Glycan processing in the Golgi – optimal information coding and constraints on cisternal number and enzyme specificity

Alkesh Yadav,[1] Quentin Vagne,[2] Pierre Sens,[2] Garud Iyengar,[3] and Madan Rao[4]

[1] *Raman Research Institute, Bangalore 560080, India*

[2] *Laboratoire Physico Chimie Curie, Institut Curie, CNRS UMR168, 75005 Paris, France*

[3] *Industrial Engineering and Operations Research, Columbia University, New York 10027, USA*

[4] *Simons Centre for the Study of Living Machines, National Centre for Biological Sciences (TIFR), Bangalore 560065, India**

(Dated: May 18, 2020)

Many proteins that undergo sequential enzymatic modification in the Golgi cisternae are displayed at the plasma membrane as cell identity markers. The modified proteins, called glycans, represent a molecular code. The fidelity of this *glycan code* is measured by how accurately the glycan synthesis machinery realises the desired target glycan distribution for a particular cell type and niche. In this paper, we quantitatively analyse the tradeoffs between the number of cisternae and the number and specificity of enzymes, in order to synthesize a prescribed target glycan distribution of a certain complexity. We find that to synthesize complex distributions, such as those observed in real cells, one needs to have multiple cisternae and precise enzyme partitioning in the Golgi. Additionally, for fixed number of enzymes and cisternae, there is an optimal level of specificity of enzymes that achieves the target distribution with high fidelity. Our results show how the complexity of the target glycan distribution places functional constraints on the Golgi cisternal number and enzyme specificity.

* madan@ncbs.res.in

I. INTRODUCTION

A majority of the proteins synthesized in the endoplasmic reticulum (ER) are transferred to the Golgi cisternae for further chemical modification by glycosylation [1], a process that sequentially and covalently attaches sugar moieties to proteins, catalyzed by a set of enzymatic reactions within the ER and the Golgi cisternae. These enzymes, called glycosyltransferases, are localized in the ER and cis-medial and trans Golgi cisternae in a specific manner [2, 3]. Glycans, the final products of this glycosylation assembly line are delivered to the plasma membrane (PM) conjugated with proteins, whereupon they engage in multiple cellular functions, including immune recognition, cell identity markers, cell-cell adhesion and cell signalling [2–6]. This *glycan code* [7, 8], representing information [9] about the cell, is generated dynamically, following the biochemistry of sequential enzymatic reactions and the biophysics of secretory transport [4, 10, 11].

In this paper, we will focus on the role of glycans as markers of cell identity. For the glycans to play this role, they must inevitably represent a molecular code [4, 7, 11]. While the functional consequences of glycan alterations have been well studied, the glycan code has remained an enigma [7, 11–13]. In this paper, we study one aspect of molecular coding, namely the *fidelity* of this molecular code generation. While it has been recognised that fidelity of the glycan code is necessary for reliable cellular recognition [14], a quantitative measure of fidelity of the code and its implications for cellular structure and organization are lacking.

There are two aspects of the cell-type specific glycan code that have an important bearing on quantifying fidelity. The first is that extant glycan distributions have high *complexity*, owing to evolutionary pressures arising from (a) reliable cell type identification amongst a large set of different cell types in a complex organism, the preservation and diversification of "self-recognition" [5], (b) pathogen-mediated selection pressures [2, 4, 6], and (c) *herd immunity* within a heterogenous population of cells of a community [15] or within a single organism [5]. Here, we will interpret this to mean that the *target distribution* of glycans of a given cell type is complex; in Sect. II we define a quantitative measure for complexity and demonstrate its implications in the context of *human* T-cells. The second is that the cellular machinery for the synthesis of glycans, which involves sequential chemical processing via cisternal resident enzymes and cisternal transport, is subject to variation and noise [4, 10, 11]; the *synthesized glycan distribution* is, therefore, a function of cellular parameters such as the number and specificity of enzymes, inter-cisternal transfer rates, and number of cisternae. We will discuss an explicit model of the cellular synthesis machinery in Sect. III.

In this paper, we define fidelity as the minimum achievable Kullback-Leibler (KL) divergence [16, 17] between the synthesized distribution of glycans and the target glycan distribution. This KL divergence is a function of the cellular parameters governing glycan synthesis, such as the number and specificity of enzymes, inter-cisternal transfer rates, and number of cisternae (Sect. V). We analyze the tradeoffs between the number of cisternae and the number and specificity of enzymes, in order to achieve a prescribed target glycan distribution with high fidelity (Sect. VI). Our analysis leads to a number of interesting results, of which we list a few here: (i) In order to construct an accurate representation of a complex target distribution, such as those observed in real cells, one needs to have multiple cisternae and precise enzyme partitioning. Low complexity target distributions can be achieved with fewer cisternae. (ii) This definition of fidelity of the glycan code, allows us to provide a quantitative argument for the evolutionary requirement of multiple-compartments. (iii) For fixed number of enzymes and cisternae, there is an optimal level of specificity of enzymes that achieves the complex target distribution with high fidelity. (iv) Keeping the number of enzymes fixed, having low specificity or sloppy enzymes and larger cisternal number could give rise to a diverse repertoire of functional glycans, a strategy used in organisms such as plants and algae.

Stated another way, our results imply that the pressure to achieve the target glycan code for a given cell type, places strong constraints on the cisternal number and enzyme specificity [18]. This would suggest that a description of the nonequilibrium assembly of a fixed number of Golgi cisternae must combine the dynamics of chemical processing with membrane dynamics involving fission, fusion and transport [19, 20], opening up a new direction for future research.

II. COMPLEXITY OF GLYCAN CODE IN REAL CELLS

Since each cell type (in a niche) is identified with a distinct glycan profile [4, 7, 11], and this glycan profile is noisy because of the stochastic noise associated with the synthesis and transport [11–13], a large number of different cell types can be differentiated only if the cells are able to produce a large set of glycan profiles that are distinguishable

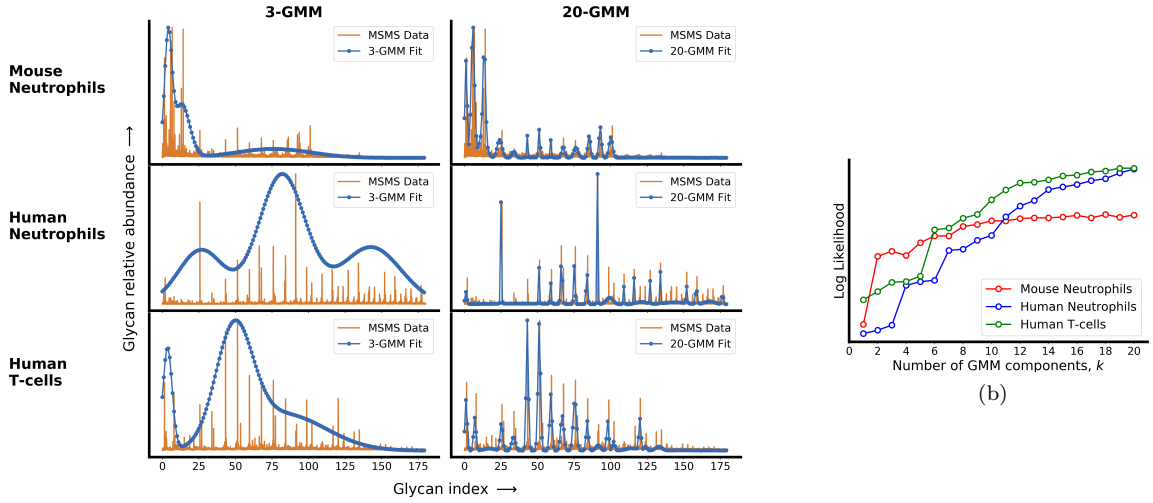


FIG. 1. Real cells display a complex glycan distribution. (a) Here we take the MSMS data from *mouse* neutrophils, *human* neutrophils and *human* T-cells and approximate these using Gaussian Mixture Models (GMM) of less complexity 3-GMM (left) and more complexity 20-GMM (right). (b) The change in log likelihood with increase in the number of GMM components for *mouse* and *human* neutrophils and *human* T-cells, shows a saturation at large enough values of k, indicating that these glycan distribution are complex. Details appear in Appendix G.

in the presence of this noise. A more complex or richer class of glycan profiles is able to support a larger number of well separated profiles, and therefore, a larger number cell types, or equivalently, a more complex organism[1]

In order to implement a quantitative measure of complexity, we first need a consistent way of smoothening or coarse-graining the discrete glycan distribution to remove measurement and synthesis noise. In this paper, we approximate the glycan profile as mixture of Gaussian densities with specified number of components that are supported on a finite set of indices [21]. Since the complexity of k-component Gaussian is an increasing function of k, we use the number of component k and complexity interchangeably.

Using this definition we demonstrate that the glycan profiles of typical mammalian cells are very complex. We obtain target profiles for a given cell type from Mass Spectrometry coupled with determination of molecular structure (MSMS) measurements [22]. Fig. 1 shows the the MSMS data from *human* T-cells and *human* and *mouse* neutrophils [22], and their coarse-grained representations using Gaussian mixture models (GMM) of differing complexity - a low complexity $k = 3$ GMM and high complexity $k = 20$ GMM. It is clear from Fig. 1, that the more complex $k = 20$ GMM is a better representation of the MSMS data as compared to the less complex $k = 3$ GMM. Indeed the $k = 20$ Gaussian mixture model is the best compromise between faithfulness of the representation and cost of an additional component, as seen from the saturation of the likelihood function [17]. Details of this systematic coarse-graining procedure appear in Sect. VI B and Appendix G.

Having demonstrated the complexity of the typical glycan distributions associated with a given cell type, we will now describe a general model of the cellular machinery that is capable of synthesizing glycans of the desired complexity. We expect that cells need a more elaborate mechanism to produce profiles from a more complex set.

III. SYNTHESIS OF GLYCANS IN THE GOLGI CISTERNAE

The glycan display at the cell surface is a result of proteins that flux through and undergo sequential chemical modification in the secretory pathway, comprising an array of Golgi cisternae situated between the ER and the PM,

[1] A rigorous definition of complexity can be given in terms of the Kullback-Leibler metric [16, 17] between two glycan profiles. We declare that two profiles are distinguishable only if the Kullback-Leibler distance between the profiles is more than a given tolerance. This tolerance is an increasing function of the noise. We define the *complexity* of a set of possible glycan profiles as the size of the largest subset such that the Kullback-Leibler distance of any pair of profiles is larger than the tolerance.

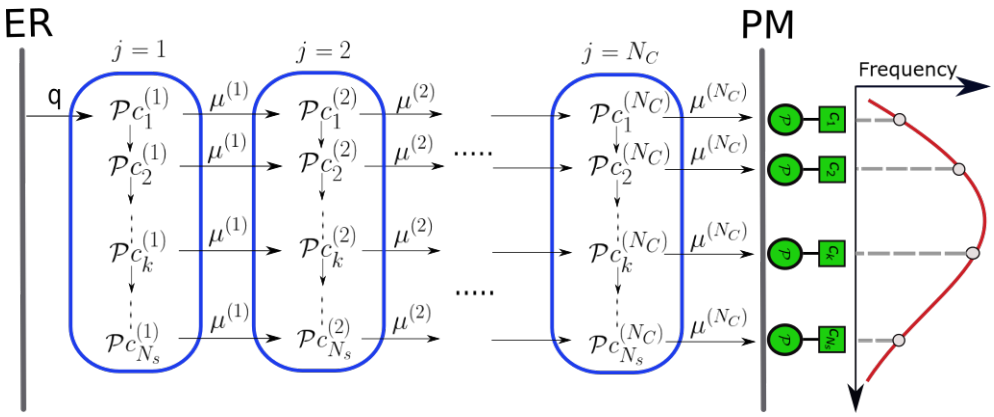


FIG. 2. Enzymatic reaction and transport network in the secretory pathway. Represented here is the array of Golgi cisternae (blue) indexed by $j = 1, \ldots, N_C$ situated between the ER and PM. Glycan-binding proteins $\mathcal{P}c_1^{(1)}$ are injected from the ER to cisterna-1 at rate q. Superimposed is transition network of chemical reactions (column) - intercisternal transfer (rows), the latter with rates $\mu^{(j)}$. $\mathcal{P}c_k^{(j)}$ denotes the acceptor substrate in compartment j and the glycosyl donor c_0 is chemostated in each cisterna. This results in a frequency distribution of glycans displayed at the PM (red curve), that is representative of the cell type.

as depicted in Fig. 2. Glycan-binding proteins (GBPs) are delivered from the ER to the first cisterna, whereupon they are processed by the resident enzymes in a sequence of steps that constitute the N-glycosylation process [2]. A generic enzymatic reaction in the cisterna involves the catalysis of a group transfer reaction in which the monosaccharide moiety of a simple sugar donor substrate, e.g. UDP-Gal, is transferred to the acceptor substrate, by a Michaelis-Menten (MM) type reaction [2]

$$\begin{aligned} &\text{Acceptor} + \text{glycosyl donor} + \text{Enzyme} \\ &\quad \underset{\omega_b}{\overset{\omega_f}{\rightleftharpoons}} \; [\text{Acceptor} \cdot \text{glycosyl donor} \cdot \text{Enzyme}] \; \xrightarrow{\omega_c} \; \text{glycosylated acceptor} + \text{nucleotide} + \text{Enzyme} \end{aligned} \tag{1}$$

From the first cisterna, the proteins with attached sugars are delivered to the second cisterna at a given inter-cisternal transfer rate, where further chemical processing catalysed by the enzymes resident in the second cisterna occurs. This chemical processing and inter-cisternal transfer continues until the last cisterna, thereupon the fully processed glycans are displayed at the PM [2]. The network of chemical processing and inter-cisternal transfer forms the basis the physical model that we will describe next.

Any physical model of such a network of enzymatic reactions and cisternal transfer needs to be augmented by reaction and transfer rates and chemical abundances. To obtain the range of allowable values for the reaction rates and chemical abundances, we use the elaborate enzymatic reaction models, such as the KB2005 model [23–25] (with a network of $22,871$ chemical reactions and 7565 oligosaccharide structures) that predict the N-glycan distribution based on the activities and levels of processing enzymes distributed in the Golgi-cisternae of mammalian cells, and compare these predictions with N-glycan mass spectrum data. For the allowable rates of cisternal transfer, we rely on the recent study by Ungar and coworkers [26, 27], whose study shows how the overall Golgi transit time and cisternal number, can be tuned to engineer a homogeneous glycan distribution.

IV. MODEL DEFINITION

A. Chemical reaction and transport network in cisternae

With this background, we now define our quantitative model for chemical processing and transport in the secretory pathway. We consider an array of N_C Golgi cisternae, labelled by $j = 1, \ldots, N_C$, between the ER and the PM (Fig. 2).

Glycan-binding proteins (GBPs), denoted as $\mathcal{P}c_1^{(1)}$, are delivered from the ER to cisterna-1 at an injection rate q. It is well established that the concentration of the glycosyl donor in the j-th cisterna is chemostated [2, 28–30], thus in our model we hold its concentration $c_0^{(j)}$ constant in time for each j. The acceptor $\mathcal{P}c_1^{(1)}$ reacts with $c_0^{(1)}$ to form the glycosylated acceptor $\mathcal{P}c_2^{(1)}$, following an MM-reaction (1) catalysed by the appropriate enzyme. The acceptor $\mathcal{P}c_2^{(1)}$ has the potential of being transformed into $\mathcal{P}c_3^{(1)}$, and so on, provided the requisite enzymes are present in that cisterna. This leads to the sequence of enzymatic reactions $\mathcal{P}c_1^{(1)} \to \mathcal{P}c_2^{(1)} \to \ldots \mathcal{P}c_k^{(1)} \to \ldots$, where k enumerates the sequence of glycosylated acceptors, using a consistent scheme (such as in [23]). The glycosylated GBPs are transported from cisterna-1 to cisterna-2 at an inter-cisternal transfer rate $\mu^{(1)}$, whereupon similar enzymatic reactions proceed. The processes of intra-cisternal chemical reactions and inter-cisternal transfer continue to the other cisternae and form a network as depicted in Fig. 2. Although, in this paper, we focus on a sequence of reactions that form a line-graph, the methodology we propose extends to tree-like reaction sequences, and more generally to reaction sequences that form a directed acyclic graphs [31].

