bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genome-scale reconstructions of the mammalian secretory pathway predict metabolic costs and limitations of protein secretion

Jahir M. Gutierrez[1,2], Amir Feizi[3], Shangzhong Li[1,2], Thomas B. Kallehauge[4], Hooman Hefzi[1,2], Lise M. Grav[4], Daniel Ley[4,5], Deniz Baycin Hizal[6], Michael J. Betenbaugh[7], Bjorn Voldborg[4], Helene Faustrup Kildegaard[4], Gyun Min Lee[4], Bernhard O. Palsson[1,2,4,8], Jens Nielsen[3,4] & Nathan E. Lewis[2,8,+]

[1]Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, United States
[2]Novo Nordisk Foundation Center for Biosustainability at the University of California, San Diego, School of Medicine, La Jolla, CA 92093, United States
[3]Department of Biology and Biological Engineering, Kemivägen 10, Chalmers University of Technology, SE41296 Gothenburg, Sweden
[4]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.
[5]Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, Denmark
[6]Pharmaceutical R&D Department, Turgut Illaclari A.S., Istanbul, Turkey
[7]Department of Chemical & Biomolecular Engineering, Johns Hopkins University, Baltimore, MD 21218-2686, United States
[8]Department of Pediatrics, University of California, San Diego, School of Medicine, La Jolla, CA 92093, United States.
* Co-first author
+ Corresponding author
Correspondence and requests for materials should be addressed to N.E.L. (email: nlewisres@ucsd.edu)

Abstract

In mammalian cells, >25% of proteins are synthesized and exported through the secretory pathway. The pathway complexity, however, obfuscates its impact on the secretion of different proteins. Unraveling its impact on diverse proteins is particularly important, since the pathway is used for biopharmaceutical production. Here we delineate the core secretory pathway functions and integrate them with genome-scale metabolic models of human, mouse, and Chinese hamster ovary (CHO) cells. The resulting reconstructions RECON2.2s, iMM1685s, and iCHO2048s, enable the computation of the cost and machinery demanded by each secreted protein. We predicted metabolic costs and maximum productivities of biotherapeutic proteins and identified protein features that most significantly impact protein secretion. By integrating additional metabolomic, glycoproteomic and ribosomal profiling data, we further found that CHO cells have adapted to reduce expression and secretion of expensive host cell proteins. Finally, the model successfully predicts the increase in titers after silencing a highly expressed selection marker. This work represents a knowledge-base of the mammalian secretory pathway that serves as a novel tool for systems biotechnology.

Keywords

Metabolic network, secretory pathway, biotherapeutic production, systems biotechnology

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Introduction

To interact with their environment, cells produce numerous signaling proteins, hormones, receptors, and structural proteins. In mammals, these include >3000 secreted proteins (e.g., enzymes, hormones, antibodies, extracellular matrix proteins, etc.) and >5500 membrane proteins[1], most of which are synthesized and processed in the secretory pathway. The secretory pathway is a complex series of processes predominantly in the endoplasmic reticulum (ER), Golgi apparatus, and other compartments of the endomembrane system. In these compartments, the synthesis of the thousands of membrane and secreted proteins is facilitated by hundreds of other proteins that make up the secretory pathway machinery. These are used to translate, fold, post-translationally modify, test for quality, sort and translocate the secreted proteins.

The secretory pathway is particularly important in biotechnology and the biopharmaceutical industry, since most therapeutic proteins are secreted. Mammalian cell lines such as HEK293, PerC6, and Chinese hamster ovary (CHO) cells are used extensively to ensure that a secreted biotherapeutic is properly folded and contains the necessary post-translational modifications (PTMs) [2]. For any given biotherapeutic, different machinery in the secretory pathway may be needed, and each step can exert a non-negligible metabolic demand on the cells. The complexity of this pathway, however, makes it unclear how the biosynthetic cost and cellular needs vary for different secreted proteins, each of which exerts different demands for cellular resources. Therefore, a detailed understanding of the biosynthetic costs of the secretory pathway could guide efforts to engineer host cells and bioprocesses for any desired product. The energetic and material demands of the mammalian secretory pathway can be accounted for by substantially extending the scope of metabolic models. Indeed, recent studies have incorporated portions of the secretory pathway in metabolic models of yeast [3–5].

Here we present the first genome-scale reconstructions of mammalian metabolism and protein secretion. Specifically, we constructed these for human, mouse, and CHO cells, called RECON2.2s, iMM1685s, and iCHO2048s, respectively. Given its dominant role in biotherapeutic production, we focus here on the biosynthetic capabilities of CHO cells, while providing models as a resource. We first demonstrate that product-specific secretory pathway models can be built to estimate CHO cell growth rates given the specific productivity of the recombinant product as a constraint. Second, we identify the features of secreted proteins that have the highest impact on protein cost and productivity rates. Third, we use our model to identify proteins that compete for cell resources, thereby presenting targets for cell engineering. Finally, we derive an expression for computing the energetic cost of synthesizing and

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

secreting a product in terms of molecules of ATP equivalents per protein molecule. We use this expression and analyze how the energetic burden of protein secretion has led to an overall suppression of more expensive secreted host cell proteins in CHO cells. Through this study we demonstrate that a systems-view of the secretory pathway now enables the analysis of many biomolecular mechanisms controlling the efficacy and cost of protein expression in mammalian cells. We envision our models as valuable tools for the study of normal physiological processes and engineering cell bioprocesses in biotechnology. Our models, Jupyter notebooks, and data are available at https://github.com/LewisLabUCSD.

2. Material and Methods

2.1 Reconstruction of the mammalian secretory pathway

A list of proteins and enzymes in the mammalian secretory pathway was compiled from literature curation, UniProt, NCBI Gene, NCBI Protein and CHOgenome.org (see Supp. File 1). To facilitate the reconstruction process, the secretory pathway was divided into twelve subsystems or functional modules (Fig. 1) to sort the components according to their function. These subsystems correspond to the major steps required to process and secrete a protein. The components from a prior yeast secretory pathway reconstruction[3] were used as a starting reference. To build species-specific models, orthologs for human, mouse and the Chinese hamster were identified and used, while yeast components and subsystems that are not present in the mammalian secretory pathway were removed. Additional subsystems were added when unique to higher eukaryotes, such as the calnexin-calreticulin cycle in the ER[6]. These were constructed *de novo* and added to the reconstruction. The databases and literature were then consulted to identify the remaining components involved in each subsystem of the mammalian secretory pathway. Since most components in the mammalian secretory pathway have been identified in mouse and human, BLAST was utilized to identify the corresponding Chinese hamster orthologs by setting human as the reference organism and a cutoff of 60% of sequence identity. See Supp. File 4 for a concise overview of the mammalian secretory pathway and its comparison with the yeast secretory pathway.

2.2 Protein Specific Information Matrix (PSIM)

The PSIM (Supp. File 2) contains the necessary information to construct a protein-specific secretory model from the template reactions in our reconstruction. The columns in the PSIM are: presence of a

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

signal peptide (SP), number of disulfide bonds (DSB), presence of Glycosylphosphatidylinositol (GPI) anchors, number of N-linked (NG) and O-linked (OG) glycans, number of transmembrane domains (TMD), subcellular location, protein length, and molecular weight. For most proteins, the information in the PSIM was obtained from the Uniprot database. When necessary, computational tools were used to predict signal peptides (PrediSi[7]) and GPI anchors (GPI-SOM[8]). Finally, additional information on the number of O-linked glycosylation sites of certain proteins were obtained from experimental data in previous studies[9,10]. The PSIMs of the CHO and human secretomes are a subset of the full PSIM and contains only the proteins with a signal peptide (predicted or confirmed in Uniprot). The distribution of all PTMs across the human, mouse and CHO proteomes can be found in Supp. Notebook 5.

