input glycan as it moves through the Golgi. The substrate and product of these enzymatic steps are both in a linear notation form, as are all glycans within this simulation. The linear notation form used here is just one possible example, but the type used was created to be tailored for the string substitutions used by the SSA while allowing all the necessary information from mass spectrometry to be encapsulated. Linear notation allows for the actions of the enzymes to be implemented using a string substitution method to build new glycans. In essence, the action of each enzyme is simulated by the code searching for the substrate sequence, and if the enzyme is chosen to act this substrate sequence will be substituted with the enzyme's product sequence. For example, MAN1 will look for the substrate sequence: '1Man2.1Man:' (Table 1 'B') and if the code finds it and the enzyme is chosen to act, it will replace '1Man2.1Man:' with the product sequence '1Man:' (Table 1 'C'), simulating the cleavage of a mannose residue. It is important to note that a single enzyme can have different substrates, which is why some enzymes have multiple entries in the table (Table 1 'A'). In some cases, different substrates are processed with a different rate; this is implemented using scale factors that alter the rate of the enzyme for a given substrate. These scale-factors initially have a value of one, meaning the enzyme's rate for both substrates is the same. If, however, after fitting the scale-factor deviates from this initial value of one, then altered substrate specificity of the enzyme is

considered to be playing a role in glycan synthesis. This type of information is an additional output of the model beyond enzyme organisation in the Golgi.

2.1 Glycan processing

A typical simulation run will use 10,000 input glycans stochastically processed one-by-one to generate a computed glycan profile. The input glycans are divided into three types based on the allocated proportions as determined by the $Man_8GlcNAc_2$ fraction and $Glc_1Man_9GlcNAc_2$ fraction. Then for each glycan the SSA will identify all possible substrate strings in the enzyme information (Table 1 'B'). The EFERs for each enzyme that can thus act on this glycan (Table 1 'D') are added up and this value becomes the Total Propensity. The Total Propensity is required for implementation of the Gillespie algorithm.

Using a pseudo-random number (we used a Mersenne Twister *(25)*) multiplied by the Total Propensity, an enzyme is chosen to act by randomly selecting from the EFERs of the reactions competing for the substrate in question. If there are multiple sites that the chosen enzyme can act upon, a similar process is iterated through to determine at which site the enzyme should act. A second pseudo random number is then used to randomly draw from an exponential distribution with a mean of $(\text{Total Propensity})^{-1}$, in order to simulate a time interval within which the reaction occurred. The randomness arising from the use of pseudo-random numbers is an essential component of the stochasticity required to mimic the competitive and heterogeneous nature of glycan biosynthesis.

2.2 String substitution to modify glycans *in silico*

The string substitution is performed in a two-part process, to ensure fidelity of the glycan string. First, the substrate sequence is replaced with a proxy string and subsequently that string is replaced with the intended product (Table 1 'C'). The enzyme "OM Quench" (Table 1) is an artificial enzyme that is used to quench the processing of oligomannose glycans (Figure 1). This is needed to mimic the action of glycans being transported retrogradely back to the endoplasmic reticulum or being phosphorylated for lysosomal targeting. Both of these

actions stop further glycan processing, and to achieve this, the string substitution adds a “P” tag to all monosaccharide residues in the chain, making this new string unrecognisable for all enzymes. A glycan is modified in an iterative process until the cumulative time interval used by the enzymatic reactions exceeds the transit time (Table 1 ‘E’). At this point the glycan moves onto the next cisterna, or out of the Golgi if it was in the final cisterna. Once a glycan has moved out of the Golgi all “P” tags present are removed. The simulation thus generates 10,000 stochastically modified glycans, which are then used to produce a simulated glycan profile by calculating the relative abundance of each of its glycan species.

2.3 Comparison with empirical data

The simulation uses an empirically determined glycan profile to fit the simulated profile. The empirical data contains a list of glycans in linear notation, and the relative abundance of each as determined by mass spectrometry, as well as the error associated with each measurement.

To compare simulated and empirical data, the empirical results have to first be aligned with the profile generated from the 10,000 simulated glycans so that a comparison of glycan abundance can be drawn. For this, the molecular weight of each glycan is calculated using the molecular weights of its monosaccharide building blocks. By calculating their molecular weights, we can merge different glycans that were computationally generated and have identical masses, to create a single virtual glycan species. This is necessary because such glycans are indistinguishable using simple MALDI mass spectrometry. Now the empirical and simulated relative abundances of each glycan of a given mass can be compared to calculate a penalty score based on the difference between empirical and simulated abundance. The sum of the individual penalty scores is the overall Score generated by the SSA.

There are a multitude of scoring methods that can be employed, and each has its own merits. For example, the square difference between the error and the absolute value of the

difference between empirical and simulated abundancies (Formula 1) provides a penalty score that places a greater weight onto the most abundant glycans in the profile. This will allow computation of the best global fit, to obtain more generalised information from the model.