Let N_s denote the maximum number of possible glycosylation reactions in each cisterna j, catalysed by enzymes labelled as $E_\alpha^{(j)}$, with $\alpha = 1, \ldots, N_E$, where N_E is the total number of enzyme species in each cisterna. Since many substrates can compete for the substrate binding site on each enzyme, one expects in general that $N_s \gg N_E$. The configuration space of the network Fig. 2 is $N_s \times N_C$. For the N-glycosylation pathway in a typical mammalian cell, $N_s = 2 \times 10^4$, $N_E = 10 - 20$ and $N_C = 4 - 8$ [23–25, 27]. We account for the fact that the enzymes have specific cisternal localisation, by setting their concentrations to zero in those cisternae where they are not present.

Now the action of enzyme $E_\alpha^{(j)}$ on the substrate $\mathcal{P}c_k^{(j)}$ in cisterna j is given by

$$\mathcal{P}c_k^{(j)} + E_\alpha^{(j)} \underset{\omega_b(j,k,\alpha)}{\overset{\omega_f(j,k,\alpha)c_0^{(j)}}{\rightleftharpoons}} \left[E_\alpha^{(j)} - \mathcal{P}c_k^{(j)} - c_0^{(j)}\right] \xrightarrow{\omega_c(j,k,\alpha)} \mathcal{P}c_{k+1}^{(j)} + E_\alpha^{(j)} \tag{2}$$

In general, the forward, backward and catalytic rates ω_f, ω_b and ω_c, respectively, depend on the cisternal label j, the reaction label k, and the enzyme label α, that parametrise the MM-reactions [32]. For instance, structural studies on glycosyltransferase-mediated synthesis of glycans [33], would suggest that the forward rate ω_f to depend on the binding energy of the enzyme $E_\alpha^{(j)}$ to acceptor substrate $\mathcal{P}c_k^{(j)}$ and a *physical variable that characterises cisterna-j*.

A potential candidate for such a cisternal variable is pH [34], whose value is maintained homeostatically in each cisterna [35]; changes in pH can affect the shape of an enzyme (substrate) or their charge properties, and in general the reaction efficiency of an enzyme has a pH optimum [32]. Another possible candidate for a cisternal variable is membrane bilayer thickness [36] - indeed both pH [37] and membrane thickness are known to have a gradient across the Golgi cisternae. We take $\omega_f(j,k,\alpha) \propto P^{(j)}(k,\alpha)$, where $P^{(j)}(k,\alpha) \in (0,1)$, is the binding probability of enzyme $E_\alpha^{(j)}$ with substrate $\mathcal{P}c_k^{(j)}$, and define the binding probability $P^{(j)}(k,\alpha)$ using a biophysical model, similar in spirit to the Monod-Wyman-Changeux model of enzyme kinetics [38, 39], of enzyme-substrate induced fit.

Let $\boldsymbol{l}_\alpha^{(j)}$ and $\boldsymbol{l}_k$ denote, respectively, the optimal "shape" for enzyme $E_\alpha^{(j)}$ and the substrate $\mathcal{P}c_k^{(j)}$. We assume that the mismatch (or distortion) energy between the substrate k and enzyme $E_\alpha^{(j)}$ is $\|\boldsymbol{l}_k - \boldsymbol{l}_\alpha^{(j)}\|$, with a binding probability given by,

$$P^{(j)}(k,\alpha) = \exp\left(-\sigma_\alpha^{(j)} \|\boldsymbol{l}_k - \boldsymbol{l}_\alpha^{(j)}\|\right) \tag{3}$$

where $\|.\|$ is a distance metric defined on the space of $\boldsymbol{l}_\alpha^{(j)}$ (e.g., the square of the ℓ_2-norm would be related to an elastic distortion model [40]) and the vector $\boldsymbol{\sigma} \equiv [\sigma_\alpha^{(j)}]$ parametrises *enzyme specificity*. A large value of $\sigma_\alpha^{(j)}$ indicates a highly specific enzyme, a small value of $\sigma_\alpha^{(j)}$ indicates a promiscuous or sloppy enzyme. It is recognised that the degree of enzyme specificity or sloppiness is an important determinant of glycan distribution [2, 41–43].

As in [23–25], our synthesis model is mean-field, in that we ignore stochasticity in glycan synthesis that may arise from low copy numbers of substrates and enzymes, multiple substrates competing for the same enzymes, and kinetics of inter-cisternal transfer. Then the usual MM-steady state condition on (2), which assumes that the concentration of the intermediate enzyme-substrate complex does not change with time, implies

$$\left[E_\alpha^{(j)} - \mathcal{P}c_k^{(j)} - c_0^{(j)}\right] = \frac{\omega_f(j,k,\alpha)\, c_0^{(j)}}{\omega_b(j,k,\alpha) + \omega_c(j,k,\alpha)} E_\alpha^{(j)} c_k^{(j)}.$$

where $c_k^{(j)}$ is the *concentration* of the acceptor substrate $\mathcal{P}c_k^{(j)}$ in compartment j.

Together with the constancy of the total enzyme concentration, $\left[E_\alpha^{(j)}\right]_{tot} = E_\alpha^{(j)} + \sum_{k=1}^{N_s}\left[E_\alpha^{(j)} - \mathcal{P}c_k^{(j)} - c_0^{(j)}\right]$, this immediately fixes the kinetics of product formation (not including inter-cisternal transport),

$$\frac{dc_{k+1}^{(j)}}{dt} = \sum_{\alpha=1}^{N_E} \frac{V(j,k,\alpha)P^{(j)}(k,\alpha)c_k^{(j)}}{M(j,k,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)} \tag{4}$$

where

$$M(j,k,\alpha) = \frac{\omega_b(j,k,\alpha)+\omega_c(j,k,\alpha)}{\omega_f(j,k,\alpha)c_0^{(j)}}P^{(j)}(k,\alpha)$$

and

$$V(j,k,\alpha) = \omega_c(j,k,\alpha)\left[E_\alpha^{(j)}\right]_{tot}.$$

From the above, the experimentally measurable parameters V_{max} and MM-constant K_M, for each (j,k,α) can be easily read out. As is the usual case, the maximum velocity V_{max} is not an intrinsic property of the enzyme, because it is dependent on the enzyme concentration $\left[E_\alpha^{(j)}\right]_{tot}$; while $K_M(j,k,\alpha) = M(j,k,\alpha)c_0^{(j)}/P^{(j)}(k,\alpha)$ is an intrinsic parameter of the enzyme and the enzyme-substrate interaction. The enzyme catalytic efficiency, the so-called "k_{cat}/K_M" $\propto P^{(j)}(k,\alpha)$ and is high for *perfect* enzymes [44] with minimum mismatch.

We now add to this chemical reaction kinetics, the rates of injection (q) and inter-cisternal transport $\mu^{(j)}$ from the cisterna j to $j+1$; in Appendix A we display the complete set of equations that describe the changes in the substrate concentrations $c_k^{(j)}$ with time. These kinetic equations automatically obey the conservation law for the protein concentration (p). Rescaling the kinetic parameters in terms of the injection rate q, i.e. $V(j,k,\alpha) = V(j,k,\alpha)/q$ and $\mu^{(j)} = \mu^{(j)}/q$, we see that the steady state concentrations of the glycans in each cisterna satisfy the following recursion relations (see, Appendix A). In the first cisterna,

$$\begin{aligned}
c_1^{(1)} &= \frac{1}{\mu^{(1)} + \sum_{\alpha=1}^{N_E}\frac{V(1,1,\alpha)P^{(1)}(1,\alpha)c_1^{(1)}}{M(1,1,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(1)}(k',\alpha)c_{k'}^{(1)}}{M(1,k',\alpha)}\right)}} \\
c_k^{(1)} &= \frac{\sum_{\alpha=1}^{N_E}\frac{V(1,k-1,\alpha)P^{(1)}(k-1,\alpha)c_{k-1}^{(1)}}{M(1,k-1,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(1)}(k',\alpha)c_{k'}^{(1)}}{M(1,k',\alpha)}\right)}}{\mu^{(1)} + \sum_{\alpha=1}^{N_E}\frac{V(1,k,\alpha)P^{(1)}(k,\alpha)c_k^{(1)}}{M(1,k,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(1)}(k',\alpha)c_{k'}^{(1)}}{M(1,k',\alpha)}\right)}} \\
c_{N_s}^{(1)} &= \frac{\sum_{\alpha=1}^{N_E}\frac{V(1,N_s-1,\alpha)P^{(1)}(N_s-1,\alpha)c_{N_s-1}^{(1)}}{M(1,N_s-1,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(1)}(k',\alpha)c_{k'}^{(1)}}{M(1,k',\alpha)}\right)}}{\mu^{(1)}}
\end{aligned} \tag{5}$$

and in cisternae $j \geq 2$,

$$
\begin{aligned}
c_1^{(j)} &= \frac{\mu^{(j-1)} c_1^{(j-1)}}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} \frac{V(j,1,\alpha) P^{(j)}(1,\alpha) c_1^{(j)}}{M(j,1,\alpha)\left(1+\sum_{k'=1}^{N_s} \frac{P^{(j)}(k',\alpha) c_{k'}^{(j)}}{M(j,k',\alpha)}\right)}} \\
c_k^{(j)} &= \frac{\mu^{(j-1)} c_k^{(j-1)} + \sum_{\alpha=1}^{N_E} \frac{V(j,k-1,\alpha) P^{(j)}(k-1,\alpha) c_{k-1}^{(j)}}{M(j,k-1,\alpha)\left(1+\sum_{k'=1}^{N_s} \frac{P^{(j)}(k',\alpha) c_{k'}^{(j)}}{M(j,k',\alpha)}\right)}}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} \frac{V(j,k,\alpha) P^{(j)}(k,\alpha) c_k^{(j)}}{M(j,k,\alpha)\left(1+\sum_{k'=1}^{N_s} \frac{P^{(j)}(k',\alpha) c_{k'}^{(j)}}{M(j,k',\alpha)}\right)}} \\
c_{N_s}^{(j)} &= \frac{\mu^{(j-1)} c_{N_s}^{(j-1)} + \sum_{\alpha=1}^{N_E} \frac{V(j,N_s-1,\alpha) P^{(j)}(N_s-1,\alpha) c_{N_s-1}^{(j)}}{M(j,N_s-1,\alpha)\left(1+\sum_{k'=1}^{N_s} \frac{P^{(j)}(k',\alpha) c_{k'}^{(j)}}{M(j,k',\alpha)}\right)}}{\mu^{(j)}}
\end{aligned}
\tag{6}
$$

Equations (5)-(6) automatically imply that the total steady state glycan concentration in each cisterna $j = 1, \ldots, N_c$ is given by

$$
\sum_{k=1}^{N_s} c_k^{(j)} = \frac{1}{\mu^{(j)}}.
$$

These nonlinear recursion equations (5)-(6) have to be solved numerically to obtain the steady state glycan concentrations, $\mathbf{c} \equiv c_k^{(j)}$, as a function of the independent vectors $\mathbf{M} \equiv [M(j,k,\alpha)]$, $\mathbf{V} \equiv [V(j,k,\alpha)]$, and $\mathbf{L} \equiv [P^{(j)}(k,\alpha)]$, the transport rates $\boldsymbol{\mu} \equiv [\mu^{(j)}]$ and specificity, $\boldsymbol{\sigma} \equiv [\sigma_\alpha^{(j)}]$.

V. OPTIMIZATION PROBLEM

Now, with both the protocol for determining the target glycan distribution and the sequential chemical processing model in hand, we can precisely define the optimization problem referred to in the Introduction. Let $\mathbf{c}^*$ denote the "target" concentration distribution[2] for a particular cell type, i.e. the goal of the sequential synthesis mechanism described in Sect. IV A is to approximate $\mathbf{c}^*$. Let $\bar{\mathbf{c}}$ denote the steady state glycan concentration distribution displayed on the PM - (6) implies that $\bar{c}_k = \mu^{(N_C)} c_k^{(N_C)}$, $k = 1, \ldots, N_s$. We measure the fidelity between the $\mathbf{c}^*$ and $\bar{\mathbf{c}}$ by the Kullback-Leibler metric [16, 17],

$$
D_{KL}(\mathbf{c}^* \| \bar{\mathbf{c}}) = \sum_{k=1}^{N_s} c_k^* \ln\left(\frac{c_k^*}{\bar{c}_k}\right) = \sum_{k=1}^{N_s} c_k^* \ln\left(\frac{c_k^*}{c_k^{(N_C)} \mu^{(N_C)}}\right) \tag{7}
$$

Thus, the problem of designing a sequential synthesis mechanism that approximates $\mathbf{c}^*$ for a given enzyme specificity $\boldsymbol{\sigma}$, transport rate $\boldsymbol{\mu}$, number of enzymes N_E, and number of cisternae N_C is given by

$$
\textit{Optimization A}: \quad \min_{\mathbf{M},\, \mathbf{V},\, \mathbf{L} \geq \mathbf{0}} D_{KL}(\mathbf{c}^* \| \bar{\mathbf{c}}) \tag{8}
$$

There is separation of time scales implicit in Optimization A – the chemical kinetics of the production of glycans and their display on the PM happens over cellular time scales, while the issues of tradeoffs and changes of parameters are driven over evolutionary timescales.

---

[2] We normalize the distribution so that $\sum_{k=1}^{N_s} c_k^* = 1$.

Optimization A, though well-defined, is a hard problem, since the steady state concentrations (6) are not *explicitly* known in terms of the parameters $(\mathbf{M}, \mathbf{V}, \mathbf{L})$. In Appendix B, we formulate an alternative problem *Optimization B* in which the steady state concentrations are defined explicitly in terms of a new parameters $\mathbf{R}$ and $\mathbf{L}$, and in Appendix C we prove that Optimization A and Optimization B are exactly equivalent. This is a crucial insight that allows us to obtain all the results that follow.

In Appendix H, we describe the variant of the Sequential Quadratic Programming (SQP) [45], that we use to numerically solve the optimization problem.

VI. RESULTS OF OPTIMIZATION

To start with, the dimension of the optimization search space is extremely large $\approx O(N_s \times N_E \times N_C)$. To make the optimization search more manageable, we ignore the k-dependence of the vectors $(\mathbf{M}, \mathbf{V})$, (or, alternatively of $\mathbf{R}$, see Appendix B for details). The dependence on the reaction rates on the glycosyl substrate is still present in the forward reactions via the enzyme-substrate binding probability $P^{(j)}(k, \alpha)$. We further assume that shape function is a number, $\boldsymbol{l}_\alpha^{(j)} = l_\alpha^{(j)}$ and that $\boldsymbol{l}_k = k$. Finally we will drop the dependence of the specificity on α and j, and take it to be a scalar σ. To fix our model, we will take the distortion energy that appears in (3) to be the linear form $|l_k - l_\alpha^{(j)}|$. Other metrics, such as $|l_k - l_\alpha^{(j)}|^2$, corresponding to the elastic distortion model [40], do not pose any computational difficulties, and we see that the results of our optimization remain qualitatively unchanged.

These restrictions significantly reduce the dimension of the optimization search, so much so that in certain limits, we can solve the problem analytically[3]. This helps us obtain some useful heuristics (Appendix E) on how to tune the parameters, e.g. N_E, N_C, σ, and others, in order to generate glycan distributions $\mathbf{c}$ of a given complexity. These heuristics inform our more detailed optimization using "realistic" target distributions.