2.3 Detection of N-linked glycosylation sites via mass spectrometry

The number of N-linked glycosylation sites in the PSIM was determined computationally and experimentally as follows. CHO-K1 cells (ATCC) were lysed, denatured, reduced, alkylated and digested by trypsin. Desalted peptides were incubated with 10 mM sodium periodate in dark for 1 hour before coupling to 50 µL of (50% slurry) hydrazide resins. After incubation overnight, non-glycosylated peptides were washed with 1.5 M NaCl and water. The N-glycosylated peptides were released with PNGaseF at 37 °C and desalted by using a C18 SepPak column. Strong cation exchange (SCX) chromatography was used to separate the sample into 8 fractions. Each fraction was analyzed on an LTQ-Orbitrap Velos (Thermo Electron, Bremen, Germany) mass spectrometer. During the mass spectrometry data analysis, carbamidomethylation was set as a fixed modification while oxidation, pyroglutamine and deamidation were variable modifications.

2.4 Construction of secretory models and constraint-based analysis

We wrote Jupyter Notebooks in both Python and MATLAB (see Supp. Jupyter Notebooks 1 and 2) that take a row from the PSIM as input to produce an expanded iCHO2048s metabolic model with the product-specific secretory pathway of the corresponding protein. Flux balance analysis (FBA[11]) and all other constraint-based analyses were done using the COBRA toolbox[12] in MATLAB R2014a and the Gurobi solver version 6.0.0. The analyses in Figs. 2, 3, and 4 were done using the constraints in Supp. File 3. For the iCHO2048s models secreting human proteins, we set the same constraints in all models and computed the theoretical maximum productivity (max_{qp}) while maintaining a growth rate (in units of 1/h) of 0.01. Finally, since the exact glycoprofiles of most proteins in CHO are unknown and some even

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

change over time in culture[13], we simplified our models by only adding the core N-linked and O-linked glycans to the secreted proteins.

2.5 Batch cultivation

Two isogenic CHO-S cell lines adapted to grow in suspension, one producing Enbrel (Etanercept) and the other producing human plasma protease C1 inhibitor (SERPING1), were seeded at 3 x 10^5 cells/mL in 60 mL CD-CHO medium (Thermo Fisher Scientific, USA) supplemented with 8 mM L-Glutamine and 1 µL/mL anti-clumping agent, in 250 mL Erlenmeyer shake flasks. Cells were incubated in a humidified incubator at 37°C, 5% CO_2 at 120 rpm. Viable cell density and viability were monitored every 24 hours for 7 days using the NucleoCounter NC-200 Cell Counter (ChemoMetec). Daily samples of spent media were taken for extracellular metabolite concentration and titer measurements by drawing 0.8 mL from each culture, centrifuging it at 1000 g for 10 minutes and collecting the supernatant and discarding the cell pellet.

2.6 Titer determination

To quantify Enbrel/SERPING1, biolayer interferometry was performed using an Octet RED96 (Pall Corporation, Menlo Park, CA). ProA biosensors (Fortebio 18-5013) were hydrated in PBS and preconditioned in 10 mM glycine pH 1.7. A calibration curve was prepared using Enbrel (Pfizer) or SERPING1 at 200, 100, 50, 25, 12.5, 6.25, 3.13, 1.56, 0.78 µg/ml. Culture spent media samples were collected after centrifugation and association was performed for 120 s with a shaking speed of 200 rpm at 30 °C. Octet System Data Analysis 7.1 software was used to calculate binding rates and absolute protein concentrations.

2.7 Extracellular metabolite concentration measurements

The concentrations of glucose, lactate, ammonium (NH_4^+), and glutamine in spent media were measured using the BioProfile 400 (Nova Biomedical). Amino acid concentrations were determined via High Performance Liquid Chromatography using the Dionex Ultimate 3000 autosampler at a flow rate of 1mL/min. Briefly, samples were diluted 10 times using 20 µL of sample, 80 µL MiliQ water, and 100 µL of an internal amino acid standard. Derivatized amino acids were monitored using a fluorescence detector. OPA-derivatized amino acids were detected at 340ex and 450em nm and FMOC-derivatized amino acids at 266ex and 305em nm. Quantifications were based on standard curves derived from dilutions of a mixed amino acid standard (250 ug/mL). The upper and lower limits of quantification were 100 and 0.5 ug/mL, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.8 Estimation of the energetic cost of secreting a protein as the number of ATP equivalent molecules

We estimated the energetic cost of synthesizing and secreting all 5,641 endogenous CHO cell proteins. These proteins were chosen for containing a signal peptide in their sequence and/or for being localized in the cell membrane (according to the UniProt database). The energetic cost (in units of number of ATP equivalents) of secreting each protein (length *L*) was computed using the following formulas and assumptions:

1. **Energy cost of translation**. For each protein molecule produced, ***2L*** ATP molecules are cleaved to AMP during charging of the tRNA with a specific amino acid; ***1*** GTP molecule is consumed during initiation and ***1*** GTP molecule for termination; L **-** ***1*** GTP molecules are required for the formation of L-1 peptide bonds; ***L - 1*** GTP molecules are necessary for L-1 ribosomal translocation steps. Thus, the total cost of translation (assuming no proofreading) is ***4L***.
2. **Average cost of signal peptide degradation**. On average, signal peptides have a length of 22 amino acids. Thus, the average cost of degrading all peptide bonds in the signal peptide is ***22***. This average cost was assigned to all proteins analyzed.
3. **Energetic cost of translocation across the ER membrane**. During activation of the translocon, ***2*** cytosolic GTP molecules are hydrolyzed. From there, a GTP molecule bound to the folding-assisting chaperone BiP is hydrolyzed to GDP for every 40 amino acids that pass through the translocon pore[14]. Thus, the cost of translocation is **L/40 + 2**.
4. **Energetic cost of vesicular transport and secretion**. We used published data[15–17] (see Supp. File 1) to compute stoichiometric coefficients for reactions involving vesicular transport. That is, the number of GTP molecules bound to RAB and coat proteins in each type of vesicle (COPII and secretory vesicles). We found that a total of 192 and 44 GTPs must be hydrolyzed to transport one COPII or secretory (i.e. clathrin coated) vesicle from the origin membrane to the target membrane, respectively. Since vesicles do not transport a single protein molecule at a time, we estimated the number of secreted protein molecules that would fit inside a spherical vesicle (see estimated and assumed diameters in Supp. File 1). For that, we assumed that the secreted protein is globular and has a volume V_P (nm^3) that is directly proportional to its molecular weight MW[18]:

$$V_P = MW \times 0.00121$$

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, we assumed that only 70 percent of the vesicular volume can be occupied by the target protein. Thus, the cost of vesicular transport via COPII vesicles with Volume V_{COPII} is:

$$192\ GTPs\ \div (V_{COPII}\ \times 0.7\ \div V)$$

Similarly, the cost of vesicular secretion is:

$$44\ GTPs\ \div (V_{Secretory}\ \times 0.7\ \div V)$$

2.9 Constraints used in models and Pareto optimality frontiers

All models were constrained using different sets of experimental uptake rates, which can be found in Supp. File 3. To construct Pareto optimality frontiers, we used the robustAnalysis function from the COBRA Toolbox in Matlab 2015b using biomass as the control and secretion of the recombinant protein as the objective reactions, respectively.

2.10 Analysis of gene expression versus protein cost

Ribosome-profiling data[19] were used to quantify the ribosomal occupancy of each transcript in CHO cells. A cutoff of 1 RPKM was used to remove genes with low expression (10,045 genes removed from day 3 analysis and 10,411 from day 6 analysis). We used Spearman correlation to assess the variation of expression levels with respect to protein ATP cost.