Formula 1:

In contrast, using the coefficient of variation (Formula 2) as the score method, puts much more focus on the less abundant glycan species. These can often be of great functional interest, such as some low abundance sialylated or fucosylated glycans, and therefore may require special attention.

Formula 2:

These are just two possible scoring methods, and it will be up to the experimenter to consider which penalty score calculation best reflects the needs of the specific project.

3.0 Approximate Bayesian Computation for fitting the model to experimental data

The second body of code is used to adjust the parameter set to ensure the modelled glycan profile fits the empirically determined one; this is accomplished through the application of Bayesian statistics. The aim of Bayesian methods is to compute a posterior probability $P(A|B)$, for a set of uncertain parameters A, given experimentally observed data B. Bayes formula, where $P(A)$ is the prior probability of our beliefs about the system (e.g. abundance and activity of resident Golgi enzymes) before B is observed, is given below.

In our case, the beliefs about the system are the information gathered from literature, which are used to calculate the EFER. This information is highly uncertain in most cases. Therefore, rather than treating the EFERs as single values, they are defined as probability density functions (PDFs). Two types of PDFs are chosen to describe parameter distributions:

log normal and exponential decay curves. These PDFs represent simple distributions with support on the positive half place only. The mean for each curve is set to our best estimate for the corresponding EFER. A log normal curve, when sampled from, will probabilistically yield a value that tends away from zero and it is for this reason that a log normal distribution is used to describe enzymes that we believe are active within a cisterna. In contrast, an exponential decay curve will yield values much closer to zero compared to a log normal, as smaller values have a greater probability, this type of PDF is used for simulating enzymes that we do not believe are acting in a particular cisterna. Importantly though, an exponential decay PDF still allows the model to engage an enzyme in that cisterna, so if we are wrong about the absence of the enzyme, the model will correct our error of judgment. By working through Bayes formula, we generate a posterior probability distribution and by comparing our prior knowledge to the generated posterior knowledge we could infer changes within the Golgi. However, this method would rely on calculating a likelihood function $P(B|A)$, which represents the probability of observing the data B, given the parameter values A. But as is often the case, for the system modelled here the likelihood function is intractable. Therefore, an ABC method has been adapted *(26)*. Essentially, for each EFER values are sampled from its PDF. These values are fed into the SSA, and depending on the calculated Score, the parameter value sets are accepted or rejected.

3.1 ABC Prerequisites

The ABC code loads enzyme rules and the order of the enzymes into its register from the same *.xls* file that the SSA uses. In addition, files containing 1000 x and y coordinates, describing values of the PDF, are used for each parameter. These PDF containing files are ordered in the same way as the enzymes in the register. When creating the PDFs it is important to choose an appropriate x value increment (step--size) in order for the tail of the distribution reach a y value below machine error (10^{-8}, as anything below this is effectively zero to a machine), to avoid boundary effects. Furthermore, when creating a PDF with a log normal distribution a variance must be specified, which in most cases can be defined as the

square of the mean. However, in some cases, our prior beliefs about the system are more uncertain due to a lack of literature and in these cases the variance can be increased to sample from a wider distribution.

3.2 Fitting methodology

The PDF for each enzyme within each cisterna, as well as PDFs for the starting glycan fractions and transit time, are loaded in from the prior spreadsheet files and stored in an array. A pseudo-random number is generated and from this, the values from a PDF are sequentially subtracted until the random number becomes negative. The last value subtracted becomes the EFER for that specific parameter, and repeating this for all PDFs, a parameter value set is passed onto the stochastic simulation. The SSA returns a Score, which is compared to a set threshold predetermined by the user. If the Score is below the threshold it is accepted, and the used parameter value set is stored. This process is repeated until there are n accepted sets of Scores.

After n Scores are accepted, a file is generated containing the stored parameter value sets. This file can be used to produce the posterior PDFs for each parameter. By examining the shifts in distribution between prior and posterior PDFs we can begin to understand how the organisation of the Golgi had to change to generate the glycan profile differences between the cellular states used at the start and end of the fitting process.

3.3 Threshold variability

Within our code is a sub-routine that can help set the Score-acceptance threshold. It has been demonstrated that MCMC algorithms are most efficient when the acceptance rate is set at 7.001% *(27)*. This is achieved by the code through first randomly sampling prior values and accepting or rejecting based on an initial threshold. This initial threshold is determined as the lowest Score that could be achieved in a reasonable amount of computational time (typically 24 hours). If the acceptance rate is lower than 7%, the initial threshold is deemed too low, and the code will increase the threshold by 10%. In contrast, if greater than 7% of

the prior values are being accepted, the threshold is deemed too high, and is decreased by 10%. In this way, the Score-acceptance threshold can be brought to a number close to that deemed optimally efficient. The Score-acceptance threshold is continually changed within a run until a second more stringent user-defined threshold is reached. The algorithm will then sample in that region of the parameter space until 10,000 parameter values for each variable have been accepted. During a chain of fitting runs the stringent threshold is generally lowered whilst some of the parameters are shifted from to the posterior PDF of the previous run. Decisions on which parameters to shift have to be taken by users experienced in the cell biology of glycan biosynthesis to ensure that parameter fitting yields plausible cellular states. This type of user guided fitting is a crucial aspect of the ABC approach *(28)*.