The calculations in Appendix E imply, as one might expect, that the synthesis model needs to be more elaborate, i.e., needs a larger number of cisternae N_C or a larger number of enzymes N_E, in order to produce a more complex glycan distribution. For a real cell type in a niche, the specific elaboration of the synthesis machinery, would depend on a variety of control costs associated with increasing N_E and N_C. While an increase in the number of enzymes would involve genetic and transcriptional costs, the costs involved in increasing the number of cisternae could be rather subtle.

Notwithstanding the relative control costs of increasing N_E and N_C, it is clear from the special case, that increasing the number of cisternae achieves the goal of obtaining an accurate representation of the target distribution. Let us assume that the target distribution is $c_k^* = \delta(k - M)$ for a fixed $M \gg 1$, i.e. $c_k^* = 1$ when $k = M$, and 0 otherwise, and that the N_E enzymes that catalyse the reactions are highly specific. In this limit, Optimization A reduces to a simple enumeration exercise [46]: clearly one needs $N_E = M$, with one enzyme species for each of the $k = 1, \ldots, M$ reactions, in order to generate $\mathcal{P}c_M$. For a single Golgi cisterna with a finite cisternal residence time (finite μ), the chemical synthesis network will generate a significant steady state concentration of lower index glycans $\mathcal{P}c_k$ with $k < M$, contributing to a low fidelity. To obtain high fidelity, one needs multiple Golgi cisternae with a specific enzyme partitioning $(E_1, E_2, \ldots, E_M)$ with E_j enzymes in cisterna $j = 1, \ldots, N_c$. This argument can be generalised to the case where the target distribution is a finite sum of delta-functions. The more general case, where the enzymes are allowed to have variable specificity, needs a more detailed study, to which we turn to below.

A. Target distribution from coarse-grained MSMS

As discussed in Sect. II, we obtain the target glycan distribution from glycan profiles for real cells obtained using Mass Spectrometry coupled with determination of molecular structure (MSMS) measurements [22]. The raw MSMS data,

[3] In Appendix D we show that (B4) can be solved analytically in the limit $N_s \gg 1$, since the glycan index k can be approximated by a continuous variable, and the recursion relations for the steady state glycan concentrations (5)-(6) can be cast as a matrix differential equation. This allows us to obtain an *explicit* expression for the steady state concentration in terms of the parameters $(\mathbf{R}, \mathbf{L})$.

however, is not suitable as a target distribution. This is because it is very noisy, with chemical noise in the sample and Poisson noise associated with detecting discrete events being the most relevant [47]. This means that many of the small peaks in the raw data are not part of the signal, and one has to "smoothen" the distribution to remove the impact of noise.

We use MSMS data from *human* T-cells [22] for our analysis. As discussed in Sect. II, the Gaussian mixture models (GMM) are often used to approximate distributions with a mixed number of modes or peaks [17], or in our setting, a given fixed complexity. Here, we use a variation of the Gaussian mixture models (see Appendix G for details) to create a hierarchy of increasingly complex distributions to approximate the MSMS raw data. Thus, the 3-GMM and 20-GMM approximations represent the low and high complexity benchmarks, respectively. In Appendix G, we show that the likelihood for the glycan distribution of the *human* T-cell saturates at 20 peaks. Thus, statistically speaking, the *human* T-cell glycan distribution is accurately approximated by 20 peaks.

This hierarchy allows us to study the trade-off between the complexity of the target distribution and the complexity of the synthesis model needed to generate the distribution as follows. Let $\mathbf{T}^{(i)}$ denote the i^{th}-GMM approximation for the *human* T-cell MSMS data. We sample this target distribution at indices $k = 1, \ldots, N_s$, that represent the glycan indices, and then renormalize to obtain the discrete distribution $\{T_k^{(i)}, k = 1, \ldots, N_s\}$. Let $H(\mathbf{T}^{(i)}) := -\sum_{k=1}^{N_s} T_k^{(i)} \log T_k^{(i)}$ denote the entropy [16] of the i^{th}-GMM approximation. $H(\mathbf{T}^{(i)})$ quantifies statistical information in the target distribution $\mathbf{T}^{(i)}$. We evaluate the fidelity of the distribution generated by the synthesis model to this target distribution by the ratio of the Kullback-Leibler distance to the entropy of the target distribution:

$$\bar{D}(\sigma, N_E, N_C, \mathbf{T}^{(i)}) := \frac{D(\sigma, N_E, N_C, \mathbf{T}^{(i)})}{H(\mathbf{T}^{(i)})} \tag{9}$$

This normalization allows us evaluate the fidelity of the synthesis model to the target distribution $\mathbf{T}^{(i)}$ as a fraction of the total statistical information in the target distribution $\mathbf{T}^{(i)}$.

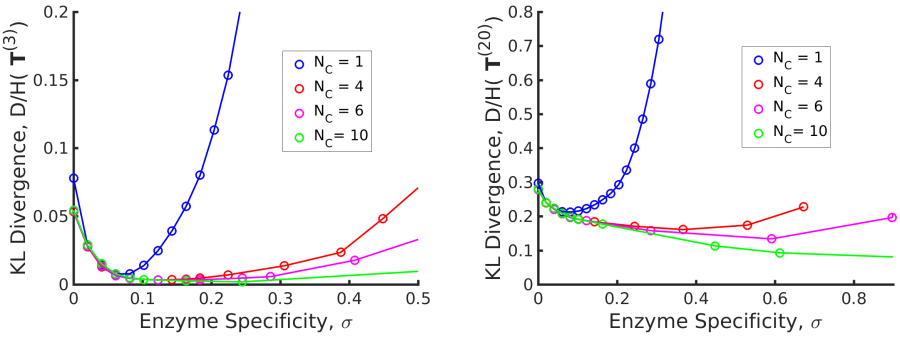
B. Tradeoffs between number of enzymes, number of cisternae and enzyme specificity to achieve given complexity

We are now in a position to catalogue the main results that follow from an optimization of the parameters of the glycan synthesis machinery to a given target distribution, Figs. 3-4

1. The normalized KL-distance $\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)$ is a convex function of σ for fixed values for other parameters (Fig. 3), i.e. it first decreases with σ and then increases beyond a critical value of $\sigma_{\min}$. $\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)$ is decreasing in N_C and N_E for fixed values of the other parameters, and increasing in the complexity of $\mathbf{c}^*$ for fixed (σ, N_C). The marginal contribution of N_C and N_E in reducing the normalised distance $\bar{D}$ is approximately equal (Figs. 4a, 4b). The lower complexity distributions can be synthesized with high fidelity with small (N_E, N_C), whereas higher complexity distributions require significantly larger (N_E, N_C), Figs. 4a, 4b. For a typical mammalian cell, the number of enzymes in the N-glycosylation pathway are in the range $N_E = 10{-}20$ [23–25, 27], Fig. 4b would then suggest that the optimal cisternal number would range from $N_C = 3 - 8$ [18].

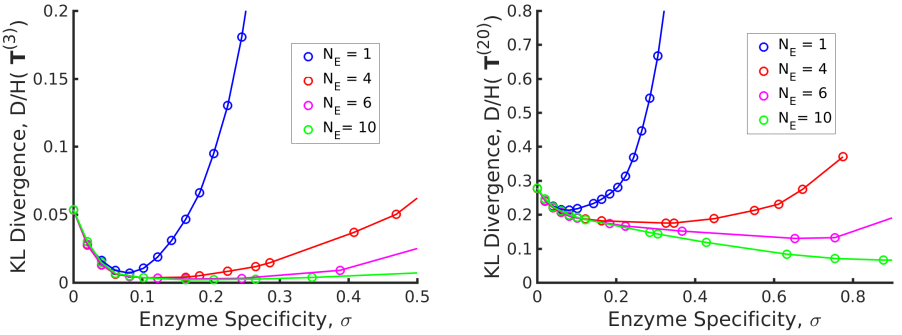
2. The optimal enzyme specificity $\sigma_{\min}(\mathbf{c}^*, N_C) = \text{argmin}_\sigma\{\bar{D}(\sigma, \bar{N}_E, N_C, \mathbf{c}^*)\}$, that minimises the error as function of $(N_C, \mathbf{c}^*)$ with N_E fixed at $\bar{N}_E$, is an increasing function of N_C and the complexity of the target distribution $\mathbf{c}^*$ (Figs. 3a, 3b, 4c, 4d). This is consistent with the results in Appendix E where we established that the width of the synthesized distribution is inversely dependent on the specificity σ: since a GMM approximation with fewer peaks has wider peaks, $\sigma_{\min}$ is low, and vice versa. Similar results hold when N_C is fixed at $\bar{N}_C$, and N_E is varied (Figs. 3c, 3d, 4c, 4d).

3. Let $\sigma_{\min}(N_C, N_E, \mathbf{c}^*)$ denote the value of σ that minimizes $\bar{D}(\sigma, N_C, N_E, \mathbf{c}^*)$. Then the second-derivative $\nabla^2_{\sigma_{\min}} \bar{D}(N_C, N_E, \mathbf{c}^*) = \frac{d^2}{d\sigma^2} \bar{D}(\sigma, N_C, N_E, \mathbf{c}^*) \mid_{\sigma=\sigma_{\min}}$ denotes the curvature at $\sigma_{\min}$, and is measure of the sensitivity of $\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)$ to σ for values close to $\sigma_{\min}(N_E, N_C, \mathbf{c}^*)$. $\nabla^2_{\sigma_{\min}} \bar{D}(N_C, N_E, \mathbf{c}^*)$ is a decreasing function of N_C (resp. N_E) for fixed values of $(N_E, \mathbf{c}^*)$ (resp. $(N_C, \mathbf{c}^*)$), see Figs. 3, 4e, 4f. Thus, for any target distribution $\mathbf{c}^*$ there is a minimal value of (N_E, N_C) such that the target can be synthesized with high fidelity



(a) Less complex target, 3-GMM approximation

(b) More complex target, 20-GMM approximation



(c) Less complex target, 3-GMM approximation

(d) More complex target, 20-GMM approximation

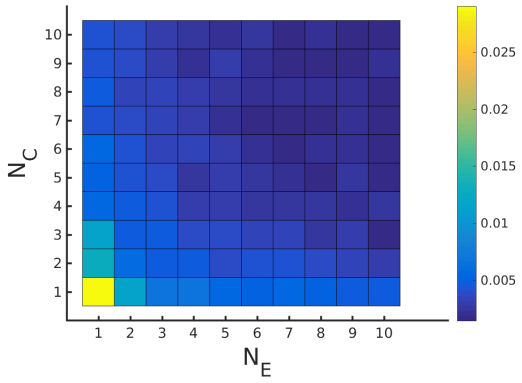
FIG. 3. Tradeoffs amongst the glycan synthesis parameters, enzyme specificity σ, cisternal number N_C and enzyme number N_E, to achieve a complex target distribution $\mathbf{c}^*$). (a)-(b) Normalised Kullback-Leibler distance $\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)$ as function of σ and N_C (for fixed $N_E = 3$), (c)-(d) $\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)$ as function of σ and N_E (for fixed $N_C = 3$), with the target distribution $\mathbf{c}^*$ set to the 3-GMM (less complex) and 20-GMM (more complex) approximations for the *human* T-cell MSMS data. $\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)$ is a convex function of σ for each $(N_E, N_C, \mathbf{c}^*)$, decreasing in N_C, N_E for each $(\sigma, \mathbf{c}^*)$, increasing in the complexity of $\mathbf{c}^*$ for fixed (σ, N_E, N_C). The specificity $\sigma_{\min}(\mathbf{c}^*, N_E, N_C) = \text{argmin}_\sigma\{\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)\}$ that minimises the error for given $(N_E, N_C, \mathbf{c}^*)$ is an increasing function of N_C, N_E and the complexity of the target distribution $\mathbf{c}^*$. Furthermore, the curvature of $\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)$ at $\sigma_{\min}(N_E, N_C, \mathbf{c}^*)$, related to *sensitivity*, is a decreasing function of N_C, N_E.

provided the sensitivity σ is tightly controlled at $\sigma_{\min}(N_C, N_E, \mathbf{c}^*)$, and there is larger value of (N_E, N_C) such that the target can be synthesized even if the control on σ is less tight.

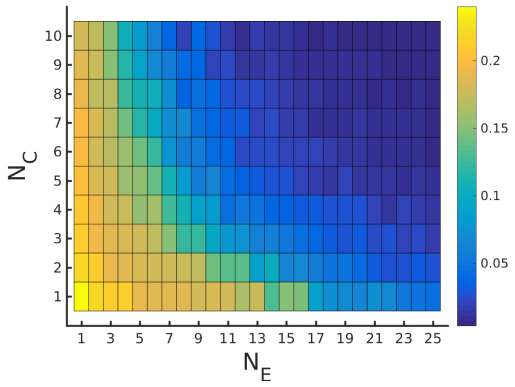
Ungar et al. [26] optimize incoming glycan ratio, transport rate and effective reaction rates in order to synthesize a narrow target distribution centred around a desired glycan. The ability to produce specific glycans without much heterogeneity is an important goal in pharma industry. They define heterogeneity as the total number of glycans synthesized, and show that increasing the number of compartments N_C decreases heterogeneity, and increases the concentration of the specific glycan. They also show that changing transport rate does not affect the heterogeneity. Our results are entirely consistent with theirs - we have shown that $\bar{D}$ decreases as we increase N_C. Thus, if the target distribution has a single sharp peak, increasing N_C will reduce the heterogeneity in the distribution.