2.11 CHO-DG44 model and prediction of neoR knock-out effect on specific productivity

Ribosome-profiling data, specific productivity, product sequence, and growth rates of an IgG-producing CHO-DG44 cell line were obtained previously[19]. From the same cultures, we obtained further cell dry weight and metabolomic data from spent culture medium for this study. The mCADRE algorithm[20,21] was used to construct a DG44 cell line-specific iCHO2048s model. The specific productivity and the RPKM values of the secreted IgG were used to estimate the translation rate for the neoR selection marker gene. We assumed that the flux (in units of mmol/gDW/h) through the neoR translation reaction (v_{neoR}) should be proportional to that of the IgG translation rate (v_{IgG}, calculated from the measured specific productivity) and related to their expression ratios (i.e. the RPKM values of their genes in the ribosome-profiling data).

$$v_{neoR} = \frac{RPKM_{neoR}}{2(RPKM_{light} + RPKM_{heavy})} v_{IgG}$$

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, a reaction of neoR peptide translation (which is expressed in cytosol and is not processed in the secretory pathway) was added to construct a neoR-specific iCHO2048s model. Uptake and secretion rates of relevant metabolites on days 3 and 6 of cell culture were used to constrain our model. Because recombinant proteins represent 20% of total cell protein[22], we scaled the coefficients of all 20 amino acids in the model's biomass reaction accordingly (i.e. each coefficient was multiplied by 0.8). We then used FBA to predict the specific productivity of IgG with or without neoR.

2.12 Cell dry weight measurements

For cell dry weight measurements, 6 tubes containing 2 mL of culture samples of known viable cell density and viability were freeze dried, weighed, washed in PBS, and weighed again. The difference in weight was used to calculate the mass per cell. The procedure resulted in an average cell dry weight of 456 pg per cell. As a simplification, we assumed that cell dry weight does not significantly differ from this average measured value during culture and thus was used when computing flux distributions in all simulations.

2.13 Calculation of amino acid uptake, growth rates and specific productivity from experimental data

Supp. File 3 contains the uptake and secretion rates used to constrain the iCHO2048s models[19,22,23]. When rates were not explicitly stated in the studies we consulted, we used a method we developed previously[24]. Briefly, appropriate viable cell density, titer, and metabolite concentration plots were digitized using WebPlot Digitizer software and we computed the corresponding rates as follows:

- Growth rate (in units of inverse hours):

$$\mu = \frac{1}{VCD}\frac{d}{dt}VCD$$

Where VCD is the viable cell density (in units of cells per milliliter)

- Specific productivity (in units of picograms per cell per hour):

$$q_p = \frac{1}{VCD}\frac{d}{dt}Titer$$

- Consumption or production rate v_x of metabolite x (in units of millimoles per gram dry weight per hour):

$$v_x = \frac{1}{VCD}\frac{d}{dt}[x]$$

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3. RESULTS

3.1 *In silico* reconstruction of the mammalian protein secretion pathway

We mapped out the core processes involved in the synthesis of secreted and membrane proteins in mammalian cells (i.e. human, mouse, and Chinese hamster). This included 261 components (gene products) in CHO cells and 271 components in both human and mouse. The components are involved in secretory reactions across 12 subsystems (i.e., functional modules of the secretory pathway; Fig. 1A). These components represent the core secretory machinery needed in the transition of a target protein from its immature state in the cytosol (i.e., right after translation) to its final form (i.e., when it contains all post-translational modifications and is secreted to the extracellular space). Each component in the reconstruction either catalyzes a chemical modification on the target protein (e.g., N-linked glycosylation inside ER lumen/Golgi) or participates in a multi-protein complex that promotes protein folding and/or transport. This distinction between catalytic enzymes and complex-forming components is important for modeling purposes as a catalytic component consumes or produces metabolites that are directly connected to the metabolic network (e.g., ATP, sugar nucleotides). Because all components of the core secretory pathway were conserved across human, mouse and hamster (Fig. 1B), we generated species-specific reconstructions and used them to expand the respective genome-scale metabolic network reconstructions (Recon 2.2[25], iMM1415[26], iCHO1766[23]) and called these metabolic-secretory reconstructions Recon 2.2s, iMM1685s, and iCHO2048s, respectively. A detailed list of the components, reactions and the associated genes can be found in Supp. File 1.