4.0 Proof of concept

We are providing one example of validation here to help the reader understand how the modelling is used. For more examples the reader is referred to Fisher *et al* papers *(12, 29)*. To explore if the model has the ability to make rational predictions, initially the computed glycan profile was fitted to experimental profiles obtained from wild type HEK293T cell lines. Following this, HEK293T cells were treated with the MAN2 inhibitor swainsonine *(30)*, and the altered glycan profile determined. As expected, the resulting glycan profile shows a significant increase in hybrid glycans. By starting with the parameter values obtained from the wild type fit, we reasoned that any changes in parameters required to fit to the swainsonine treated glycan profile, reflected changes in the glycosylation machinery due to swainsonine treatment. After iterating through several rounds of simulation and fitting, the model predicted a large decrease in the EFER of MAN2 (Figure 3) *(12)*.

5.0 Further applications of the computational model

The implication of a model that can deduce the organisational changes of the glycosylation machinery using experimentally determined glycan profiles is far-reaching and important. The ability to determine the changes necessary to get from one cell-type to another, or more

importantly wild type to mutant cell is incredibly significant with regards to disease. For example, many congenital disorders of glycosylation (CDGs) show complex alterations in glycan biosynthesis *(31, 32)*. The model could potentially help the diagnostics of orphan CDGs by comparing the glycan profiles obtained from patients and healthy controls, which would highlight critical changes in the biosynthetic machinery. In addition, computational modelling could pinpoint critical changes in the glycosylation machinery that may correct the glycan profile in patients, and even perform *in silico* drug tests on healthy and patient cells to suggest novel therapeutic directions. Similar approaches could be used for other diseases, such as some cancers where altered glycosylation is known to influence the pathology *(5)*. For example, using the model it was found that decreased fucosylation flux in a Cog4KO cell line compared to wild type was the result of reduced MGAT1 activity by restricting the amount of complex glycans which were the ideal substrates for fucosylation *(12)*.

Another possible application of the model is within the biopharmaceutical industry. As previously mentioned, glycan heterogeneity can have a detrimental effect on drug efficacy. Therefore, being able to control glycosylation would be of enormous benefit both economically and from a health perspective. By understanding how the organisation of enzymes in the Golgi needs to change in order to get a specific glycan profile, it should be possible to plan glycoengineering strategies to control the glycosylation of biopharmaceuticals *(29)*.

6.0 References

1. Takahashi M, Tsuda T, Ikeda Y, et al (2003) Role of N-glycans in growth factor signaling. Glycoconj. J. 20:207–212

2. Gu J, Isaji T, Xu Q, et al (2012) Potential roles of N-glycosylation in cell adhesion. Glycoconj J 29:599–607

3. Xu C, Ng DTW (2015) Glycosylation-directed quality control of protein folding. Nat. Rev. Mol. Cell Biol. 16:742–752

4. Chang IJ, He M, Lam CT (2018) Congenital disorders of glycosylation. Ann Transl Med 6:477–477

5. Pinho SS, Reis CA (2015) Glycosylation in cancer: Mechanisms and clinical implications. Nat. Rev. Cancer 15:540–555

6. Nairn A V., Aoki K, Dela Rosa M, et al (2012) Regulation of glycan structures in murine embryonic stem cells: Combined transcript profiling of glycan-related genes and glycan structural analysis. J Biol Chem 287:37835–37856

7. Fisher P, Ungar D (2016) Bridging the gap between glycosylation and vesicle traffic. Front. Cell Dev. Biol. 4:15

8. Rabouille C, Hui N, Hunte F, et al (1995) Mapping the distribution of Golgi enzymes involved in the construction of complex oligosaccharides. J Cell Sci 108:1617–1627

9. Umaña P, Bailey JE (1997) A mathematical model of N-linked glycoform biosynthesis. Biotechnol Bioeng 55:890–908

10. Krambeck FJ, Betenbaugh MJ (2005) A mathematical model of N-linked glycosylation. Biotechnol Bioeng 92:711–728

11. Spahn PN, Hansen AH, Hansen HG, et al (2016) A Markov chain model for N-linked protein glycosylation - towards a low-parameter tool for model-driven

glycoengineering. Metab Eng 33:52–66

12. Fisher P, Spencer H, Thomas-Oates J, et al (2019) Modeling Glycan Processing Reveals Golgi-Enzyme Homeostasis upon Trafficking Defects and Cellular Differentiation. Cell Rep 27:1231–1243

13. Glick BS, Elston T, Oster G (1997) A cisternal maturation mechanism can explain the asymmetry of the Golgi stack. FEBS Lett 414:177–181

14. Ungar D, Oka T, Brittle EE, et al (2002) Characterization of a mammalian Golgi-localized protein complex, COG, that is required for normal Golgi morphology and function. J Cell Biol 157:405–415