C. Optimal partitioning of enzymes in cisternae

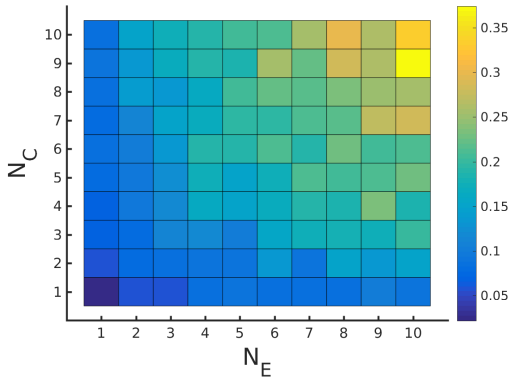
Having studied the optimum N_E, N_C, σ to attain a given target distribution with high fidelity, we ask what is the optimal partitioning of the N_E enzymes in these N_C cisternae? Answering this within our chemical reaction model (Sect. IV A) requires some care, since it incorporates the following enzymatic features: (a) enzymes with a finite specificity σ can catalyse several reactions, although with an efficiency that varies with both the substrate index k and cisternal index j, and (b) every enzyme appears in each cisternae; however their reaction efficiencies depend on



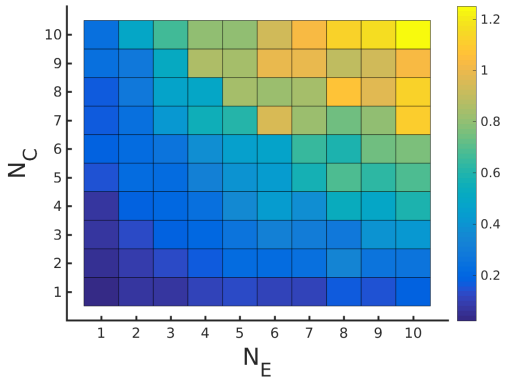
(a) Fidelity for less complex target, $\mathbf{c}^* = 3$-GMM approximation



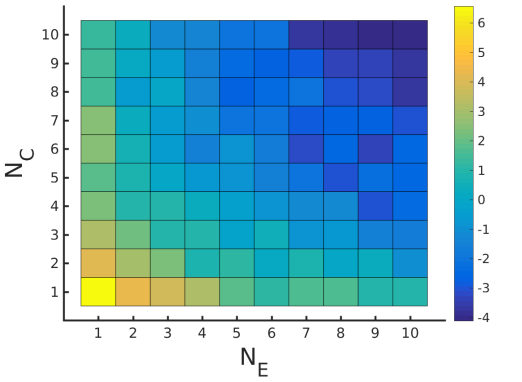
(b) Fidelity for more complex target $\mathbf{c}^* = 20$-GMM approximation



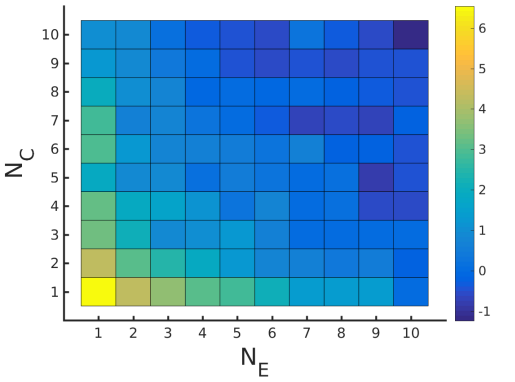
(c) Optimal enzyme specificity for less complex target, $\mathbf{c}^* = 3$-GMM approximation



(d) Optimal enzyme specificity for more complex target $\mathbf{c}^* = 20$-GMM approximation

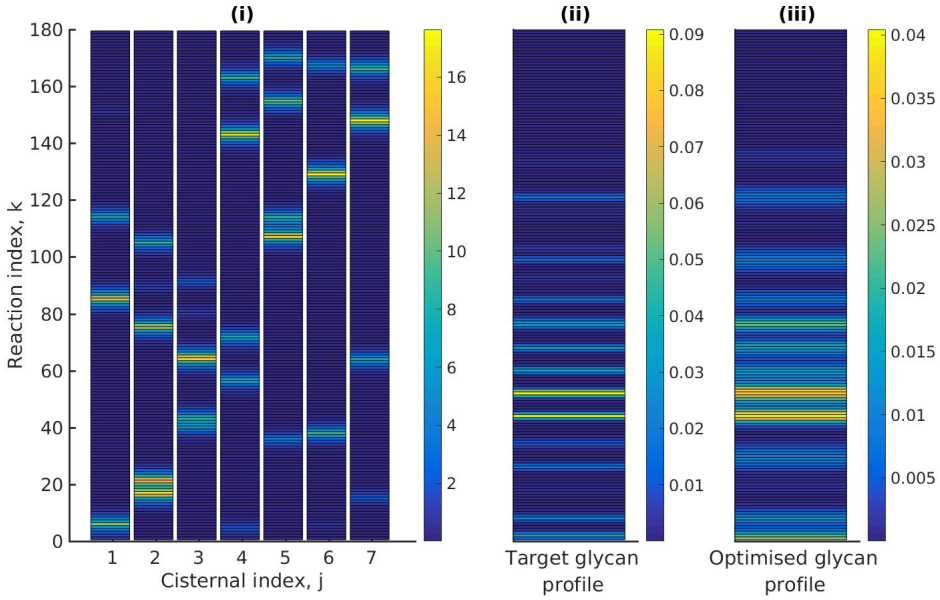


(e) Sensitivity for less complex target, $\mathbf{c}^* = 3$-GMM approximation

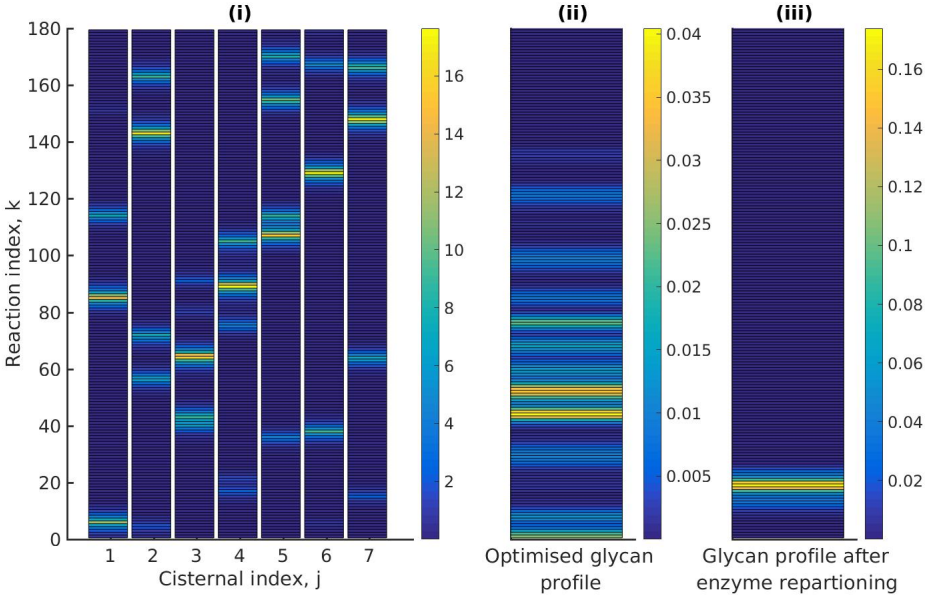


(f) Sensitivity for more complex target $\mathbf{c}^* = 20$-GMM approximation

FIG. 4. Fidelity of glycan distribution and optimal enzyme properties to achieve a complex target distribution. The target $\mathbf{c}^*$ is taken from 3-GMM (less complex) and 20-GMM (more complex) approximations of the *human* T-cell MSMS data. (a)-(b) Minimum normalised KL divergence $\min_\sigma\{\bar{D}(\sigma, N_C, N_E, \mathbf{c}^*)\}$ as a function of (N_E, N_C). More complex distributions require either a larger value N_E or N_C. The marginal impact of increasing N_E and N_C on $\bar{D}$ is approximately equal. (c)-(d) Optimum enzyme specificity $\sigma_{\min}$ obtained from $\min_\sigma\{\bar{D}(\sigma, N_C, N_E, \mathbf{c}^*)\}$ as a function of (N_E, N_C). $\sigma_{\min}$ increases with increasing N_E or N_C. To synthesize the more complex 20 GMM approximation with high fidelity requires enzymes with higher specificity $\sigma_{\min}$ compared to those needed to synthesize the broader, less complex 3-GMM approximation. (e) -(f) Sensitivity $\ln \frac{d^2}{d\sigma^2}\bar{D}\Big|_{\sigma_{\min}}$ of the normalised Kullback-Leibler distance $\bar{D}(\sigma, N_C, N_E, \mathbf{c}^*)$ as a function of (N_E, N_C). Increasing N_E or N_C decreases this sensitivity implying the specificity σ does not need to be tuned very carefully if N_E, N_C are high.



(a)

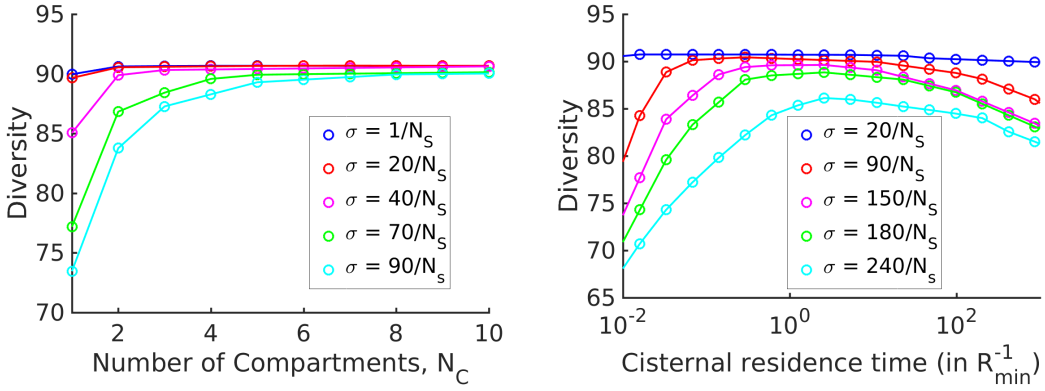


(b)

FIG. 5. Optimal enzyme partitioning in cisternae. (a) Heat map of the effective reaction rates in each cisterna (representing the optimal enzyme partitioning) and the steady state concentration in the last compartment ($\mathbf{c}^{(N_C)}$) for the 20-GMM target distribution. Here $N_E = 5$, $N_C = 7$, normalised $D_{KL}(\mathbf{T}^{(20)}\|\mathbf{c}^{(N_C)})/H(\mathbf{T}^{(20)}) = 0.11$. (b) Effective Reaction rates after swapping the optimal enzymes of the fourth and second cisternae. The displayed glycan profile is considerably altered from the original profile.

the enzyme levels, the enzymatic reaction rates and the enzyme matching function $\mathbf{L}$, all of which depend on the cisternal index j.

Thus, rather than determining the cisternal partitioning of enzymes, we instead identify chemical reactions that occur with high propensity in each cisternae. For this we define an effective reaction rate $\bar{R}(j,k)$ for $\mathcal{P}c_k \to \mathcal{P}c_{k+1}$ in the



(a) (b)

FIG. 6. Strategies for achieving high glycan diversity. Diversity versus N_C and transport rate μ at various values of specificity σ for fixed $N_E = 3$. (a) Diversity vs. N_C at optimal transport rate μ. Diversity initially increases with N_C, but eventually levels off. The levelling off starts at a higher N_C when σ is increased. These curves are bounded by the $\sigma = 0$ curve. (b) Diversity vs. cisternal residence time (μ^{-1}) in units of the reaction time ($R_{\min}^{-1}$) at various value of σ, for fixed $N_C = 4$ and $N_E = 10$. This has implications for glycoengineering (see text) where the task is to produce a particular glycan profile with low heterogeneity [26, 46].

j-th cisterna as

$$\bar{R}(j,k) = \sum_{\alpha=1}^{N_E} R_\alpha^{(j)} P^{(j)}(k,\alpha). \tag{10}$$

According to our model presented in Sect. IV A, the list of reactions with high effective reaction rates in each cisterna, corresponds to a cisternal partitioning of the perfect enzymes. In a future study, we will consider a Boolean version of a more complex chemical model, to address more clearly, the optimal enzyme partitioning amongst cisternae.

Figure 5 (a) (i) shows the heat map of the effective reaction rates in each cisterna for the optimal N_E, N_C, σ that minimises the normalised KL-distance to the 20 GMM target distribution $\mathbf{T}^{(20)}$ (see Fig. 5 (a) (ii)). The optimized glycan profile displayed in Fig. 5 (a) (iii) is very close to the target. An interesting observation from Fig. 5a(i) is that the same reaction can occur in multiple cisternae.

Keeping everything else fixed at the optimal value, we ask whether simply repartitioning the optimal enzymes amongst the cisternae, alters the displayed glycan distribution. In Fig. 5 (b) (i), we have exchanged the enzymes of the fourth and second cisterna. The glycan profile after enzyme partitioning (see Fig. 5 (b) ((iii)) is now completely altered (compare Fig. 5 (b) (ii) with Fig. 5 (b) (iii)). Thus one may achieve a different glycan distribution by repartitioning enzymes amongst the same number of cisternae [46].

VII. STRATEGIES TO ACHIEVE HIGH GLYCAN DIVERSITY

So far we have studied how the complexity of the target glycan distribution places constraints on the evolution of Golgi cisternal number and enzyme specificity. We now take up another issue, namely, how the physical properties of the Golgi cisternae, namely cisternal number and inter-cisternal transport rate, may drive diversification of glycans [48, 49]. There is substantial correlative evidence to support the idea that cell types that carry out extensive glycan processing employ larger numbers of Golgi cisternae. For example, the salivary Brunners gland cells secrete mucous that contains heavily O-glycosylated mucin as its major component [50]. The Golgi complex in these specialized cells contain $9-11$ cisternae per stack. Additionally, several organisms such as plants and algae secrete a rather diverse repertoire of large, complex glycosylated proteins, for a variety of functions [51–60]. These organisms possess enlarged Golgi complexes with multiple cisternae per stack [61–65].

In this section, we study how changing the physical parameters in our chemical synthesis model can lead to changes in the diversity of glycan distributions.

We define *diversity* as the total number of glycan species produced above a specified threshold abundance c_{th}. This last condition is necessary because very small peaks will not be distinguishable in the presence of noise. In computing the diversity from our chemical synthesis model, we have chosen the threshold to be $c_{th} = 1/N_s$, where N_s is the total number of glycan species. We have checked that the qualitative results do not depend on this choice, Fig. A6.

Using the sigmoid function $(1 + e^{-x/\tau})^{-1}$ as a continuous approximation to the Heaviside function $\Theta(x)$, we define the following optimization to achieve the maximal diversity for a given set of parameter values, N_E, N_C, σ,

$$\begin{aligned} \text{Diversity}(\sigma, N_C, N_E) := \quad & \max_{\boldsymbol{\mu},\mathbf{R},\mathbf{L}} \ \sum_{i=1}^{N_s} \left(1 + e^{-N_s(c_i - c_{th})}\right)^{-1} \\ & \text{s.t.} \ \ R_{\min} \le R_\alpha^{(j)} \le R_{\max}, \\ & \quad \mu_{\min} \le \mu^{(j)} \le \mu_{\max}, \end{aligned}$$

where, as before, $(\mu_{\max}, \mu_{\min}) = (1, 0.01)/\text{min}$, and $(R_{\max}, R_{\min}) = (20, 0.018)/\text{min}$, and $c_{th} = 1/N_s$ is the threshold. See Appendix F for details on the parameter estimation.

The results displayed in Fig. 6(a), show that for a fixed specificity σ, the diversity at first increases with the number of cisternae N_C, and then saturates at a value that depends on σ. For very high specificity enzymes, one can achieve very high diversity by appropriately increasing N_C. This establishes the link between glycan diversity and cisternal number. However, this link is correlative at best, since there are many ways to achieve high glycan diversity - notably by increasing the number of enzymes.

On the other hand, one of the goals of glycoengineering is to produce a particular glycan profile with low heterogeneity [26, 46]. For low specificity enzymes, the diversity remains unchanged upon increasing the cisternal residence time. For enzymes with high specificity, the diversity typically shows a non-monotonic variation with the cisternal residence time. At small cisternal residence time, the diversity decreases from the peak because of early exit of incomplete oligomers. At large cisternal residence time the diversity again decreases as more reactions are taken to completion. Note that the peak is generally very flat, this is consistent with the results of [26]. To get a sharper peak, as advocated for instance by [46], one might need to increase the number of high specificity enzymes N_E further.

VIII. DISCUSSION

The precision of the stereochemistry and enzymatic kinetics of these N-glycosylation reactions [2], has inspired a number of mathematical models [23–25] that predict the N-glycan distribution based on the activities and levels of processing enzymes distributed in the Golgi-cisternae of mammalian cells, and compare these predictions with N-glycan mass spectrum data. Models such as the KB2005 model [23–25] are extremely elaborate (with a network of 22, 871 chemical reactions and 7565 oligosaccharide structures) and require many chemical input parameters. These models have an important practical role to play, that of being able to predict the impact of the various *chemical parameters* on the glycan distribution, and to evaluate appropriate metabolic strategies to recover the original glycoprofile. Additionally, a recent study by Ungar and coworkers [26, 27] shows how *physical parameters*, such as overall Golgi transit time and cisternal number, can be tuned to engineer a homogeneous glycan distribution. Overall, such models may help predict glycosylation patterns and direct glycoengineering projects to optimize glycoform distributions.

In this paper, we have been interested in the role of glycans as a marker or molecular code of cell identity [4, 7, 11]. In particular, we have studied one aspect of molecular coding, namely the *fidelity* of this glycan code generated by enzymatic and transport processes located in the secretory apparatus of the cell. This involves a method of analysis that draws on many different fields, and so it might be useful to provide a short summary of the assumptions, methods and results of the paper:

1. The distribution of glycans at the cell surface is a marker of *cell-type identity* [2, 4, 7, 11]. This glycan distribution can be very complex; it is believed that there is an evolutionary drive for having glycan distributions of high *complexity* arising from the following considerations,

   (a) Reliable cell type identification amongst a large set of different cell types in a complex organism, the preservation and diversification of "self-recognition" [5].