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

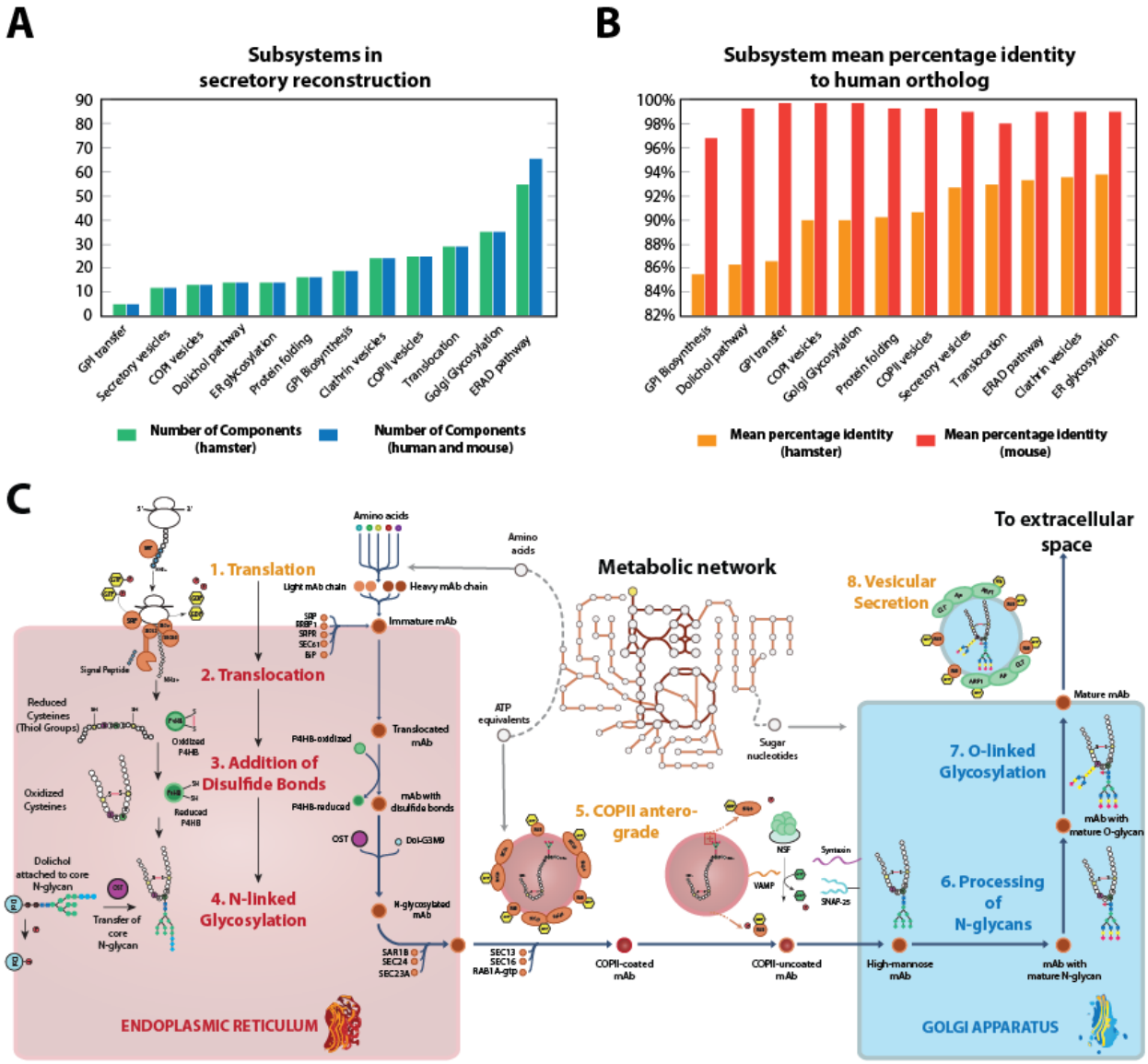


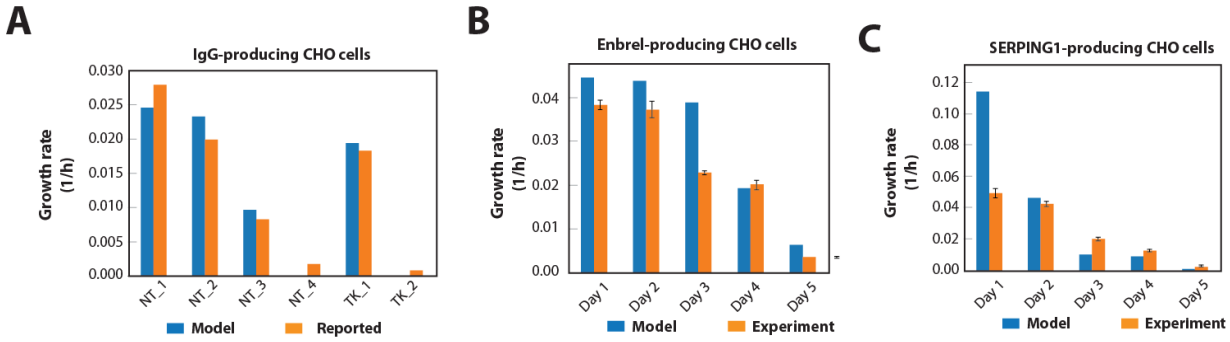
Figure 1 - Components in the reconstruction of secretory pathway in mammalian cells. (A) The reconstruction comprises 261 proteins in CHO cells and 271 proteins in human and mouse that are distributed across 12 subsystems. The different component numbers arise from the fact that the Chinese hamster proteome annotation only contains one alpha and one beta proteasome subunits, whereas the human and mouse contain 12 subunits of different subtypes. The detailed description of all components can be found in Supp. File 1. **(B)** High similarities were seen for proteins in CHO and human, with a high mean percentage identity in each subsystem (calculated with the sequence alignment tool BLAST). **(C)** Simplified schematic of reactions and subsystems involved in the secretion of a monoclonal antibody (mAb). A total of eight subsystems are necessary to translate, fold, transport, glycosylate, and secrete a mAb. The color of the subsystem names indicates if the reactions occur in the cytoplasm

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(yellow), the ER lumen (red) or the Golgi apparatus (blue). GPI = Glycosylphosphatidylinositol, ER = Endoplasmic Reticulum, ERAD = ER associated degradation.

3.2 iCHO2048s predicts measured growth rates for recombinant-protein-producing cells

We first validated the accuracy of iCHO2048s predictions using published growth and specific productivity rates of IgG-producing CHO cell lines from two independent studies[19,27]. For this, we built an IgG-secreting iCHO2048s model using the information in the PSIM matrix for the therapeutic monoclonal antibody (mAb) Rituximab. We then constrained the model's Rituximab-specific secretory pathway with the reported productivity value in each study and used FBA to predict growth (Fig. 2A). Later, to assess the ability of iCHO2048s to predict growth rates in cases when CHO cells are producing non-antibody proteins, we collected data from two batch culture experiments using Enbrel- and SERPING1-producing isogenic CHO cell lines. We constructed two iCHO2048s models for each case and predicted growth rates on days 1 (early exponential growth phase) through 5 (late exponential growth phase) of culture while constraining the protein secretion rate to the measured specific productivity value (Fig. 2B-C). The model's predictions agreed well with the reported/measured values. There were cases where iCHO2048s predicted a much higher growth rate than what was measured in the first days of batch culture (Fig. 2B-C). Since FBA computes theoretical maximum growth rates given a set of constraints, these over-prediction cases point at situations where CHO cells do not direct resources towards biomass production (during very early stages of culture), a discrepancy that is attenuated in later stages of culture (days 4-5 in Fig. 2B-C). In conclusion, these results confirm the ability of protein-specific reconstructions to capture the specific energetic requirements that each recombinant product imposes on CHO cell metabolism.



bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2 – Recombinant-protein-producing models of iCHO2048s predict measured growth rates. (A) Growth rates were computed using an IgG-specific iCHO2048s model and compared to experimentally-measured growth rates from six datasets from two previous studies using IgG-producing cell lines. NT and TK specify the initials of the first author of thetwo studies (Neil Templeton, Thomas Kallehauge). **(B-C)** Additional growth, productivity, and metabolomic data were obtained from Enbrel and SERPING1-producing CHO cells, and models were constructed. **(B)** The model-predicted growth rates were consistent with experimental growth rates of of Enbrel-producing CHO cells and **(C)** SERPING1-producing CHO cells at almost all time points. Error bars in B and C represent the standard deviation of three biological replicates. In all cases, the iCHO2048s models were constrained to produce the recombinant protein at the measured specific productivity rate. The values used to constrain each of the iCHO2048s models are reported in Supp. File 3.

3.3 Protein composition and complexity significantly impact model-predicted productivity

To produce a specific product, CHO cells may utilize different modules of the secretory pathway based on the protein attributes and post-translational modifications (PTMs). For example, the synthesis of a mAb requires the use of multiple processes and consumes several different metabolites, such as amino acids for protein translation, redox equivalents for forming disulfide bonds, ATP equivalents for vesicular transport, and sugar nucleotides for protein glycosylation (Fig. 1C). Therefore, we set out to generate eight product-specific secretory pathway models for biotherapeutics commonly produced in CHO cells (Fig. 3A): bone morphogenetic proteins 2 and 7 (BMP2, BMP7), erythropoietin (EPO), Etanercept, factor VIII (F8), interferon beta 1a (IFNB1), Rituximab, and tissue plasminogen activator (tPA). The resulting iCHO2048s models were used to compute Pareto optimality frontiers between maximum cell growth (μ) and specific productivity (q_P) assuming all eight CHO cells grow under the same conditions. That is, all eight models were given the same measured glucose and amino acid uptake rates[23] as model constraints (see Supp. File 3)

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

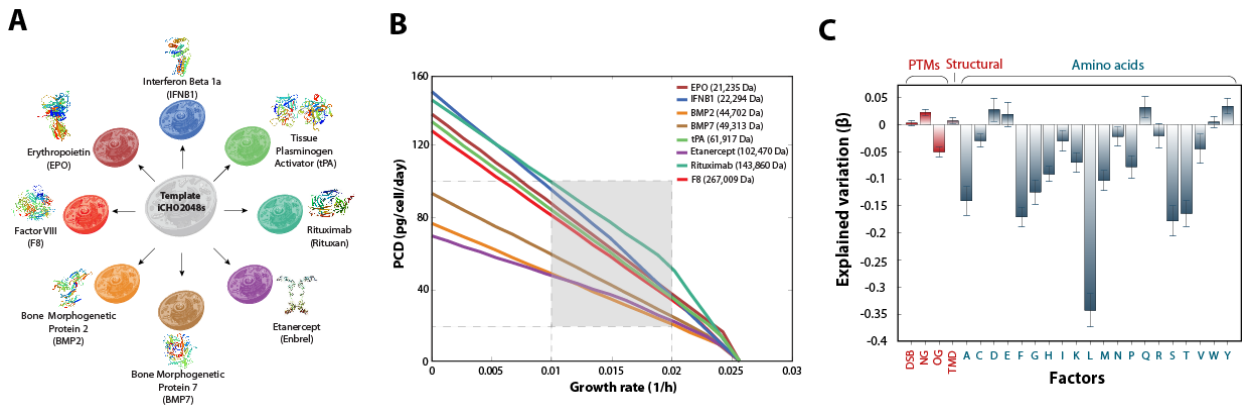


Figure 3 - Construction of product-specific iCHO2048s models. (A) Eight product-specific iCHO2048s models were constructed for biotherapeutics commonly produced in CHO cells. **(B)** Pareto optimality frontiers of growth/productivity (μ/q_p) trade-off curves were computed for the eight iCHO2048s models using the same constraints and experimental data from Supp. File 3. The shaded region corresponds to range of maximum productivity at commonly observed growth rates in CHO cell cultures. The molecular weight (in Daltons) of each biotherapeutic is shown in the legend. **(C)** All protein features (PTMs, transmembrane domains, and amino acid compositions) were analyzed to quantify their contribution to the explained variation of specific productivity.

We computed the tradeoff between growth (μ; inverse hours or 1/h) and specific productivity (q_p; pg protein produced per cell per day or PCD) as a Pareto optimal "μ/q_p curve" for each protein (Fig. 2B). This curve defines the frontier of maximum specific productivity and maximum growth rates under the assumption that CHO cells can utilize all available resources towards production of biomass and recombinant protein only. The hinges in some of the μ/q_p curves are indicative of a transition between regions in the μ/q_p that are limited by distinct protein requirements (e.g., amino acids).