15. Harris SL, Waters MG (1996) Localization of a yeast early Golgi mannosyltransferase, Och1p, involves retrograde transport. J Cell Biol 132:985–998

16. Morré DJ, Ovtracht L (1977) Dynamics of the Golgi apparatus: membrane differentiation and membrane flow. Int Rev Cytol Suppl 61–188

17. Ungar D (2009) Golgi linked protein glycosylation and associated diseases. Semin. Cell Dev. Biol. 20:762–769

18. Arigoni-Affolter I, Scibona E, Lin C-W, et al (2019) Mechanistic reconstruction of glycoprotein secretion through monitoring of intracellular N-glycan processing. Sci Adv 5:eaax8930

19. McDonald AG, Hayes JM, Bezak T, et al (2014) Galactosyltransferase 4 is a major control point for glycan branching in N-linked glycosylation. J Cell Sci 127:5014–5026

20. Sou SN, Jedrzejewski PM, Lee K, et al (2017) Model-based investigation of intracellular processes determining antibody Fc-glycosylation under mild hypothermia. Biotechnol Bioeng 114:1570–1582

21. Kotidis P, Jedrzejewski P, Sou SN, et al (2019) Model based optimization of antibody

galactosylation in CHO cell culture. Biotechnol Bioeng 116:1612–1626

22. Jimenez del Val I, Nagy JM, Kontoravdi C (2011) A dynamic mathematical model for monoclonal antibody N-linked glycosylation and nucleotide sugar donor transport within a maturing Golgi apparatus. Biotechnol Prog 27:1730–1743

23. Wada Y, Azadi P, Costello CE, et al (2007) Comparison of the methods for profiling glycoprotein glycans - HUPO human disease glycomics/proteome initiative multi-institutional study. Glycobiology 17:411–422

24. Gillespie DT (1976) A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. J Comput Phys 22:403–434

25. Matsumoto M, Nishimura T (1998) Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator. ACM Trans Model Comput Simul 8:3–30

26. Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. PNAS December 100:15324–15328

27. Sherlock C, Thiery AH, Roberts GO, Rosenthal JS (2015) On the efficiency of pseudo-marginal random walk metropolis algorithms. Ann Stat 43:238–275

28. Toni T, Stumpf MPH (2009) Simulation-based model selection for dynamical systems in systems and population biology. Bioinformatics 26:104–110

29. Fisher P, Thomas-Oates J, Wood AJ, Ungar D (2019) The N-Glycosylation Processing Potential of the Mammalian Golgi Apparatus. Front Cell Dev Biol 7:157

30. Elbein AD, Solf R, Dorling PR, Vosbeck K (1981) Swainsonine: An inhibitor of glycoprotein processing. Proc Natl Acad Sci U S A 78:7393–7397

31. Zeevaert R, Foulquier F, Jaeken J, Matthijs G (2008) Deficiencies in subunits of the Conserved Oligomeric Golgi (COG) complex define a novel group of Congenital

Disorders of Glycosylation. Mol Genet Metab 93:15–21

32. Ng BG, Freeze HH (2018) Perspectives on Glycosylation and Its Congenital Disorders. Trends Genet. 34:466–476

Figure Captions

Figure 1. An overview of the cisternal maturation model. Secreted cargo remains within the cisternae, while resident Golgi proteins are transported in a retrograde fashion to previous cisternae to induce their maturation.