   (b) Consequence of pathogen-mediated selection pressures [2, 4, 6].

   (c) Consequence of *herd immunity* within a heterogenous population of cells of a community [15] or within a single organism [5].

2. The glycans at the cell surface are the end product of a sequential chemical processing via a set of enzymes resident in the Golgi cisternae, and transport across cisternae [4, 10, 11]. Using a fairly general and tractable model for chemical synthesis and transport, we compute the *synthesized* glycan distribution at the cell surface. Parameters of our synthesis model include the number of enzymes N_E, specificity of enzymes σ, number of cisternae N_C and transport rates μ.

3. We measure the reliability or fidelity of cell identity [10, 11, 66] in terms of the error between synthesized glycan distribution on the cell surface from the its internal "target" distribution using the Kullback- Leibler distance D_{KL} [16, 17]. In our numerical study, we obtain the *target distribution* for the given cell type by suitable coarse-graining of the MSMS data for the *human* T-cells [22]. We solve a constrained optimization problem for minimising D_{KL}, and study the tradeoffs between N_E, N_C and σ.

4. The results of the optimization to achieve a given target complexity are summarised in Figs. 3-4. Here, we highlight some its direct consequences:

   (a) Keeping the number of enzymes fixed, a more elaborate transport mechanism (via control of N_C and μ) is essential for synthesising high complexity target distributions (Figs. 4a, 4b). Fewer cisternae cannot be compensated for by optimising the enzymatic synthesis (via control of parameters $\mathbf{R}, \mathbf{L}$ and σ).

   (b) Thus, our study suggests that fidelity of the glycan code generation provides a functional control of Golgi cisternal number. It also provides a quantitative argument for the evolutionary requirement of multiple-compartments, by demonstrating that the fidelity and sensitivity of the glycan code arising from a chemical synthesis that involves multiple cisternae is higher than one that involves a single cisterna (keeping everything else fixed) (Figs. 4a, 4b, 4e, 4f). This feature that with multiple cisternae and precise enzyme partitioning, one may generically achieve a highly accurate representation of the target distribution, has been highlighted in an algorithmic model of glycan synthesis [46].

   (c) Our study shows that for a fixed N_C and N_E, there is an optimal enzyme specificity that achieves the lowest distance from a given target distribution. As we see in Fig. 4d, this optimal enzyme specificity can be very high for highly complex target distributions.

   (d) Organisms such as plants and algae, have a diverse repertoire of glycans that are utilised in a variety of functions [51–60]. Our study shows that it is optimal to use low specificity enzymes to synthesize target distributions with high diversity (Fig. 6). However, this compromises on the complexity of the glycan distribution, revealing a tension between complexity and diversity. One way of relieving this tension is to have larger N_E and N_C.

   (e) Consider a situation where the environment, and hence the target glycan distribution, fluctuates rapidly. When synthesis parameters cannot change rapidly enough to track the environment, high specificity enzymes can lead to a *lowering* of the cell's fitness [67, 68]. Having slightly sloppy enzymes may give the best selective advantage in a time varying environment. This compromise, between robustness in a changing environment and the demand for complexity, is achieved by having sloppy enzymes, that allows the system to be more *evolvable* [67, 68]. However, sloppy enzymes are subject to errors from synthesising the wrong reaction products. In this case, error correcting mechanisms must be in place to ensure fidelity of the glycan code. We leave the role of intra-cellular transport in providing non-equilibrium proof-reading mechanisms to reduce such coding errors, and its optimal adaptive strategies and plasticity in a time varying environment, as a task for the future.

Admittedly the chemical network that we have considered here is much simpler than the chemical network associated with all possible protein modifications in the secretory pathway. For instance, typical N-glycosylation pathways would involve the glycosylation of a variety of GBPs. Further, apart from N-glycosylation, there are other glycoprotein, proteoglycan and glycolipid synthesis pathways [1, 2, 11]. We believe our analysis is generalisable and that the qualitative results we have arrived at would still hold.

To conclude, our work establishes the link between the cisternal machinery (chemical and transport) and optimal coding. We find that the pressure to achieve the target glycan code for a given cell type, places strong constraints on

the cisternal number and enzyme specificity [18]. An important implication is that a description of the nonequilibrium self-assembly of a fixed number of Golgi cisternae must combine the dynamics of chemical processing and membrane dynamics involving fission, fusion and transport [18–20]. We believe this is a promising direction for future research.

IX. ACKNOWLEDGMENTS

We thank M. Thattai, A. Jaiman, S. Ramaswamy, A. Varki for discussions, and S. Krishna and R. Bhat for very useful suggestions on the manuscript. We thank our group members at the Simons Centre for many incisive inputs. We are very grateful to P. Babu for consultations on the MSMS data and literature. We acknowledge the computational facilities at NCBS. MR acknowledges a JC Bose Fellowship from DST (Government of India), and thanks Institut Curie for hosting a visit under the Labex program. This work has received support under the program Investissements dAvenir launched by the French Government and implemented by ANR with the references ANR-10-LABX-0038 and ANR-10-IDEX-0001-02 PSL. QV thanks the Simons Centre (NCBS) for hosting his visit.

Appendix

Appendix A: Kinetics of sequential chemical reactions and transport

On including the rates of injection (q) and inter-cisternal transport $\mu^{(j)}$ from the cisterna j to $j+1$, into the chemical reaction kinetics, the substrate concentrations $c_k^{(j)}$ change with time as,

$$\begin{aligned}
\frac{dc_1^{(1)}}{dt} &= q - \sum_{\alpha=1}^{N_E} \frac{V(1,1,\alpha)P^{(1)}(1,\alpha)c_1^{(1)}}{M(1,1,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(1)}(k',\alpha)c_{k'}^{(1)}}{M(1,k',\alpha)}\right)} - \mu^{(1)}c_1^{(1)} \\
\frac{dc_k^{(1)}}{dt} &= \sum_{\alpha=1}^{N_E} \frac{V(1,k-1,\alpha)P^{(1)}(k-1,\alpha)c_{k-1}^{(1)}}{M(1,k-1,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(1)}(k',\alpha)c_{k'}^{(1)}}{M(1,k',\alpha)}\right)} \\
&\quad - \sum_{\alpha=1}^{N_E} \frac{V(1,k,\alpha)P^{(1)}(k,\alpha)c_k^{(1)}}{M(1,k,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(1)}(k',\alpha)c_{k'}^{(1)}}{M(1,k',\alpha)}\right)} - \mu^{(1)}c_k^{(1)} \\
\frac{dc_{N_s}^{(1)}}{dt} &= \sum_{\alpha=1}^{N_E} \frac{V(1,N_s-1,\alpha)P^{(1)}(N_s-1,\alpha)c_{N_s-1}^{(1)}}{M(1,N_s-1,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(1)}(k',\alpha)c_{k'}^{(1)}}{M(1,k',\alpha)}\right)} - \mu^{(1)}c_{N_s}^{(1)}
\end{aligned} \tag{A1}$$

for cisterna-1, and

$$\begin{aligned}
\frac{dc_1^{(j)}}{dt} &= \mu^{(j-1)}c_1^{(j-1)} - \sum_{\alpha=1}^{N_E} \frac{V(j,1,\alpha)P^{(j)}(1,\alpha)c_1^{(j)}}{M(j,1,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)} - \mu^{(j)}c_1^{(j)} \\
\frac{dc_k^{(j)}}{dt} &= \mu^{(j-1)}c_k^{(j-1)} + \sum_{\alpha=1}^{N_E} \frac{V(j,k-1,\alpha)P^{(j)}(k-1,\alpha)c_{k-1}^{(j)}}{M(j,k-1,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)} \\
&\quad - \sum_{\alpha=1}^{N_E} \frac{V(j,k,\alpha)P^{(j)}(k,\alpha)c_k^{(j)}}{M(j,k,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)} - \mu^{(j)}c_k^{(j)} \\
\frac{dc_{N_s}^{(j)}}{dt} &= \mu^{(j-1)}c_{N_s}^{(j-1)} + \sum_{\alpha=1}^{N_E} \frac{V(j,N_s-1,\alpha)P^{(j)}(N_s-1,\alpha)c_{N_s-1}^{(j)}}{M(j,N_s-1,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)} - \mu^{(j)}c_{N_s}^{(j)}
\end{aligned} \tag{A2}$$

for cisternae $j = 2, 3, \ldots, N_C$. These set of dynamical equations (A1)-(A2), with initial conditions, can be solved to obtain the concentration $c_k^{(j)}(t)$ for $t \geq 0$.

Equations (A1)-(A2) automatically obey the conservation law for the protein concentration (p), i.e., denoting the protein concentration in the j-th cisterna as $p^{(j)} = \sum_{k'=1}^{N_s} c_{k'}^{(j)}$, we automatically obtain,

$$\begin{aligned}
\frac{dp^{(1)}}{dt} &= q - \mu^{(1)}p^{(1)} \\
\frac{dp^{(j)}}{dt} &= \mu^{(j-1)}p^{(j-1)} - \mu^{(j)}p^{(j)}
\end{aligned}$$

for $j = 2, 3, \ldots N_C$.

At steady state, the left hand side of the above equations is set to zero, which after rescaling, gives the nonlinear recursion relations displayed in (5) and (6) of the main text.

Appendix B: A computationally simpler optimization equivalent to Optimization A

Define a new set of parameters,

$$R(j,k,\alpha)=\sum_{\alpha=1}^{N_E}\frac{V(j,k,\alpha)}{M(j,k,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)} \tag{B1}$$

where $\mathbf{c}$ denotes the steady state glycan concentration, corresponding to a specific $(\mathbf{M},\mathbf{V},\mathbf{L})$. Define $\mathbf{v}$ by the following set of linear equations:

$$\begin{aligned}
v_1^{(1)} =& \frac{1}{\mu^{(1)}+\sum_{\alpha=1}^{N_E}R(1,1,\alpha)P^{(1)}(1,\alpha)}\\
v_k^{(1)} =& \frac{v_{k-1}^{(1)}\sum_{\alpha=1}^{N_E}R(1,k-1,\alpha)P^{(1)}(k-1,\alpha)}{\mu^{(1)}+\sum_{\alpha=1}^{N_E}R(1,k,\alpha)P^{(1)}(k,\alpha)}\\
v_{N_s}^{(1)} =& \frac{v_{N_s-1}^{(1)}\sum_{\alpha=1}^{N_E}R(1,N_s-1,\alpha)P^{(1)}(N_s-1,\alpha)}{\mu^{(1)}}
\end{aligned} \tag{B2}$$

for $j=1$, and

$$\begin{aligned}
v_1^{(j)} =& \frac{v_1^{(j-1)}\mu^{(j-1)}}{\mu^{(j)}+\sum_{\alpha=1}^{N_E}R(j,1,\alpha)P^{(j)}(1,\alpha)}\\
v_k^{(j)} =& \frac{v_k^{(j-1)}\mu^{(j-1)}}{\mu^{(j)}+\sum_{\alpha=1}^{N_E}R(j,k,\alpha)P^{(j)}(k,\alpha)}\\
&+\frac{v_{k-1}^{(j)}\sum_{\alpha=1}^{N_E}R(j,k-1,\alpha)P^{(j)}(k-1,\alpha)}{\mu^{(j)}+\sum_{\alpha=1}^{N_E}R(j,k,\alpha)P^{(j)}(k,\alpha)}\\
v_{N_s}^{(j)} =& \frac{v_{N_s}^{(j-1)}\sum_{\alpha=1}^{N_E}R(j,N_s-1,\alpha)P^{(j)}(N_s-1,\alpha)}{\mu^{(j)}}+\frac{v_{N_s}^{(j-1)}\mu^{(j-1)}}{\mu^{(j)}}
\end{aligned} \tag{B3}$$

for $j=2,\ldots,N_C$. Then, by the definition of $\mathbf{R}$ in (B1), it trivially follows that the steady state concentration $\mathbf{c}$ corresponding to $(\mathbf{M},\mathbf{V},\mathbf{L})$ is a solution for (B2)-(B3).

In Appendix C we show that for $\mathbf{v}$ obtained from (B2)-(B3) for any parameter $(\mathbf{R},\mathbf{L})$, there exists parameter $(\mathbf{M},\mathbf{V},\mathbf{L})$ such that (5)-(6) are automatically satisfied when we set $\mathbf{c}=\mathbf{v}$, i.e. $\mathbf{v}$ is the steady state concentration for $(\mathbf{M},\mathbf{V},\mathbf{L})$. Thus, the set of all concentration profiles defined by (B2)-(B3) as a function of all possible values of the parameters $(\mathbf{R},\mathbf{L})$ is identical to the set defined by (5)-(6) as function of $(\mathbf{M},\mathbf{V},\mathbf{L})$. This is a crucial insight, since it allows us to search the entire parameter space using (B2)-(B3), where the concentration is known explicitly in terms of $(\mathbf{R},\mathbf{L})$. See Figure A1 for a flow chart of the two optimization schemes.

To pose this new optimization problem, it is convenient to define $\bar{v}_i=\mu^{(N_c)}v_i^{(N_c)}$. Then, it follows that

Optimization B:

$$D(\sigma,N_E,N_C,\mathbf{c}^*):=\min_{\mathbf{R}\geq\mathbf{0},\ \mathbf{L}}D_{KL}(\mathbf{c}^*\|\bar{\mathbf{v}}) \tag{B4}$$

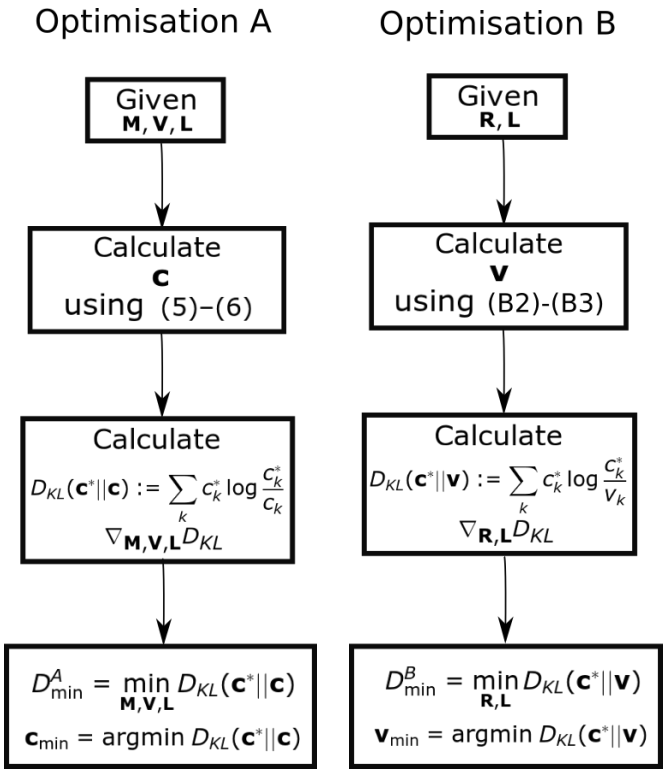


FIG. A1. Flow chart showing the optimization schemes for Optimization A and B. We prove that $D^A_{\min} = D^B_{\min}$ by showing the set of all $\mathbf{c}$ is equal to the set of all $\mathbf{v}$. We additionally establish that the optimum $\mathbf{v}_{\min} = \mathbf{c}_{\min}$.

is equivalent to (8). Since $\mathbf{v}$ is explicitly known as a function of $(\mathbf{R}, \mathbf{L})$, optimization B (B4) is a more tractable optimization problem than (8). Note that in this setting, the function $D(\sigma, N_E, N_C, \mathbf{c}^*)$ (B4) is independent of the rates $\boldsymbol{\mu}$.