An analysis of the μ/q_p curves for the eight biotherapeutics demonstrates that under the measured growth conditions, maximum productivities vary from 20-100 PCD at common growth rates (Fig. 2B, shaded region) to 70-150 PCD for senescent CHO cells. Neither the molecular weight (MW) nor product length can explain the 2-fold range differences in maximum productivity for different proteins. For example, the μ/q_p curves show tPA (MW = 61,917 Da) can express at higher PCD than BMP2 (MW = 44,702 Da) despite being larger, because the N-glycans in BMP2 reduce productivity due to the higher cost of synthesizing core N-glycans (see Table 1), consistent with previous observations in yeast[5]. Furthermore, the degree and directionality of these effects will depend on the nutrient uptake rates,

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

highlighting the need in CHO bioprocessing to tailor culture media in a host cell and product-specific manner. Thus, while intuitively larger proteins would be expected to exert more bioenergetic cost on protein secretion, we find that specific compositional attributes of both the recombinant protein and the culture media significantly impact biosynthetic capacity.

Table 1. Protein specific information matrix of biotherapeutics secreted in eight iCHO2048s models

Protein Name	Total number of amino acids in biotherapeutic	Molecular Weight [Da]	Total number of disulfide bonds in mature protein	Total number of N-glycans in mature protein	Total number of O-glycans in mature protein	Estimated secretory cost [ATP equivalents]
IFNB1	187	22294	1	1	0	777
EPO	193	21037	2	3	1	801
BMP2	396	44702	4	5	0	1618
BMP7	431	49313	4	4	0	1759
tPA	562	61917	17	3	1	2286
Etanercept**	934	102470	7	6	26	3784
Rituximab*	1328	143860	17	2	0	5370
F8	2351	267009	8	22	0	9488

** = Rituximab is a tetramer (2 light and 2 heavy chains)*

*** = Etanercept is a dimer*

3.4 iCHO2048s accurately predicts protein productivity increase following gene knock-down

Kallehauge et al.[19] demonstrated that a CHO-DG44 cell line producing an antiviral mAb[28] also expressed high levels of the neoR selection-marker gene (Fig. 4A-B). Upon neoR knockdown, the titer and maximum viable cell densities of the CHO-DG44 cell line were increased. To test if iCHO2048s could replicate these results, we constructed a model for the Kallehauge et al. DG44 cell line and measured exometabolomics, and dry cell weight to parameterize the model. Since expression of neoR uses resources that could be used for antibody production, we predicted how much additional antibody could be synthesized with the elimination of the neoR gene. We simulated antibody production following a complete knockout of neoR (see Table 2 and Fig. 4B) and predicted that the deletion of neoR could increase specific productivity by up to 4% and 29% on days 3 (early exponential phase) and 6 (late phase) of culture, respectively (Fig. 4C). This was consistent with the experimentally observed values of 2% and 14%. We then computed the μ/q_p curves for both the control and the neoR *in silico* knockout conditions on day 6. We found that the length of the μ/q_p curve (i.e. the size of the set of Pareto optimal

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

flux distributions, here denoted by delta (Δ)) increased by 18% percent when neoR production is eliminated (Fig. 4D). Thus, iCHO2048s can quantify how much non-essential gene knockouts can boost growth and productivity in CHO cells by freeing energetic and secretory resources. In fact, the ribosome-profiling data from Kallehauge et al. revealed that only 30 secretory proteins in CHO cells account for more than 50% of the ribosomal load directed towards translation of protein bearing a signal peptide (Fig. 4E). An analysis of other potential host cell gene knockouts using the method proposed here can be found in Supp. File 5.

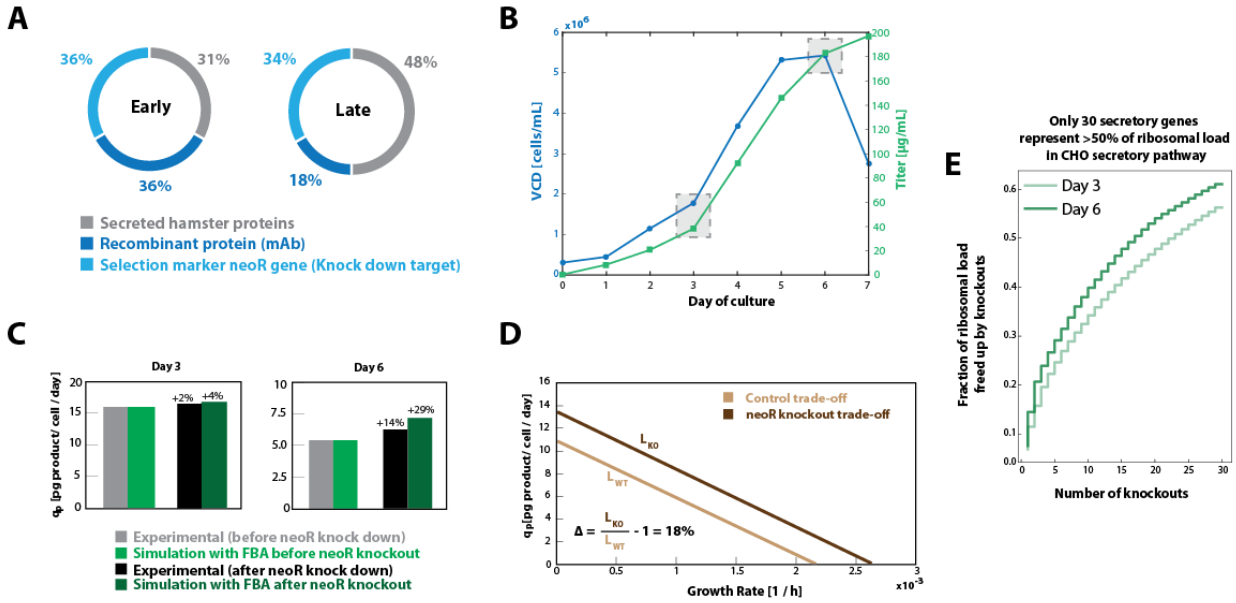


Figure 4 - iCHO2048s recapitulates experimental results of neoR knock-down *in silico*. (A) Ribosome occupancy was measured with ribosomal profiling during early (left) and late (right) exponential growth phases[19]. **(B)** Time profiles are shown for viable cell density (VCD) and titer in experimental culture. Shaded boxes indicate the time points corresponding to early (day 3) and late (day 6) growth phases. **(C)** Flux balance analysis was used to predict specific productivity (q_p) with the iCHO2048s model before and after *in silico* knockout of neoR gene. **(D)** Growth/productivity (μ/q_p) trade-offs were predicted by iCHO2048s and demonstrated a potential 18% increase after the neoR *in silico* knockout. The formula for calculating the trade-off improvement (Δ) is shown in the plot. L_{WT} = length of μ/q_p curve before knockout, L_{KO} = length of μ/q_p curve after knockout. **(E)** Ribosmal occupancy for all mRNA sequences bearing a signal peptide sequence were analyzed from the Kallehauge et al. study, and demonstrated that the top 30 secreted proteins accounted >50% of the ribosomal occupancy of secreted proteins.