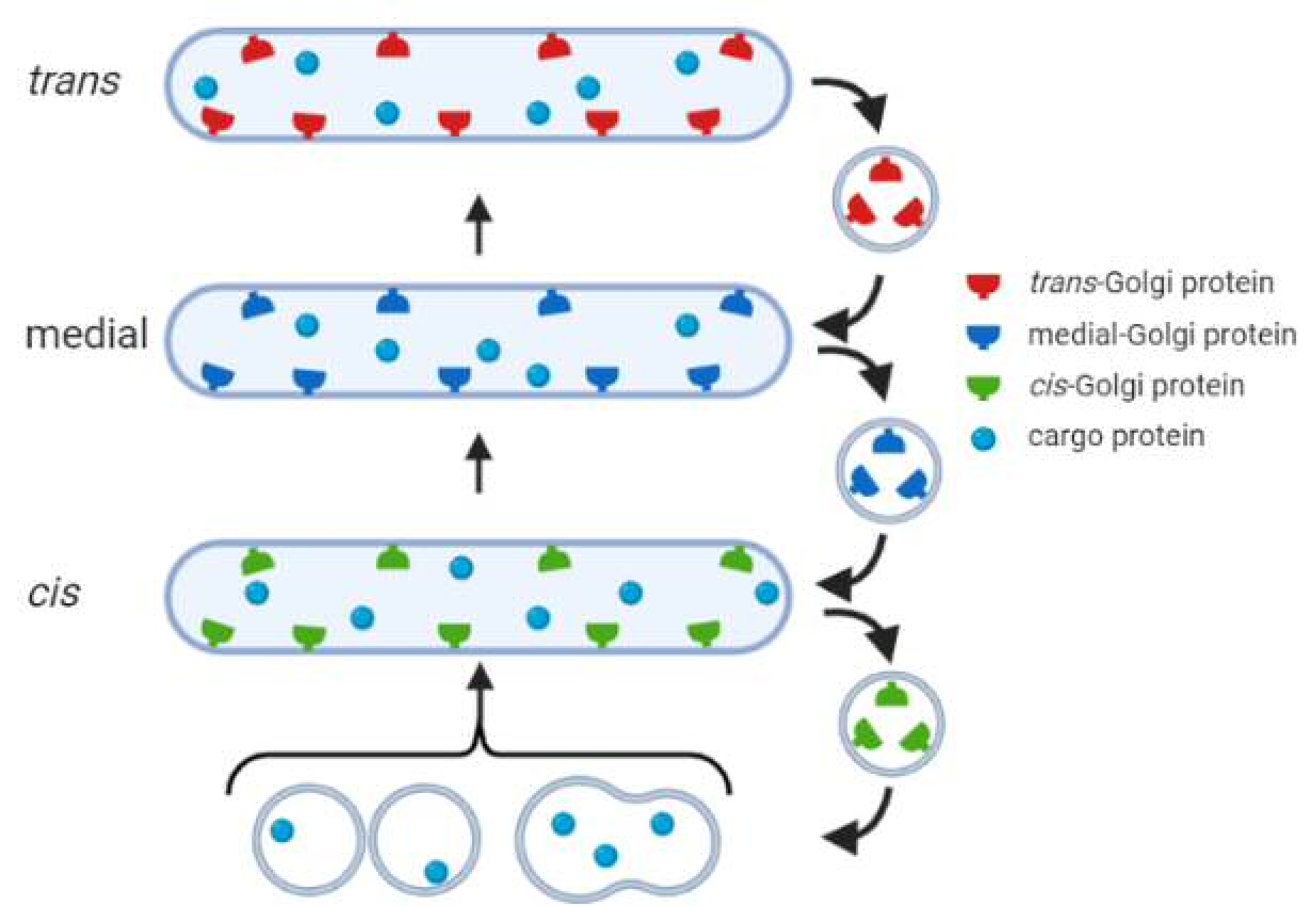
Figure 2. The route a glycan can take on its biosynthetic journey from the ER to the trans-Golgi. The blue box denotes glycans that are oligomannose, green box denotes hybrid glycans, and the red box denotes complex. The first four glycans in the blue box are those that can be subject to the oligomannose quench in the stochastic simulation of glycan processing, which targets them for removal from further processing in the Golgi. Enzyme abbreviations: mannosidase I (MAN1), mannosidase