While this optimization is easy to implement, we note that the parameters (e.g., reaction rates, specificity) are not constrained to take only physically relevant values; a legitimate concern is that the absence of such physico-chemical constraints might drive this optimization to physically unrealistic solutions.

There are two possible ways to impose these parameter constraints. One is to impose constraints on the "microscopic" chemical parameters, such as the rate of individual reactions $R(j,k,\alpha)$ and the inter-cisternal transport rate $\mu^{(j)}$. These take into consideration constraints arising from molecular enzymatic processes. The other is to impose constraints on "global" physical parameters, such as the total transport time across the Golgi cisternae and the average enzymatic reaction time. Here, we impose constraints on the microscopic reaction and transport parameters.

Optimization C :

$$\begin{aligned} D(\sigma, N_C, N_E, \mathbf{c}^*) := \quad & \min_{\boldsymbol{\mu},\mathbf{R},\mathbf{L}} \; D_{KL}(\mathbf{c}^* \| \bar{\mathbf{v}}) \\ & \text{s.t.} \; R_{\min} \le R(j,k,\alpha) \le R_{\max}, \\ & \quad \mu_{\min} \le \mu^{(j)} \le \mu_{\max}. \end{aligned}$$

The upper and lower bounds on the rates $\mathbf{R}$ and $\boldsymbol{\mu}$ are estimated in Appendix F : $\mu_{\max} = 1/\text{min}$ (resp. $\mu_{\min} = .01/\text{min}$) and $R_{\max} = 20/\text{min}$ (resp. $R_{\min} = .018/\text{min}$).

Appendix C: Equivalence of Optimizations A and B

Let

$$\mathcal{A} = \left\{ [c_k^{(j)}]_{j,k} : \begin{array}{l} M(j,k,\alpha) \geq 0, V(j,k,\alpha) \geq 0, l_\alpha^{(j)} \geq 0, \\ [c_k^{(j)}]_{jk} \text{ given by (5) and (6)} \end{array} \right\}$$

denote the set of concentrations achievable in Optimization A. Similarly, let

$$\mathcal{B} = \left\{ [v_k^{(j)}]_{j,k} : \begin{array}{l} R(j,k,\alpha) \geq 0, l_\alpha^{(j)} \geq 0 \\ [v_k^{(j)}]_{j,k} \text{ given by (B2) and (B3)} \end{array} \right\}$$

denote the set of concentrations achievable in the Optimization B.

Our task is to show that $\mathcal{A} = \mathcal{B}$. Suppose $[c_k^{(j)}]_{j,k} \in \mathcal{A}$. Let $[M(j,k,\alpha)]$, $[V(j,k,\alpha)]$ and $[l_\alpha^{(j)}]$ be the corresponding parameters. Define

$$R(j,k,\alpha) = \sum_{\alpha=1}^{N_E} \frac{V(j,k,\alpha)}{M(j,k,\alpha)\left(1 + \sum_{k'=1}^{N_s} \frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)} \geq 0$$

Then $[c_k^{(j)}]_{j,k} \in \mathcal{B}$.

Suppose $[v_k^{(j)}]_{j,k} \in \mathcal{B}$. Let $[R(j,k,\alpha)]$, $[l_\alpha^{(j)}]$ denote the corresponding parameters. Since $\sum_{k=1}^{N_s} v_k^{(j)} = 1/\mu^{(j)} < \infty$, it follows that $\sum_{k=1}^{N_s} P^{(j)}(k,\alpha)v_k^{(j)} < 1/\mu^{(j)} < \infty$. Thus, there exists parameters $[M(j,k,\alpha)]$, $[V(j,k,\alpha)]$ and $[l_\alpha^{(j)}]$ such that

$$R(j,k,\alpha) = \sum_{\alpha=1}^{N_E} \frac{V(j,k,\alpha)}{M(j,k,\alpha)\left(1 + \sum_{k'=1}^{N_s} \frac{P^{(j)}(k',\alpha)v_{k'}^{(j)}}{M(j,k',\alpha)}\right)} \tag{C1}$$

Therefore, $[v_k^{(j)}]_{j,k}$ satisfy (5) and (6), i.e. $[v_k^{(j)}]_{j,k} \in \mathcal{A}$.

Moreover, suppose $\mathbf{v}$ satisfies (B2)-(B3) for a given set of parameters $(\mathbf{R}, \mathbf{L})$. Then there exist $(\mathbf{M}, \mathbf{V}, \mathbf{L})$ such that $\mathbf{v}$ satisfies (5)-(6), i.e. $\mathbf{v}$ is the steady state concentration for $(\mathbf{M}, \mathbf{V}, \mathbf{L})$.

Appendix D: Analytical solution when $N_s \gg 1$

It is possible to obtain analytical expressions for the steady state glycan distribution, in the limit $N_s \gg 1$, when the glycan index k can be approximated by a continuous variable. In this case, (5)-(6) can be cast as differential equations,

$$\begin{aligned} &\frac{dc_k^{(1)}}{dk} \approx c_k^{(1)} - c_{k-1}^{(1)} \\ &= \left(\frac{\sum_{\alpha=1}^{N_E} R(1,k-1,\alpha)\exp(-\sigma|k-1-l_\alpha^{(1)}|)}{\mu^{(1)} + \sum_{\alpha=1}^{N_E} R(1,k,\alpha)\exp(-\sigma|k-l_\alpha^{(1)}|)} - 1 \right) c_{k-1}^{(1)} \\ &\approx -\left(\frac{\mu^{(1)} + \frac{d}{dk}\sum_{\alpha=1}^{N_E} R(1,k,\alpha)\exp(-\sigma|k-l_\alpha^{(1)}|)}{\mu^{(1)} + \sum_{\alpha=1}^{N_E} R(1,k,\alpha)\exp(-\sigma|k-l_\alpha^{(1)}|)} \right) c_k^{(1)}, \end{aligned} \tag{D1}$$

and

$$\begin{aligned}\frac{dc_k^{(j)}}{dk} &\approx c_k^{(j)} - c_{k-1}^{(j)} \\ &= \frac{\mu^{(j-1)}}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)} c_k^{(j-1)} \\ &- \left(\frac{\mu^{(j)} + \frac{d}{dk}\sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)} \right) c_k^{(j)}\end{aligned} \tag{D2}$$

for $j = 2, \ldots, N_C$. In (D1) and (D2),

$$\begin{aligned}&\frac{d}{dk}\sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|) \\ &= \sum_{\alpha=1}^{N_E} R(j,k,\alpha)\sigma\exp(-\sigma|k - l_\alpha^{(j)}|)(1 - 2\mathbb{I}(k \geq l_\alpha)) + \; R'(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)\end{aligned} \tag{D3}$$

where the indicator function $\mathbb{I}(\cdot)$ is equal to 1 if the argument is true, and zero otherwise and $R'(j,k,\alpha)$ is the derivative of $R(j,k,\alpha)$ with respect to k.

Define a vector function $C(k) \in \mathbb{R}_c^N$ of the continuous variable k by $C(k) = [c_k^{(1)}, c_k^{(2)}, \ldots c_k^{(N_C)}]$. Then (D1) and (D2) can be written as:

$$\frac{dC(k)}{dk} = M(k)C(k) \tag{D4}$$

where the matrix $M(k)$ is given by

$$M(k) = \begin{bmatrix} A^{(1)}(k) & 0 & 0 & 0 & \ldots 0 \\ B^{(2)}(k) & A^{(2)}(k) & 0 & 0 & \ldots 0 \\ 0 & B^{(3)}(k) & A^{(3)}(k) & 0 & \ldots 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & B^{(N_C)}(k) & A^{(N_C)}(k) \end{bmatrix} \tag{D5}$$

with

$$\begin{aligned}A^{(j)}(k) &= -\frac{\mu^{(j)} + \frac{d}{dk}\sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)} \\ B^{(j)}(k) &= \frac{\mu^{(j-1)}}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)}\end{aligned}$$

The functions $A^{(j)}(k)$ and $B^{(j)}(k)$ involve absolute value and indicator functions; therefore, the differential equation has to be solved in a piecewise manner assuming continuity of solution $C(k)$.

The general solution of (D4)

$$C(k) = C_0 \exp\left(\Omega(k)\right) \tag{D6}$$

is written in terms of the Magnus Function $\Omega(k) = \sum_{n=1}^{\infty} \Omega(n,k)$, obtained from the Baker-Campbell-Hausdorff formula [69],

$$\begin{aligned}
\Omega(1,k) &= \int_0^k M(k_1)dk_1 \\
\Omega(2,k) &= \frac{1}{2}\int_0^k dk_1 \int_0^{k_1} dk_2\,[M(k_1), M(k_2)] \\
\Omega(3,k) &= \frac{1}{6}\int_0^k dk_1 \int_0^{k_1} dk_2 \int_0^{k_2} dk_3\,[M(k_1),[M(k_2),M(k_3)]] + [M(k_3),[M(k_2),M(k_1)]] \\
&\ldots\ldots
\end{aligned}$$

where $[M(k_1), M(k_2)] := M(k_1)M(k_2) - M(k_2)M(k_1)$ is the commutator, and the higher order terms in $\ldots$ contain higher order nested commutators.

Here, we establish conditions under which the series $\sum_{n=1}^{\infty} \Omega(n,k)$ that defines solution $C(k)$ to the differential equation (D4) converges. We also solve (D4) for some special cases.

The commutator

$$[M(k_1), M(k_2)] = \begin{bmatrix} 0 & 0 & 0 & 0 & \ldots & 0 \\ a_{21} & 0 & 0 & 0 & \ldots & 0 \\ a_{31} & a_{32} & 0 & 0 & \ldots & 0 \\ 0 & a_{42} & a_{43} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \ldots & & a_{n,n-2} & a_{n,n-1} & 0 \end{bmatrix}$$

where

$$\begin{aligned}
a_{i,i-1} &= A^{(i-1)}(k_2)B^{(i)}(k_1) + A^{(i)}(k_1)B^{(i)}(k_2) - A^{(i-1)}(k_1)B^{(i)}(k_2) + A^{(i)}(k_2)B^{(i)}(k_1) \\
a_{i,i-2} &= B^{(i-1)}(k_2)B^{(i)}(k_1) - B^{(i-1)}(k_1)B^{(i)}(k_2)
\end{aligned}$$

The general form of $\Omega(n,k)$ is given by [69]

$$\Omega(n,k) = \frac{z_n}{n!}\int_0^k dk_1 \int_0^{k_1} dk_2 \ldots \int_0^{k_{n-2}} dk_{n-1} \int_0^{k_{n-1}} dk_n \sum_l W_l M(k_{p_1^l}) M(k_{p_2^l}) \ldots M(k_{p_n^l}) \tag{D7}$$

where $(p_1^{(l)}, p_2^{(l)} \ldots p_n^{(l)})$ is a permutation of $(1,2,3,\ldots n)$, $W_l \in \{-1,1\}$, and $z_n \in 1,\ldots n$.

Let $\bar{A} = \max_{k,l,m} |M_{l,m}(k)|$. Define

$$\bar{M} = \begin{bmatrix} \bar{A} & 0 & 0 & 0\ldots & 0 \\ \bar{A} & \bar{A} & 0 & 0\ldots & 0 \\ 0 & \bar{A} & \bar{A} & 0\ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \ldots & 0 & \bar{A} & \bar{A} \end{bmatrix}$$

We can bound all the matrix elements of $\Omega(n,k)$ in the following way

$$\begin{aligned}
\Omega_{lm}(n,k) &\le z_n \bar{M}^n_{l,m} \int_0^k dk_1 \int_0^{k_1} dk_2 \ldots \int_0^{k_{n-1}} dk_n \\
&= z_n \bar{M}^n \Big|_{lm} \frac{k^n}{n!}
\end{aligned} \tag{D8}$$

The matrix

$$\bar{M}^n = \begin{bmatrix} a_{11} & 0 & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & 0 & 0 & \dots & 0 \\ a_{31} & a_{32} & a_{33} & 0 & \dots & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & \dots & & a_{n,n-2} & a_{n,n-1} & a_{nn} \end{bmatrix}$$

where $a_{lm} = S_{lm}(n)\bar{A}^n$ for appropriately defined polynomials $S_{l,m}(n)$. Thus, it follows that $\Omega_{lm} \leq z_n S_{lm}(n)(A^*)^n \frac{k^n}{n!}$ and $\Omega_{l,m}(k) \leq \sum_{n=1}^{\infty} z_n S_{l,m}(n)(A^*)^n \frac{k^n}{n!}$. Consequently, the series will converge if $\bar{A}k < 1$, i.e. $k \leq \frac{1}{\bar{A}}$. Assuming $\mu^{(j)} = \mu \; \forall j$, we can bound $\bar{A}$ as

$$\bar{A} \leq \max_{j,k} \left(\frac{\mu + \sigma \sum_{\alpha=1}^{N_E} R(j,k,\alpha) \exp(-\sigma|k - l_\alpha^{(j)}|)}{\mu + \sum_{\alpha=1}^{N_E} R(j,k,\alpha) \exp(-\sigma|k - l_\alpha^{(j)}|)} + \frac{\sum_{\alpha=1}^{N_E} R'(j,k,\alpha) \exp(-\sigma|k - l_\alpha^{(j)}|)}{\mu + \sum_{\alpha=1}^{N_E} R(j,k,\alpha) \exp(-\sigma|k - l_\alpha^{(j)}|)} \right) \tag{D9}$$

Since the parameters μ, σ, $R(j,k,\alpha)$, $l_\alpha^{(j)}$ and N_E are finite and positive, and $R'(j,k,\alpha)$ is finite, $\bar{A}$ has a finite upper bound, implying k is always greater than zero, and the series has finite radius of convergence.

While in principle we can obtain the glycan profile for any N_E and N_C with arbitrary accuracy, assuming $R(j,k,\alpha) = R_\alpha^{(j)}$, we provide explicit formulae for a few representative cases : (i) $(N_E = 1, N_C = 1)$ and (ii) $(N_E = 1, N_C = 2)$.

(i) $N_E = 1, N_C = 1$: The solution of the differential equation is given by

$$c(k) = \begin{cases} c_0 e^{-k} \left(\frac{\mu + R\exp(-\sigma(l-k))}{\mu + R\exp(-\sigma l)} \right)^{(1/\sigma)-1} & k \leq l \\ c(l) e^{-(k-l)} \left(\frac{\mu + R}{\mu + R\exp(-\sigma(k-l))} \right)^{(1/\sigma)+1} & k > l \end{cases} \tag{D10}$$

A representative concentration profile is plotted in Fig. A2(a). The concentration profile consists of two distinct components: an initial exponential decay, and then an exponential rise and fall concentrated around l. The relative weight of these two components is controlled by the sensitivity σ and the rate R. Such explicit formulae can be obtained for any $N_E > 1$, as long as $N_C = 1$.