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2 - Experimental data from Kallenhauge et al. used for testing predictive capabilities of iCHO2048s

Experimental value description	Day 3 (early exponential growth phase)	Day 6 (late growth phase)
Growth Rate [1/day]	0.44	0.02
*Specific Productivity [Picograms of IgG/cell/day]	16	5.5
**Total IgG ribosomal footprint [RPKM]	40258	13356
Total neoR ribosomal footprint [RPKM]	36952	25679

* Average cell dry weight = 456.3 pg/cell

** Sum of light and heavy chains ribosomal footprints

3.5 CHO cells have suppressed expression of expensive proteins in their secretome

In any cell, the secretory machinery is concurrently processing thousands of secreted and membrane proteins, which all compete for secretory pathway resources and pose a metabolic burden. To quantify this burden, we estimated the energetic cost of synthesizing and secreting all 5,641 endogenous proteins in the CHO secretome and membrane proteome in terms of total number of ATP equivalent molecules consumed. These protein costs were compared to the cost of the eight recombinant proteins previously analyzed. To refine estimates, we experimentally measured the number of N-linked glycans in the CHO proteome and integrated published numbers of O-linked glycans in CHO proteomic data[9]. Across the CHO secretome, protein synthesis cost varies substantially, and recombinant products are on average more expensive (Fig. 5A). For example, F8 is a “difficult-to-express” protein in CHO cells due to its propensity to aggregate in the ER, which promotes its premature degradation[29,30]. Our analysis further highlights that each molecule of F8 requires an excessive amount of ATP for its production (9488 ATP molecules). This imposes a significant burden to the secretory machinery of CHO cells, which typically expresses much less expensive proteins. With the broad range of biosynthetic costs for different proteins, we wondered if gene expression in CHO cells has been influenced by the ATP cost of secreted proteins, by suppressing host cell protein expression to more efficiently allocate nutrients. Unless specific proteins are essential, CHO cells may preferentially suppress energetically expensive proteins. To test this, we analyzed ribosomal profiling (Ribo-seq) data from a mAb-producing CHO cell line[19] and compared translation of each transcript against the ATP cost of the associated secreted protein (see Methods). Indeed, there was a significant negative correlation of -0.43 and -0.36 (Spearman

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

R_s, p value < $1x10^{-20}$) between ribosomal occupancy and ATP cost during early and late phases of culture, respectively (Fig. 5B). Wondering if the reduced translation was regulated transcriptionally, we further analyzed RNA-seq data from the same mAb-producing cell line and from another, non-producing CHO-K1 cell line[31]. The RNA expression also negatively correlated with ATP cost (see Supp. Figure 2). Finally, we analyzed RNA-seq data from human tissues and immortalized cell lines in the Human Protein Atlas (HPA)[1]. All RNA-seq datasets in the HPA samples also negatively correlated with ATP cost (Supp. Figure 3). Interestingly, we found that highly secretory tissues such as liver, pancreas and salivary gland had the strongest correlations, although none as strong as that of the mAb-producing CHO cells (Supp. Figure 3a). Feizi and colleagues recently found that these tissues fine-tune the expression of protein disulfide isomerase genes[32], suggesting that a similar regulatory process may take place in the ER of CHO cells as the secreted mAb contains a relatively high number (17) of disulfide bonds. In conclusion, there is a clear preference in CHO cells to suppress the expression and translation of proteins that are costly to synthesize, fold, and secrete.

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

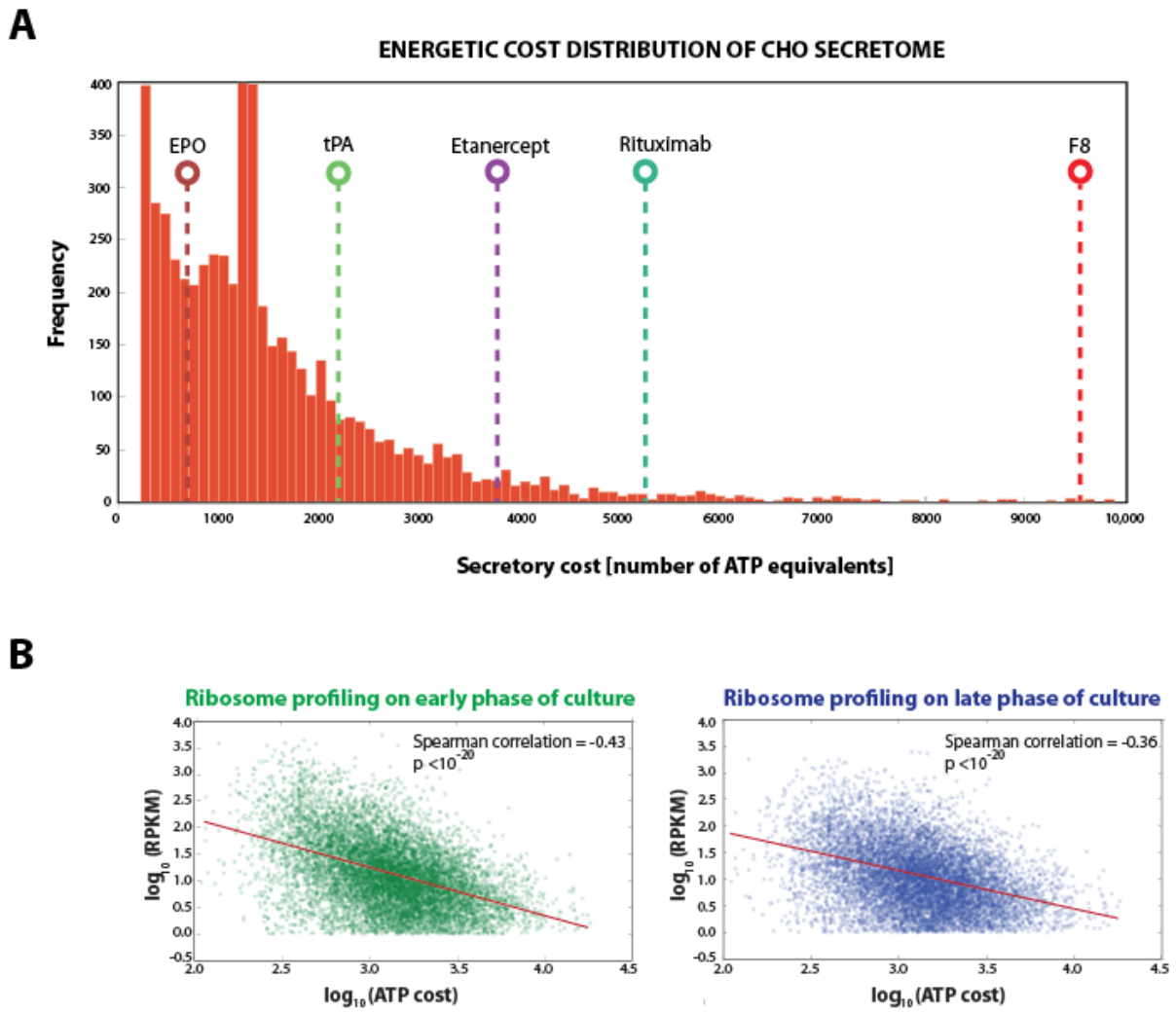


Figure 5 – CHO cells preferentially suppress more expensive proteins. (A) The bioenergetic cost of each secreted CHO protein was computed. 28 of the 5641 proteins in CHO secretome had a cost > 10,000 ATP equivalents and were therefore not included in the histogram for the sake of ease of visualization (these 28 proteins are listed in Supp. Notebook 3). The biosynthetic costs of 5 representative biotherapeutics are shown for comparison purposes (see Table 1). **(B)** Scatter plots and Spearman correlation of gene expression and protein cost (in number of ATP per protein) from Kallehauge et al.[19] during early (left) and late (right) phases of culture. RPKM = reads per kilobase of transcript per million.

4. Discussion

Mammalian cells synthesize and process thousands of proteins through their secretory pathway. Many of these proteins, including hormones, enzymes, and receptors, are essential for mediating mammalian cell interactions with their environment. Therefore, many have therapeutic importance either as drugs

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

or as targets. The expression and secretion of recombinant proteins represents a significant anabolic demand that drains several substrates from cellular metabolism (e.g., amino acids, sugar nucleotides, ATP)[33,34]. Furthermore, the recombinant proteins demand adequate expression of proteins involved in their transcription, translation, folding, modification, and secretion. Thus, there has been an increasing interest in engineering the mammalian secretory pathway to boost protein production[35–38]. Despite important advances in the field[24], current strategies to engineer the secretory pathway have remained predominantly empirical[39,40]. Recent modeling approaches, however, have enabled the analysis of the metabolic capabilities of important eukaryotic cells (including CHO) under different genetic and environmental conditions[23,41–43]. With the development of genome-scale models of protein-producing cells, such as the genome-scale model of CHO cell metabolism[23], it is now possible to gain a systems-level understanding of the CHO phenotype[44].