II (MAN2), fucosyltransferase 8 (FUT8), mammalian glucosamine transferase I-V (MGAT1-5), galactosyltransferases (GalT), sialyltransferases (SiaT).

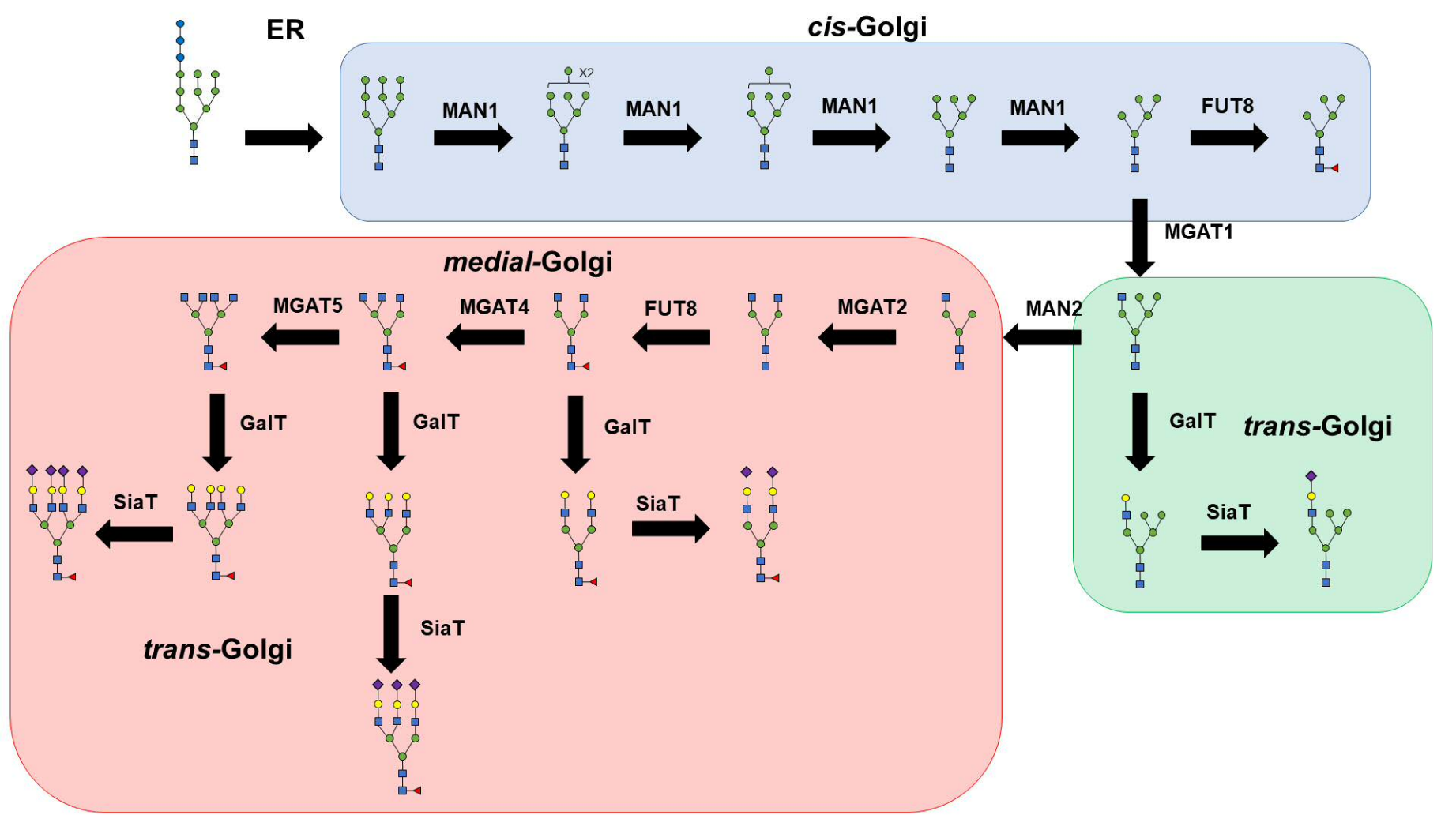
Figure 3. Proof of concept using drug treatment of HEK293T cells. Total “effective” enzymatic rate changes for seven different enzymes, obtained when fitting a glycan profile of untreated cells to a profile obtained following swainsonine treatment. Error bars are SD for n = 16 individual fitting runs. Figure adapted from Fisher et al. (12).