(ii) $N_E = 1, N_C = 2$: The concentration profile $c^{(2)}$ in cisterna 2 can be obtained from the following calculation. Let $l^{(j)}$ denote the "length" of the enzyme in cisterna $j = 1, 2$. For $k \leq \min\{l^{(1)}, l^{(2)}\}$

$$\begin{aligned} c^{(2)}(k) = c_0 \mu^{(1)} e^{-k} \Big(\frac{\mu^{(2)} + R^{(2)} \exp(-\sigma(l^{(2)} - k))}{\mu^{(1)} + R^{(1)} e^{-\sigma l^{(1)}}} \Big)^{(1/\sigma)-1} \int_0^k \frac{(\mu^{(1)} + R^{(1)} \exp(-\sigma(l^{(1)} - k)))^{(1/\sigma)-1}}{(\mu^{(2)} + R^{(2)} \exp(-\sigma(l^{(2)} - k)))^{1/\sigma}} dk \\ + c^{(2)}(0) e^{-k} \left(\frac{\mu^{(2)} + R^{(2)} e^{-\sigma(l^{(2)} - k)}}{\mu^{(2)} + R^{(2)} e^{-\sigma l^{(2)}}} \right)^{(1/\sigma)-1} \end{aligned} \tag{D11}$$

Next, consider the case where $l^{(1)} \leq l^{(2)}$. Then, for $l^{(1)} < k \leq l^{(2)}$

$$\begin{aligned} c^{(2)}(k) = c^{(1)}(l^{(1)}) \mu^{(1)} e^{-(k - l^{(1)})} (\mu^{(1)} + R^{(1)})^{(1/\sigma)+1} (\mu^{(2)} + R^{(2)} \exp(-\sigma(l^{(2)} - k)))^{(1/\sigma)-1} \\ \int_{l^{(1)}}^k \frac{(\mu^{(2)} + R^{(2)} \exp(-\sigma(l^{(2)} - k)))^{-1/\sigma}}{(\mu^{(1)} + R^{(1)} \exp(-\sigma(k - l^{(1)})))^{(1/\sigma)+1}} dk \\ + c^{(2)}(l^{(1)}) e^{-(k - l^{(1)})} \left(\frac{\mu^{(2)} + R^{(2)} e^{-\sigma(l^{(2)} - k)}}{\mu^{(2)} + R^{(2)} e^{-\sigma(l^{(2)} - l^{(1)})}} \right)^{(1/\sigma)-1} \end{aligned} \tag{D12}$$

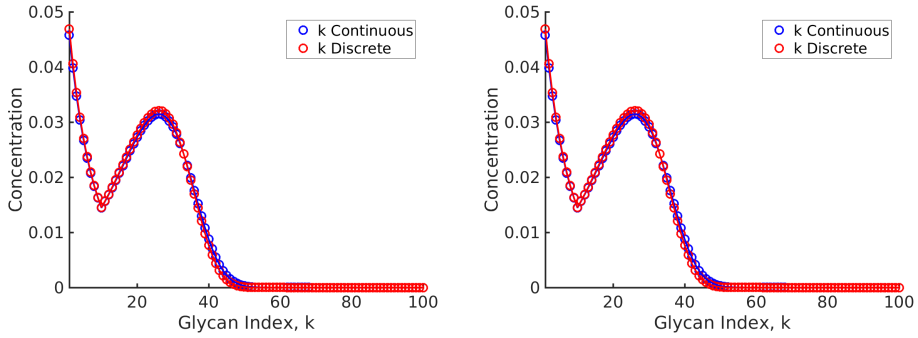


FIG. A2. Glycan concentration profile calculated from the model using (a) formula (D10) for $N_E = N_C = 1$ and (b) formulae (D11)-(D15) for $N_E = 1, N_C = 2$.

and for $l^{(1)} \leq l^{(2)} < k$,

$$c^{(2)}(k) = c^{(1)}(l^{(1)})\mu^{(1)}e^{-(k-l^{(1)})}\left(\frac{\mu^{(1)}+R^{(1)}}{\mu^{(2)}+R^{(2)}\exp(-\sigma(k-l^{(2)}))}\right)^{(1/\sigma)+1} \int_{l^{(2)}}^{k} \frac{(\mu^{(2)}+R^{(2)}\exp(-\sigma(k-l^{(2)})))^{1/\sigma}}{(\mu^{(1)}+R^{(1)}\exp(-\sigma(k-l^{(1)})))^{(1/\sigma)+1}}dk + c^{(2)}(l^{(2)})e^{-(k-l^{(2)})}\left(\frac{\mu^{(2)}+R^{(2)}}{\mu^{(2)}+R^{(2)}e^{-\sigma(k-l^{(2)})}}\right)^{(1/\sigma)+1} \tag{D13}$$

Next, the case where $l^{(1)} \geq l^{(2)}$. For $l^{(2)} < k \leq l^{(1)}$,

$$c^{(2)}(k) = c_0\mu^{(1)}e^{-k}\frac{(\mu^{(1)}+R^{(1)}e^{-\sigma l^{(1)}})^{1-(1/\sigma)}}{(\mu^{(2)}+R^{(2)}\exp(-\sigma(k-l^{(2)})))^{(1/\sigma)+1}}\int_{l^{(2)}}^{k}\frac{(\mu^{(1)}+R^{(1)}\exp(-\sigma(l^{(1)}-k)))^{(1/\sigma)-1}}{(\mu^{(2)}+R^{(2)}\exp(-\sigma(k-l^{(2)})))^{-1/\sigma}}dk + c^{(2)}(l^{(2)})e^{l^{(2)}-k}\left(\frac{\mu^{(2)}+R^{(2)}}{\mu^{(2)}+R^{(2)}e^{-\sigma(k-l^{(2)})}}\right)^{(1/\sigma)+1} \tag{D14}$$

For $^{(2)} \leq l^{(1)} < k$,

$$c^{(2)}(k) = c^{(1)}(l^{(1)})\mu^{(1)}e^{-(k-l^{(1)})}\left(\frac{\mu^{(1)}+R^{(1)}}{\mu^{(2)}+R^{(2)}\exp(-\sigma(k-l^{(2)}))}\right)^{(1/\sigma)+1} \int_{l^{(2)}}^{k} \frac{(\mu^{(2)}+R^{(2)}\exp(-\sigma(k-l^{(2)})))^{1/\sigma}}{(\mu^{(1)}+R^{(1)}\exp(-\sigma(k-l^{(1)})))^{(1/\sigma)+1}}dk + c^{(2)}(l^{(1)})e^{-(k-l^{(1)})}\left(\frac{\mu^{(2)}+R^{(2)}e^{-\sigma(l^{(1)}-l^{(2)})}}{\mu^{(2)}+R^{(2)}e^{-\sigma(k-l^{(2)})}}\right)^{(1/\sigma)+1} \tag{D15}$$

The integrals in (D11) to (D15) can evaluated numerically. The result of the numerical computation is shown in Fig. A2.

Appendix E: Capability of the chemical network model to generate complex distributions

Is our glycan synthesis model capable of generating concentration distributions of arbitrary complexity? In what way do we need to change the parameters $N_E, N_C, \sigma, \ldots$, in order to generate glycan distributions $\mathbf{c}$ of a given complexity?

The purpose of this section, is to obtain some heuristics for this task.

We show in Appendix D that (B4) can be solved analytically in the limit $N_s \gg 1$, because in this limit the glycan index k can be approximated by a continuous variable, and the recursion relations for the steady state glycan concentrations (5)-(6) can be cast as a matrix differential equation. This allows us to obtain an *explicit* expression for the steady state concentration in terms of the parameters $(\mathbf{R}, \mathbf{L})$.

We derive our heuristics from a semi-analytical treatment in the limit $N_s \gg 1$ (Appendix D), which apart from being simple to implement in general, provides an explicit formula for c_k for the case $N_E = N_C = 1$ (D10). Figures A3(a)-(d) show the glycan profile c_k *vs.* k as one varies the enzyme specificity σ, the reaction rates R and transport rates μ, for two different values of N_E and N_C. The results in the plots lead us to the following general observations:

- Very low specificity enzymes cannot generate complex glycan distributions. Keeping everything else fixed, intermediate or high specificity enzymes can generate glycan distributions of higher complexity by increasing N_E or N_C (Figs. A3(a),(c)).

- Decreasing the specificity σ or increasing the rates R increases the proportion of higher index glycans. Keeping everything else fixed, changes in the rate R have a stronger impact on the relative weights of the higher index glycans to lower index glycans. The relative weight of the higher index glycans increases with increasing N_E and N_C (Figs. A3(b)-(d)).

- Keeping everything else fixed, decreasing enzyme specificity increases the spread of the distribution around the peaks (Figs. A3(a),(c)).

Appendix F: Parameter estimation

The typical transport time of glycoproteins across the Golgi complex is estimated to be in the range 15-20 mins. [23], which corresponds to the transport rate, $\mu = .18$/min. We bound the transport rate for our optimization between 0.01/min and 1/min.

Next, we estimate the range of values for the chemical reaction rates. The injection rate q is in the range $100 - 1500$ pmol/10^6 cell 24 h [23, 24]. For our calculation we set $q = 387.30$ pmol/10^6 cells 24 hr $= 0.27$ pmol/10^6 cells min, the geometric mean of 100 and 1500. We set the range for the enzymatic rate R to be

$$R_{\min} = \min_{\alpha} \left\{ \frac{V_{\max}^{(\alpha)}/\nu}{K_M^{(\alpha)} + \frac{1}{\nu}\frac{q}{\mu}} \right\} \leq R \leq R_{\max} = \max_{\alpha} \left\{ \frac{V_{\max}^{(\alpha)}/\nu}{K_M^{(\alpha)}} \right\}.$$

where $K_M^{(\alpha)}$ and $V_{\max}^{(\alpha)}$ denote the Michaelis constants and $V_{\max}$ of the α-th enzyme. The conversion from 1 pmoles/10^6 cells to concentration can be obtained by taking cisternal volume (ν) to be $2.5\mu m^3$ [23, 24]. This gives

$$1 \text{ pmoles}/10^6 \text{ cells} = \frac{10^{-12}\text{moles}}{10^6 \times 2.5 \times 10^{-18} \times 10^3 \text{litre}} = 400\mu M. \tag{F1}$$

In Table I we report the parameters for the 8 enzymes taken from Table 3 in [23]. From these parameters it follows that

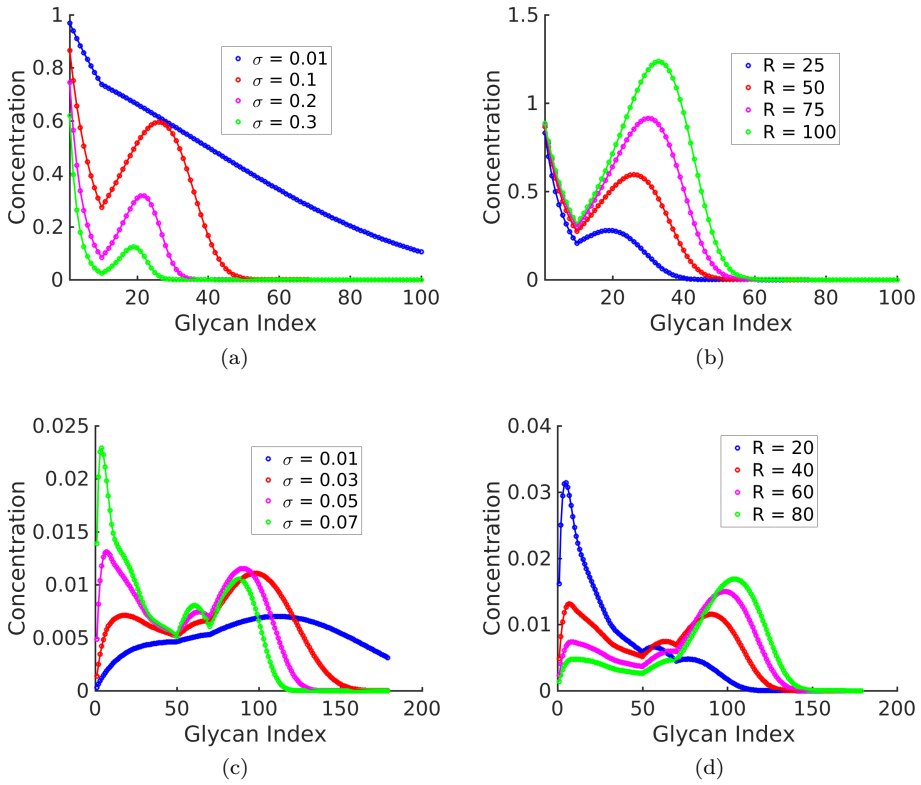


FIG. A3. Glycan profile $\{c_k : k = 1, \ldots, N_s\}$ as a function of specificity σ (Fig. (a), (c)), and reaction rates R (Fig. (b), (d)).
Fig. (a): $N_E = N_C = 1, (R = 50, \mu = 1, l = 10)$. c_k decreases exponentially with k for very low and very high σ; however, the decay rate is lower at low σ. For intermediate values of σ, the distribution has *exactly* two peaks, one of which is at $k = 0$, and eventually decays exponentially. The width of the distribution is a decreasing function of σ.
Fig. (b): $N_E = N_C = 1, (\sigma = 0.1, \mu = 1, l = 10)$. At low R, c_k is concentrated at low k. The proportion of higher index glycans in an increasing function of R.
Fig. (c): $N_E = N_C = 2, (R = 40, \mu = 1, [l_1^{(1)}, l_2^{(1)}, l_1^{(2)}, l_2^{(2)}] = [10, 30, 50, 70])$. As σ increases, the distribution becomes more complex – from a single peaked distribution at low σ to a maximum of four-peaked distribution at high σ. The peaks gets sharper, and more well defined as σ increases.
Fig. (d): $N_E = N_C = 2, (R = 40, \mu = 1, [l_1^{(1)}, l_2^{(1)}, l_1^{(2)}, l_2^{(2)}] = [10, 30, 50, 70])$. As in the plots in Fig. (b), increasing R shifts the peaks towards higher index glycans and the proportion of higher index glycan increases.

$$
\begin{aligned}
R_{\min} &= \min_{\alpha} \left\{ \frac{V_{\max}^{(\alpha)}/\nu}{K_M^{(\alpha)} + \frac{1}{\nu}\frac{q}{\mu}} \right\} \\
&= \frac{V_{\max}^{(7)}/\nu}{K_M^{(7)} + \frac{1}{\nu}\frac{q}{\mu}} = \frac{.16 \times 400\mu M/\text{min}}{3400\mu M + 149.4\mu M} = 0.018\text{min}^{-1} \\
R_{\max} &= \max_{\alpha} \left\{ \frac{V_{\max}^{(\alpha)}/\nu}{K_M^{(\alpha)}} \right\} \\
&= \frac{V_{\max}^{(1)}/\nu}{K_M^{(1)}} = \frac{5 \times 400\mu M/\text{min}}{100\mu M} = 20\text{min}^{-1}
\end{aligned}
$$

α	$K_M^{(\alpha)}$ (μmol)	$V_{\max}^{(\alpha)}$ (pmol/10^6 cell-min)
1	100	5
2	260	7.5
3	200	5
4	100	5
5	190	2.33
6	130	.16
7	3400	.16
8	4000	9.66

TABLE I. Enzyme parameters taken from Table 3 in [23] that we use to calculate the bounds on the reaction rate R. Here $K_M^{(\alpha)}$ and $V_{\max}^{(\alpha)}$ denote the Michaelis constant and $V_{\max}$ of the α-th enzyme.

Appendix G: Constructing target distributions for glycans of a given cell type

The target distribution of the glycans on the cell surface is obtained via mass spectrometry. The x-axis of mass spectroscopy (MS) graphs is mass/charge of the ionised sample molecules and the y-axis is relative intensity corresponding to each mass/charge value, taking the highest intensity as 100%.

This relative intensity roughly correlates with the relative abundances of the molecules in the sample.

This raw MS data is noisy and cannot be directly used as the target distribution in our optimization problem. There are three major sources of noise in the MS data [47]: the chemical noise in the sample, the Poisson noise associated with detecting discrete events, and the Nyquist-Johnson noise associated with any charge system. We propose a simple model that accounts for the chemical noise and the Poisson sampling noise. Using this noise model and the available MS data, we generate parametric bootstrap samples of glycan measurements, and fit a Gaussian Mixture Model (GMM) on this sample to approximate the glycan distribution. This GMM probability distribution is used as the target distribution in our numerical experiments.

The MS data obtained from [22] had mass ranging between 500 to 5000 Daltons with intensity reported at every 0.0153 Daltons. We first bin this MS data into 180 bins and take the maximum value within each bin as the value of intensity for that bin. Fig. A4 shows the raw MS data and the binned distribution.