Efforts have been underway to enumerate the machinery needed for protein production. For example, Lund and colleagues[45] recently reconstructed the most comprehensive network of the mouse secretory pathway to date. By comparing the mouse and CHO-K1 genomes and mapping CHO gene expression data onto this network, the authors identified potential targets for CHO cell engineering, demonstrating the great potential of systems biology to interrogate and understand protein secretion in animal cells. This network reconstruction, although useful for contextualizing omics data (e.g., RNA-seq), is not set up for simulations of protein production, nor integrated with additional cellular processes such as metabolism. Therefore, to quantify the cost and cellular capacity for protein production, it is important to delineate the mechanisms of all biosynthetic steps and bioenergetic processes in the cell.

Here we have presented the first genome-scale reconstruction of the secretory pathway in mammalian cells. We connected this to current metabolic networks, yielding models of protein secretion and metabolism for human, mouse and CHO cells. These models compile decades of research in biochemistry and cell biology of higher eukaryotes and present it in a mathematical model. Using our model, we quantitatively estimated the energetic cost of producing several therapeutic proteins and all proteins in the CHO cell and human secretomes. We also identified factors limiting the secretion of individual products and observed that these depend on both the complexity of the product and the composition of the culture media. Furthermore, by integrating ribosomal profiling data with our model we found that CHO cells have selectively suppressed the expression of energetically expensive secreted

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

proteins. Expanding upon this observation, we demonstrated that specific productivities can be predictably increased following the knock-down of an energetically expensive, non-essential protein.

It is important to note that while our models capture major features of secreted proteins, there are additional PTMs (e.g., phosphorylation, gamma carboxylation), pathway machinery (e.g., chaperones), and cell processes that could possibly be captured in further expansions of the modeling framework[45] (e.g, the unfolded protein response). These could be included as energetic costs associated with building and maintaining the secretory machinery (chaperones[3], glycosyltransferases[46]); protein stability and turnover rates[47]; solubility constraints[48] and molecular crowding effects[49]. As these are captured by the models in a protein product-specific manner, predictions of protein production capacity will improve, and the models could provide further insights for cell engineering for biotechnology or to obtain a deeper understanding of mechanisms underlying amyloid diseases. Finally, a simplification of our secretory model is that it only computes the bioenergetic cost of synthesizing and attaching single representative N- and O-linked glycans to secreted proteins (i.e., it does not include the microheterogeneity and diversity of glycan structures of different proteins). Thus, an immediate potential expansion of our secretory model would involve coupling it to existing computational models of protein glycosylation (recently reviewed by Spahn and Lewis[50]). For example, given an N-glycan reaction network that captures the glycoform complexity of a target protein[51], one could build secretory reactions for the specific glycoforms of interest and compute the metabolic demands associated with each of them as to identify potential targets and nutrient supplementations for glycoengineering.

In conclusion, the results of our study have important implications regarding the ability to predict protein expression based on protein specific attributes and energetic requirements. The secretory pathway models here stand as novel tools to study mammalian cells and the energetic trade-off between growth and protein secretion in a product- and cell-specific manner. We presented algorithms that provide novel insights with our models, and expect that many other methods can be developed to answer a wide array of questions surrounding the secretory pathway, as seen for metabolism[52]. To facilitate further use of these models, we provide our code and detailed instructions on how to construct protein-specific models in the Supp. Jupyter Notebooks available at https://github.com/LewisLabUCSD.

Acknowledgements

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The authors would like to thank Philipp Spahn, Austin Chiang, and Chih-Chung Kuo for their insightful comments on this manuscript. This work was supported by generous funding from the Novo Nordisk Foundation provided to the Center for Biosustainability at the Technical University of Denmark (NNF10CC1016517 and NNF16CC0021858), and from NIGMS (R35 GM119850), and a fellowship from the Government of Mexico (CONACYT) and the University of California Institute for Mexico and the United States (UC-MEXUS).

References

1. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347,** (2015).
2. Walsh, G. Biopharmaceutical benchmarks 2014. *Nat. Biotechnol.* **32,** 992–1000 (2014).
3. Feizi, A., Österlund, T., Petranovic, D., Bordel, S. & Nielsen, J. Genome-Scale Modeling of the Protein Secretory Machinery in Yeast. *PLoS One* **8,** e63284 (2013).
4. Liu, L., Feizi, A., Österlund, T., Hjort, C. & Nielsen, J. Genome-scale analysis of the high-efficient protein secretion system of Aspergillus oryzae. *BMC Syst. Biol.* **8,** 73 (2014).
5. Irani, Z. A., Kerkhoven, E. J., Shojaosadati, S. A. & Nielsen, J. Genome-scale metabolic model of Pichia pastoris with native and humanized glycosylation of recombinant proteins. *Biotechnol. Bioeng.* **113,** 961–969 (2016).
6. Araki, K. & Inaba, K. Structure, mechanism, and evolution of Ero1 family enzymes. *Antioxid. Redox Signal.* **16,** 790–9 (2012).
7. Hiller, K., Grote, A., Scheer, M., Munch, R. & Jahn, D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* **32,** W375–W379 (2004).
8. Fankhauser, N. & Maser, P. Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics* **21,** 1846–1852 (2005).
9. Yang, Z. *et al.* The GalNAc-type O-Glycoproteome of CHO cells characterized by the SimpleCell strategy. *Mol. Cell. Proteomics* **13,** 3224–35 (2014).
10. Baycin-Hizal, D. *et al.* Proteomic Analysis of Chinese Hamster Ovary Cells. *J. Proteome Res.* **11,** 5265–5276 (2012).
11. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28,** 245–8 (2010).
12. Schellenberger, J. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* **6,** 1290–1307 (2011).
13. Grainger, R. K. & James, D. C. CHO cell line specific prediction and control of recombinant

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

monoclonal antibody *N* -glycosylation. *Biotechnol. Bioeng.* **110,** 2970–2983 (2013).

14. Araki, K. & Nagata, K. Protein Folding and Quality Control in the ER. *Cold Spring Harb. Perspect. Biol.* **3,** a007526–a007526 (2011).

15. Borner, G. H. H. *et al.* Multivariate proteomic profiling identifies novel accessory proteins of coated vesicles. *J. Cell Biol.* **197,** 141–160 (2012).

16. Cheng, Y., Boll, W., Kirchhausen, T., Harrison, S. C. & Walz, T. Cryo-electron Tomography of Clathrin-coated Vesicles: Structural Implications for Coat Assembly. *J. Mol. Biol.* **365,** 892–899 (2007).

17. Takamori, S. *et al.* Molecular Anatomy of a Trafficking Organelle. *Cell* **127,** 831–846 (2006).

18. Liu, J. K. *et al.* Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. *BMC Syst. Biol.* **8,** 110 (2014).

19. Kallehauge, T. B. *et al.* Ribosome profiling-guided depletion of an mRNA increases cell growth rate and protein secretion. *Sci. Rep.* **7,** 40388 (2017).

20. Wang, Y., Eddy, J. A. & Price, N. D. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst. Biol.* **6,** 153 (2012).

21. Opdam, S. *et al.* A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. *Cell Syst.* **4,** 318–329.e6 (2017).

22. González-Leal, I. J. *et al.* Use of a Plackett-Burman statistical design to determine the effect of selected amino acids on monoclonal antibody production in CHO cells. *Biotechnol. Prog.* **27,** 1709–1717 (2011).

23. Hefzi, H. *et al.* A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. *Cell Syst.* **3,** 434–443.e8 (2016).

24. Golabgir, A. *et al.* Quantitative feature extraction from the Chinese hamster ovary bioprocess bibliome using a novel meta-analysis workflow. *Biotechnol. Adv.* **34,** 621–633 (2016).

25. Swainston, N. *et al.* Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* **12,** 109 (2016).

26. Sigurdsson, M. I., Jamshidi, N., Steingrimsson, E., Thiele, I. & Palsson, B. Ø. A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. *BMC Syst. Biol.* **4,** 140 (2010).