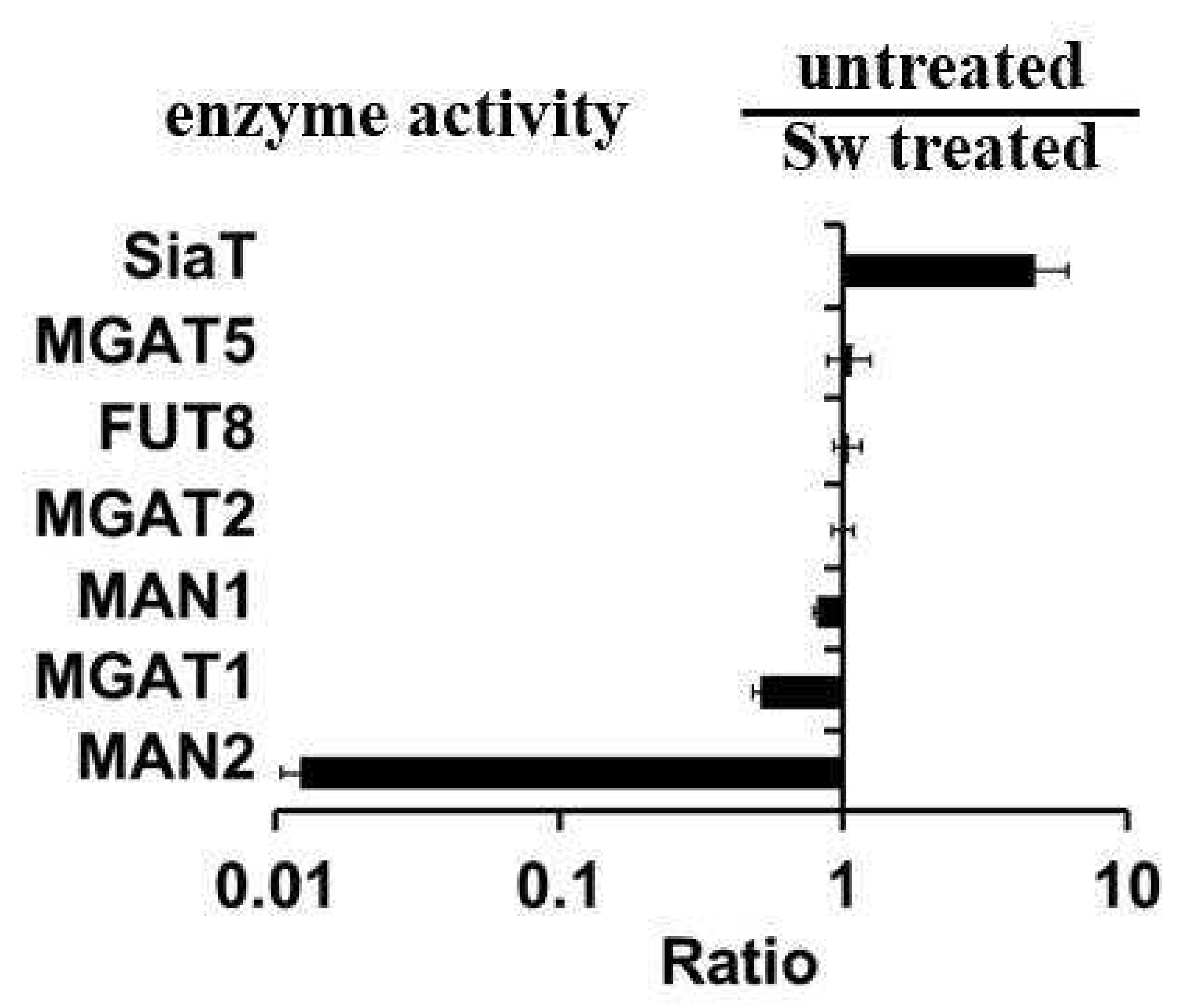
Table Captions

Table 1. A table showing the.xls file used as an input for the stochastic simulation. The table contains information on the Golgi enzymes required for glycosylation. (A) The first column denotes the names of the different enzymes. It is worth noting that the same enzyme is present across multiple lines to account for instances where multiple different glycan structures serve as substrates for the same enzyme. In some cases these different entries have a different EFER. E.g. compare ST3Gal2.1 and ST3Gal2.2. (B) The enzyme substrate is a string sequence that the simulation will search for and this will be replaced with a different string, the product (C). The linkage between residues is denoted by the numbers in a conventional manner. Residues enclosed in brackets represent single branches. The

underscore and lowercase letters represent the continuation from the previous residue not enclosed within the brackets. The ":" represents the termination of the branch and the "@" denotes the end of the N-glycan string. (D) The EFER for a particular enzyme across the three different cisternae used in this example. More cisternae can be added, and we have successfully run the model with four. (E) Three extra parameters that are required for the model: Man8GlcNAc2 fraction, Glc1Man9GlcNAc2 fraction, and transit time, respectively.


trans
medial
cis
trans-Golgi protein
medial-Golgi protein
cis-Golgi protein
cargo protein


ER
cis-Golgi
MAN1
X2
MAN1
MAN1
MAN1
FUT8
MGAT1
medial-Golgi
MGAT5
MGAT4
FUT8
MGAT2
MAN2
GalT
GalT
GalT
GalT
trans-Golgi
SiaT
SiaT
SiaT
SiaT
trans-Golgi


enzyme activity
untreated
Sw treated
SiaT
MGAT5
FUT8
MGAT2
MAN1
MGAT1
MAN2
0.01
0.1
1
10
Ratio