Let $\bar{I}_k$ represents the relative intensity of the k-th bin in the binned MS graph. We generate a sample population of glycans using the MS data in the following way:

1. Poisson sampling noise: The MS data does not have absolute count information. We assume an arbitrary maximum count $I_{\max}$, and define the intensity $I_k = I_{\max}\bar{I}_k$. The plots in Fig. A5(a) show that the results are not sensitive to the specific value of $I_{\max}$.

2. Chemical noise: The sample used for MS analysis also contains small amounts of molecules that are not glycans. These appear as the very small peaks in the MS data. We assume that the probability p_k that the peak at index k corresponds to a glycan is given by

$$p_k = 1 - e^{-\frac{I_k}{I_{max}}} = 1 - e^{-\bar{I}_k}$$

   which adequately suppresses this chemical noise.

3. Bootstrapped glycan data: The count n_k at the glycan index k is distributed according to the following distribution:

$$n_k = \begin{cases} 0 & (1-p_k) & n=0 \\ n & p_k e^{-I_k}\frac{(I_k)^n}{n!} & n \geq 1. \end{cases}$$

   We assume that the MS data was generated from N different cells. Thus, the total count at glycan index k is given by the sum of N_i i.i.d. samples distributed according to the distribution above. We in Fig. A5 (b) show that results are insensitive to N_i.

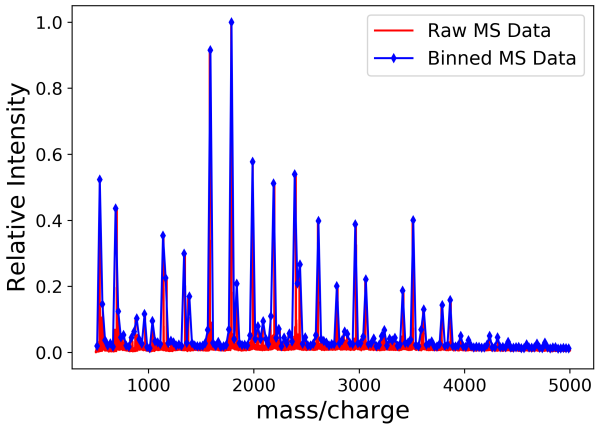
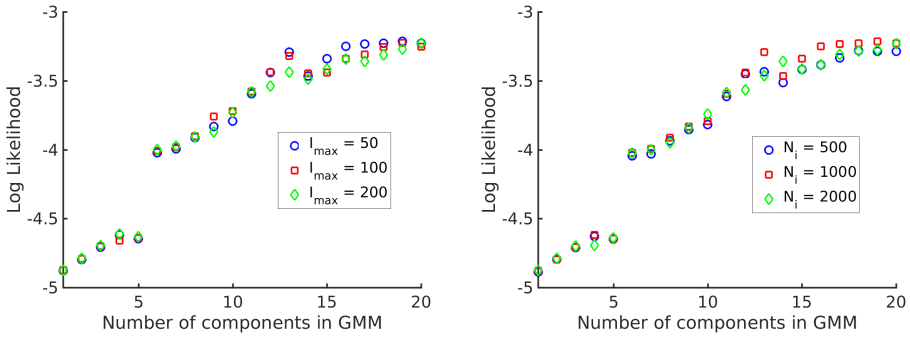


FIG. A4. The binned MS data (blue) approximates the raw MS data (red) very well. We use this binned data for GMM approximation of the MS data.



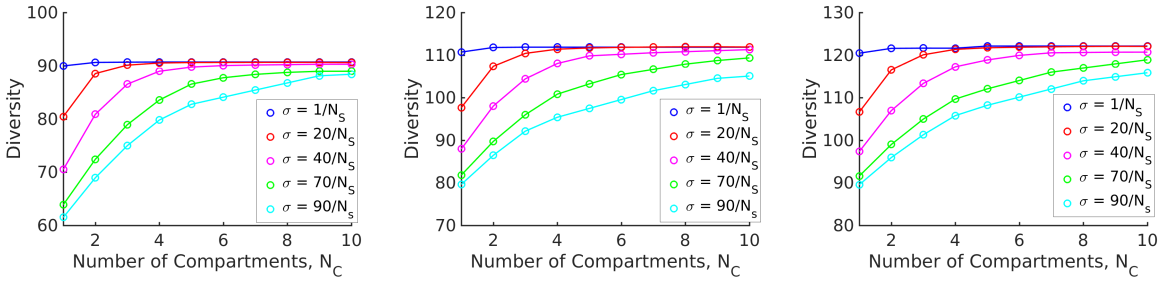
(a) (b)

FIG. A5. Log likelihood vs. number of components (N) in the GMM. We see that the log likelihood saturates at around $N = 20$, thus 20-GMM is a very good representation of the MS-data from *human* T-cells. The different symbols are for (a) different values of the maximum intensity $I_{max} = 50, 100, 200$ and (b) different values of the number of i.i.d samples $N_i = 500, 1000, 2000$, showing the insensitivity of the log likelihood to the value of I_{max} and N_i.

Next, we interpret the counts as samples from a "spatial" distribution f. We approximate this distribution as a Gaussian mixture, i.e. $f(x) = \sum_{\ell=1}^{N} \gamma_\ell \, \eta(x \mid \mu_\ell, \sigma_\ell)$, where $\eta(x \mid \mu, \sigma)$ denotes the density of a normally distributed random variable with mean μ and standard deviation σ at the location x. In this setting, we assume that each count is a sample from the distribution $\eta(x \mid \mu_\ell, \sigma_\ell)$ with probability γ_ℓ. Thus, each count is classified as coming from one of the Gaussian components.

Appendix H: Numerical scheme for performing the non-convex optimization

We solve Optimization C using the numerical scheme detailed below. The optimization problem consists of minimising a non-convex objective with linear box constraints. We use the MATLAB FMINCON function to solve this optimization. We use Sequential Quadratic Programming (SQP), a gradient based iterative optimization scheme for solving optimizations with non-linear differentiable objective and constraints. Since our problem is non-convex and SQP only gives local minima, we initialise the algorithm with many random initial points. We use SOBOLSET function of MATLAB to generate space filling pseudo random numbers. We have taken 1000 initialisations for each N_E, N_C and σ value. We have taken 50 equally spaced points between 0 and 1 to explore the σ-space for Fig. 3. Some minor fluctuations in D due to non-convexity of the objective function in the final results were smoothed out by taking the convex hull of the D vs. σ graph. The results for $\sigma_{min}(N_E, N_C)$ and $D(\sigma_{min}, N_E, N_C)$ (Fig. 4) were obtained by



(a) $c_{th} = \frac{1}{N_s}$ (b) $c_{th} = \frac{1}{2N_s}$ (c) $c_{th} = \frac{1}{4N_s}$

FIG. A6. Diversity vs. N_C for different values of σ keeping $N_E = 1$ fixed, for three different values of the threshold, $c_{th} = \frac{1}{N_s}, \frac{1}{2N_s}, \frac{1}{4N_s}$. Changing the value of the threshold c_{th}, only changes the saturation value of the diversity curve.

adding σ to the optimization vector and then performing the optimization. The sensitivity results (Figs. 4e and 4e) were obtained by approximating the D vs σ graph around σ_{min} with a parabola, the coefficient of the quadratic term being the curvature of the graph at σ_{min}.

A similar numerical scheme was used to optimize diversity.

[1] B. Alberts *et al.*, *Molecular Biology of the Cell* (Garland Science, 2002).
[2] A. Varki *et al.*, *Essentials of Glycobiology* (Cold Spring Harbor Laboratory Press, 2009).
[3] R. D. Cummings and J. M. Pierce, Chemistry & biology **21**, 1 (2014).
[4] A. Varki, Glycobiology **27**, 3 (2017).
[5] K. Drickamer and M. E. Taylor, Trends in biochemical sciences **23**, 321 (1998).
[6] P. Gagneux and A. Varki, Glycobiology **9**, 747 (1999).
[7] H.-J. Gabius, BioSystems **164**, 102 (2018).
[8] R. A. Dwek, Chemical reviews **96**, 683 (1996).
[9] P. Winterburn and C. Phelps, Nature **236**, 147 (1972).
[10] A. Varki, Trends in cell biology **8**, 34 (1998).
[11] P. Pothukuchi, I. Agliarulo, D. Russo, R. Rizzo, F. Russo, and S. Parashuraman, FEBS letters **593**, 2390 (2019).
[12] F. Bard and J. Chia, Trends in cell biology **26**, 379 (2016).
[13] G. D'Angelo, S. Capasso, L. Sticco, and D. Russo, The FEBS journal **280**, 6338 (2013).
[14] M. Demetriou, M. Granovsky, S. Quaggin, and J. W. Dennis, Nature **409**, 733 (2001).
[15] C. WILLS and D. R. GREEN, Immunological reviews **143**, 263 (1995).
[16] T. M. Cover and J. A. Thomas, *Elements of information theory* (John Wiley & Sons, 2012).
[17] D. J. MacKay, *Information theory, inference and learning algorithms* (Cambridge university press, 2003).
[18] D. Sengupta and A. D. Linstedt, Annual review of cell and developmental biology **27**, 57 (2011).
[19] H. Sachdeva, M. Barma, and M. Rao, Scientific reports **6**, 1 (2016).
[20] P. Sens and M. Rao, in *Methods in cell biology*, Vol. 118 (Elsevier, 2013) pp. 299–310.
[21] A. Bacharoglou, Proceedings of the American Mathematical Society **138**, 2619 (2010).
[22] R. D. Cummings and P. Crocker, *Functional Glycomics Database, Consortium for Functional Glycomics,* `http://www.functionalglycomics.org` (2020).
[23] P. Umaña and J. E. Bailey, Biotechnology and bioengineering **55**, 890 (1997).
[24] F. J. Krambeck, S. V. Bennun, S. Narang, S. Choi, K. J. Yarema, and M. J. Betenbaugh, Glycobiology **19**, 1163 (2009).
[25] F. J. Krambeck and M. J. Betenbaugh, Biotechnology and Bioengineering **92**, 711 (2005).
[26] P. Fisher, H. Spencer, J. Thomas-Oates, A. J. Wood, and D. Ungar, Cell reports **27**, 1231 (2019).
[27] P. Fisher and D. Ungar, Frontiers in cell and developmental biology **4**, 15 (2016).
[28] C. B. Hirschberg, P. W. Robbins, and C. Abeijon, *Transporters of nucleotide sugars, ATP, and nucleotide sulfate in the endoplasmic reticulum and Golgi apparatus,* (1998).
[29] C. E. Caffaro and C. B. Hirschberg, Accounts of chemical research **39**, 805 (2006).
[30] P. M. Berninsone and C. B. Hirschberg, Current opinion in structural biology **10**, 542 (2000).
[31] N. Trinajstic, *Chemical graph theory* (Routledge, 2018).
[32] N. Price and L. Stevens, *Fundamentals of Enzymology: The cell and molecular biology of catalytic proteins* (Oxford University Press, 1999).
[33] K. W. Moremen and R. S. Haltiwanger, Nature chemical biology **15**, 853 (2019).

[34] S. Kellokumpu, Frontiers in cell and developmental biology **7**, 93 (2019).
[35] J. R. Casey, S. Grinstein, and J. Orlowski, Nature reviews Molecular cell biology **11**, 50 (2010).
[36] S. Dmitrieff, M. Rao, and P. Sens, Proceedings of the National Academy of Sciences **110**, 15692 (2013).
[37] J. Llopis, J. M. McCaffery, A. Miyawaki, M. G. Farquhar, and R. Y. Tsien, Proceedings of the National Academy of Sciences **95**, 6803 (1998).
[38] J. Monod, J. Wyman, and J.-P. Changeux, J Mol Biol **12**, 88 (1965).
[39] J.-P. Changeux and S. J. Edelstein, Science **308**, 1424 (2005).
[40] Y. Savir and T. Tlusty, PloS one **2**, e468 (2007).
[41] S. Roseman, Journal of Biological Chemistry **276**, 41527 (2001).
[42] P. Hossler, B. C. Mulukutla, and W.-S. Hu, PloS one **2** (2007).
[43] M. Yang, C. Fehl, K. V. Lees, E.-K. Lim, W. A. Offen, G. J. Davies, D. J. Bowles, M. G. Davidson, S. J. Roberts, and B. G. Davis, Nature chemical biology **14**, 1109 (2018).
[44] A. Bar-Even, R. Milo, E. Noor, and D. S. Tawfik, Biochemistry **54**, 4969 (2015).
[45] S. Boyd and L. Vandenberghe, *Convex optimization* (Cambridge university press, 2004).
[46] A. Jaiman and M. Thattai, BioRxiv , 440792 (2018).
[47] P. Du, G. Stolovitzky, P. Horvatovich, R. Bischoff, J. Lim, and F. Suits, Bioinformatics **24**, 1070 (2008).
[48] A. Varki, Cold Spring Harbor perspectives in biology **3**, a005462 (2011).
[49] J. W. Dennis, I. R. Nabi, and M. Demetriou, Cell **139**, 1229 (2009).
[50] H. van Halbeek, G. J. Gerwig, J. F. Vliegenthart, H. L. Smits, P. J. Van Kerkhof, and M. F. Kramer, Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology **747**, 107 (1983).
[51] H. E. McFarlane, A. Döring, and S. Persson, Annual review of plant biology **65**, 69 (2014).
[52] B. E. Koch, J. Stougaard, and H. P. Spaink, Glycobiology **25**, 469 (2015).
[53] M. A. O'Neill, T. Ishii, P. Albersheim, and A. G. Darvill, Annu. Rev. Plant Biol. **55**, 109 (2004).
[54] T. Hayashi and R. Kaida, Molecular Plant **4**, 17 (2011).
[55] P. Kumar, M. Yang, B. C. Haynes, M. L. Skowyra, and T. L. Doering, Current opinion in structural biology **21**, 597 (2011).
[56] N. A. Gow and B. Hube, Current opinion in microbiology **15**, 406 (2012).
[57] M. A. Atmodjo, Z. Hao, and D. Mohnen, Annual review of plant biology **64** (2013).
[58] S. J. Free, in *Advances in genetics*, Vol. 81 (Elsevier, 2013) pp. 33–82.
[59] M. Pauly, S. Gille, L. Liu, N. Mansoori, A. de Souza, A. Schultink, and G. Xiong, Planta **238**, 627 (2013).
[60] R. A. Burton and G. B. Fincher, Frontiers in plant science **5**, 456 (2014).
[61] B. Becker and M. Melkonian, Microbiol. Mol. Biol. Rev. **60**, 697 (1996).
[62] A. A. Mironov, I. S. Sesorova, E. V. Seliverstova, and G. V. Beznoussenko, Tissue and Cell **49**, 186 (2017).
[63] B. S. Donohoe, B.-H. Kang, and L. A. Staehelin, Proceedings of the National Academy of Sciences **104**, 163 (2007).
[64] S. Mogelsvang, N. Gomez-Ospina, J. Soderholm, B. S. Glick, and L. A. Staehelin, Molecular biology of the cell **14**, 2277 (2003).
[65] M. S. Ladinsky, C. C. Wu, S. McIntosh, J. R. McIntosh, and K. E. Howell, Molecular biology of the cell **13**, 2810 (2002).
[66] P. Stanley, Cold Spring Harbor perspectives in biology **3**, a005199 (2011).
[67] H. Nam, N. E. Lewis, J. A. Lerman, D.-H. Lee, R. L. Chang, D. Kim, and B. O. Palsson, Science **337**, 1101 (2012).
[68] A. Peracchi, Trends in biochemical sciences (2018).
[69] S. Blanes, F. Casas, J. Oteo, and J. Ros, Physics Reports **470**, 151 (2009).