A

Enzyme Name	Enzyme Target	Enzyme Result	EFER 1	EFER 2	EFER 3
MAN1A1	1Man2.1Man:	1Man:	0.19449	1.75169	0.22350
MAN1A2	GlcNAc4.1GlcNAc4.1Man(3.1ManSS2.1Man:)_m6.1Man(3.1Man:)_m6.1Man:	GlcNAc4.1GlcNAc4.1Man(3.1Man:)_m6.1Man(3.1Man:)_m6.1Man:	0.02193	0.19751	0.02520
EndoMAN	GlcNAc4.1GlcNAc4.1Man(3.1ManSS2.1Man2.1Man3.1Gluc:)_m6.1Man(3.1Man2.1ManGG:)_m6.1Man2.1ManGG:	GlcNAc4.1GlcNAc4.1Man(3.1ManSS2.1Man:)_m6.1Man(3.1Man2.1Man:)_m6.1Man2.1Man:	0.03064	0.00533	0.00408
MAN2A	Man(3.1Man2.1GlcNAc:)_m6.1Man(3.1Man:)_m6.1Man:	Man(3.1Man2.1GlcNAc:)_m6.1Man:	0.50131	2.94670	1.91448
MGAT1	4.1Man(3.1Man:)_m6.1Man(3.1Man:)_m6.1Man:	4.1Man(3.1Man2.1GlcNAc:)_m6.1Man(3.1Man:)_m6.1Man:	0.00712	0.63680	0.09882
MGAT2.1	GlcNAc:)_m6.1Man:	GlcNAc:)_m6.1Man2.1GlcNAc:	1.79680	6.50778	5.89688
MGAT2.2	GlcNAc4.1Gal:)_m6.1Man:	GlcNAc4.1Gal:)_m6.1Man2.1GlcNAc:	1.79680	6.50778	5.89688
MGAT2.3	GlcNAc4.1Gal6.2Sia:)_m6.1Man:	GlcNAc4.1Gal6.2Sia:)_m6.1Man2.1GlcNAc:	1.79680	6.50778	5.89688
MGAT4B	3.1Man2.1GlcNAc:	3.1Man(2.1GlcNAc:)_m4.1GlcNAc:	0.00418	0.02006	0.01951
MGAT5B	6.1Man2.1GlcNAc:	6.1Man(2.1GlcNAc:)_m6.1GlcNAc:	0.02005	0.16863	0.24529
FUT8.1	GlcNAc4.1GlcNAc4.1Man(3.1Man2.1GlcNAc:)_m6.1Man	GlcNAc(6.1Fuc:)_g4.1GlcNAc4.1Man(3.1Man2.1GlcNAc:)_m6.1Man	0.04918	0.17729	0.07166
FUT8.2	GlcNAc4.1GlcNAc4.1Man(3.1Man(2.1GlcNAc:)_m4.1GlcNAc	GlcNAc(6.1Fuc)_g4.1GlcNAc4.1Man(3.1Man(2.1GlcNAc:)_m4.1GlcNAc	0.04918	0.17729	0.07166
FUT8.3	GlcNAc4.1GlcNAc4.1Man(3.1Man2.1GlcNAc4.1Gal:)_m6.1Man	GlcNAc(6.1Fuc:)_g4.1GlcNAc4.1Man(3.1Man2.1GlcNAc4.1Gal:)_m6.1Man	0.04918	0.17729	0.07166
FUT8.4	GlcNAc4.1GlcNAc4.1Man(3.1Man2.1GlcNAc4.1Gal6.2Sia:)_m6.1Man	GlcNAc(6.1Fuc:)_g4.1GlcNAc4.1Man(3.1Man2.1GlcNAc4.1Gal6.2Sia:)_m6.1Man	0.04918	0.17729	0.07166
FUT8.5	GlcNAc4.1GlcNAc4.1Man(3.1Man(2.1GlcNAc4.1Gal:)_m4.1GlcNAc	GlcNAc(6.1Fuc:)_g4.1GlcNAc4.1Man(3.1Man(2.1GlcNAc4.1Gal:)_m4.1GlcNAc	0.04918	0.17729	0.07166
FUT8.6	GlcNAc4.1GlcNAc4.1Man(3.1Man(2.1GlcNAc4.1Gal6.2Sia:)_m4.1GlcNAc	GlcNAc(6.1Fuc:)_g4.1GlcNAc4.1Man(3.1Man(2.1GlcNAc4.1Gal6.2Sia:)_m4.1GlcNAc	0.04918	0.17729	0.07166
FUT8.7	GlcNAc4.1GlcNAc4.1Man(3.1Man:)_m6.1Man(3.1Man:)_m6.1Man:	GlcNAc(6.1Fuc:)_g4.1GlcNAc4.1Man(3.1Man:)_m6.1Man(3.1Man:)_m6.1Man:	0.00271	0.00978	0.00395
FUT8OFF	F:	:	0.09945	0.09945	0.09945
ANTFUT	GlcNAc4.1Gal	GlcNAc(3.1Fuc:)_g4.1Gal	0.00100	0.01408	0.01105
B4GALT1	2.1GlcNAc:)_m6.1Man:	2.1GlcNAc4.1Gal:)_m6.1Man:	0.01002	0.07202	0.20616
B4GALT1H	2.1GlcNAc:)_m6.1Man(3.1Man:)_m6.1Man:	2.1GlcNAc4.1Gal:)_m6.1Man(3.1Man:)_m6.1Man:	0.01002	0.07202	0.20616
B4GALT1/2	2.1GlcNAc:)_m6.1Man2.1GlcNAc	2.1GlcNAc4.1Gal:)_m6.1Man2.1GlcNAc	0.01002	0.07202	0.20616
B4GALT2	6.1Man2.1GlcNAc:	6.1Man2.1GlcNAc4.1Gal:	0.01002	0.07202	0.20616
B4GALT1/3	2.1GlcNAc:)_m6.1Man(2.1GlcNAc	2.1GlcNAc4.1Gal:)_m6.1Man(2.1GlcNAc	0.00356	0.02560	0.07330
B4GALT1/4	2.1GlcNAc:)_m4.1GlcNAc	2.1GlcNAc4.1Gal:)_m4.1GlcNAc	0.00356	0.02560	0.07330
B4GALT2/3	6.1Man(2.1GlcNAc:	6.1Man(2.1GlcNAc4.1Gal:	0.00356	0.02560	0.07330
B4GALT4	4.1GlcNAc:	4.1GlcNAc4.1Gal:	0.00356	0.02560	0.07330
B4GALT3	6.1GlcNAc:	6.1GlcNAc4.1Gal:	0.00356	0.02560	0.07330
ST3Gal2	3.1Man2.1GlcNAc4.1Gal:	3.1Man2.1GlcNAc4.1Gal6.2Sia:	0.00600	0.00671	0.08412
ST3Gal2.1	3.1Man(2.1GlcNAc4.1Gal:)_m6.1GlcNAc	3.1Man(2.1GlcNAc4.1Gal6.2Sia:)_m6.1GlcNAc	0.00600	0.00671	0.08412
ST3Gal2.2	.1Gal:	.1Gal6.2Sia:	0.00275	0.00308	0.03862
OM Quench	1Man2.1Man:	1Man2.1ManP:	0.05286	0.01177	0.01404
			0.11329	0.03100	7.25802

B

C

D

E