27. Templeton, N., Dean, J., Reddy, P. & Young, J. D. Peak Antibody Production is Associated With Increased Oxidative Metabolism in an Industrially Relevant Fed-Batch CHO Cell Culture. *Biotechnol. Bioeng* **110,** 2013–2024 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28. Kim, S. J., Kim, N. S., Ryu, C. J., Hong, H. J. & Lee, G. M. Characterization of chimeric antibody producing CHO cells in the course of dihydrofolate reductase-mediated gene amplification and their stability in the absence of selective pressure. *Biotechnol. Bioeng.* **58,** 73–84 (1998).

29. Kaufman, R. J. *et al.* Effect of von Willebrand Factor Coexpression on the Synthesis and Secretion of Factor VIII in Chinese Hamster Ovary Cells. *Mol. Cell. Biol.* **9,** 1233–1242 (1989).

30. Pipe, S. W., Morris, J. A., Shah, J. & Kaufman, R. J. Differential interaction of coagulation factor VIII and factor V with protein chaperones calnexin and calreticulin. *J. Biol. Chem.* **273,** 8537–44 (1998).

31. van Wijk, X. M. *et al.* Whole-Genome Sequencing of Invasion-Resistant Cells Identifies Laminin α2 as a Host Factor for Bacterial Invasion. *MBio* **8,** e02128-16 (2017).

32. Feizi, A., Gatto, F., Uhlen, M. & Nielsen, J. Human protein secretory pathway genes are expressed in a tissue-specific pattern to match processing demands of the secretome. *npj Syst. Biol. Appl.* **3,** 22 (2017).

33. Gu, M. B., Todd, P. & Kompala, D. S. Metabolic burden in recombinant CHO cells: effect of dhfr gene amplification and lacZ expression. *Cytotechnology* **18,** 159–166 (1996).

34. Gu, M. B., Todd, P. & Kompala, D. S. Analysis of foreign protein overproduction in recombinant CHO cells. Effect of growth kinetics and cell cycle traverse. *Ann. N. Y. Acad. Sci.* **721,** 194–207 (1994).

35. Hansen, H. G., Pristovšek, N., Kildegaard, H. F. & Lee, G. M. Improving the secretory capacity of Chinese hamster ovary cells by ectopic expression of effector genes: Lessons learned and future directions. *Biotechnol. Adv.* **35,** 64–76 (2017).

36. Delic, M., Göngrich, R., Mattanovich, D. & Gasser, B. Engineering of Protein Folding and Secretion—Strategies to Overcome Bottlenecks for Efficient Production of Recombinant Proteins. *Antioxid. Redox Signal.* **21,** 414–437 (2014).

37. Le Fourn, V., Girod, P.-A., Buceta, M., Regamey, A. & Mermod, N. CHO cell engineering to prevent polypeptide aggregation and improve therapeutic protein secretion. *Metab. Eng.* **21,** 91–102 (2014).

38. Kuo, C. C. *et al.* The emerging role of systems biology for engineering protein production in CHO cells. *Curr. Opin. Biotechnol.* **51,** 64–69 (2018).

39. Borth, N., Mattanovich, D., Kunert, R. & Katinger, H. Effect of Increased Expression of Protein Disulfide Isomerase and Heavy Chain Binding Protein on Antibody Secretion in a Recombinant CHO Cell Line. *Biotechnol. Prog.* **21,** 106–111 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

40. Ku, S. C. Y., Ng, D. T. W., Yap, M. G. S. & Chao, S.-H. Effects of overexpression of X-box binding protein 1 on recombinant protein production in Chinese hamster ovary and NS0 myeloma cells. *Biotechnol. Bioeng.* **99,** 155–164 (2008).

41. Yusufi, F. N. K. *et al.* Mammalian Systems Biotechnology Reveals Global Cellular Adaptations in a Recombinant CHO Cell Line. *Cell Syst.* **4,** 530–542.e6 (2017).

42. Selvarasu, S. *et al.* Combined in silico modeling and metabolomics analysis to characterize fed-batch CHO cell culture. *Biotechnol. Bioeng.* **109,** 1415–1429 (2012).

43. Gutierrez, J. M. & Lewis, N. E. Optimizing eukaryotic cell hosts for protein production through systems biotechnology and genome-scale modeling. *Biotechnol. J.* **10,** 939–49 (2015).

44. Galleguillos, S. N. *et al.* What can mathematical modelling say about CHO metabolism and protein glycosylation? *Comput. Struct. Biotechnol. J.* **15,** 212–221 (2017).

45. Lund, A. M. *et al.* Network reconstruction of the mouse secretory pathway applied on CHO cell transcriptome data. *BMC Syst. Biol.* **11,** 37 (2017).

46. Jimenez Del Val, I., Polizzi, K. M. & Kontoravdi, C. A theoretical estimate for nucleotide sugar demand towards Chinese Hamster Ovary cellular glycosylation. *Sci. Rep.* **6,** (2016).

47. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. O. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9,** 693–693 (2014).

48. Vazquez, A. & Oltvai, Z. N. Macromolecular crowding explains overflow metabolism in cells. *Sci. Rep.* **6,** (2016).

49. Beg, Q. K. *et al.* Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 12663–8 (2007).

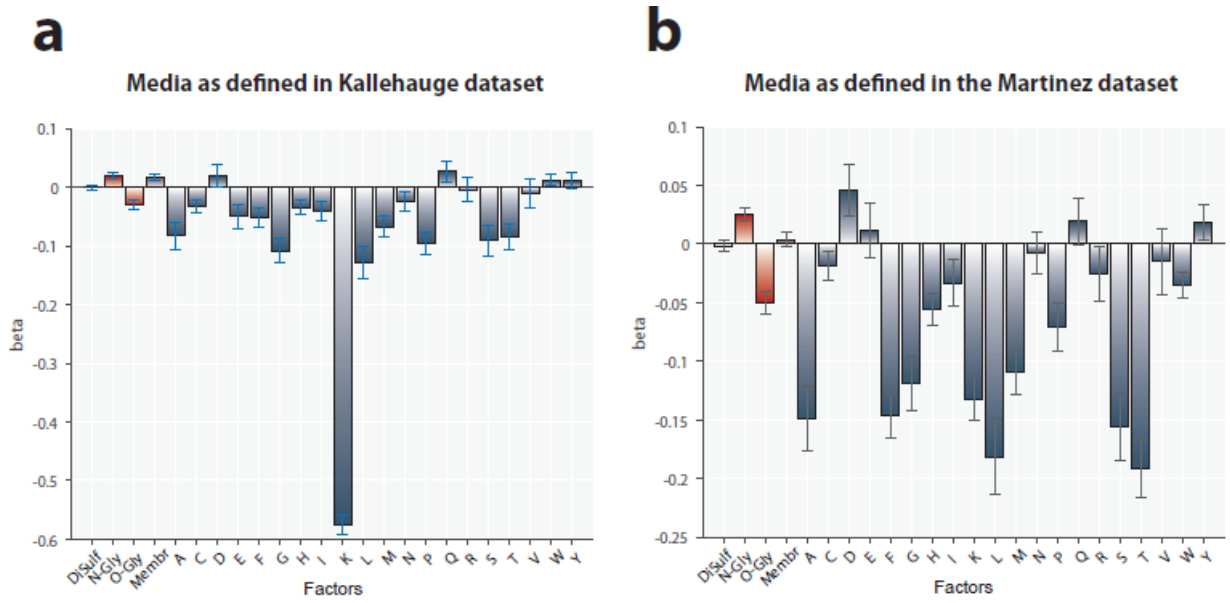
50. Spahn, P. N. & Lewis, N. E. Systems glycobiology for glycoengineering. *Curr. Opin. Biotechnol.* **30C,** 218–224 (2014).

51. Spahn, P. N. *et al.* A Markov chain model for N-linked protein glycosylation - towards a low-parameter tool for model-driven glycoengineering. *Metab. Eng.* **33,** 52–66 (2016).

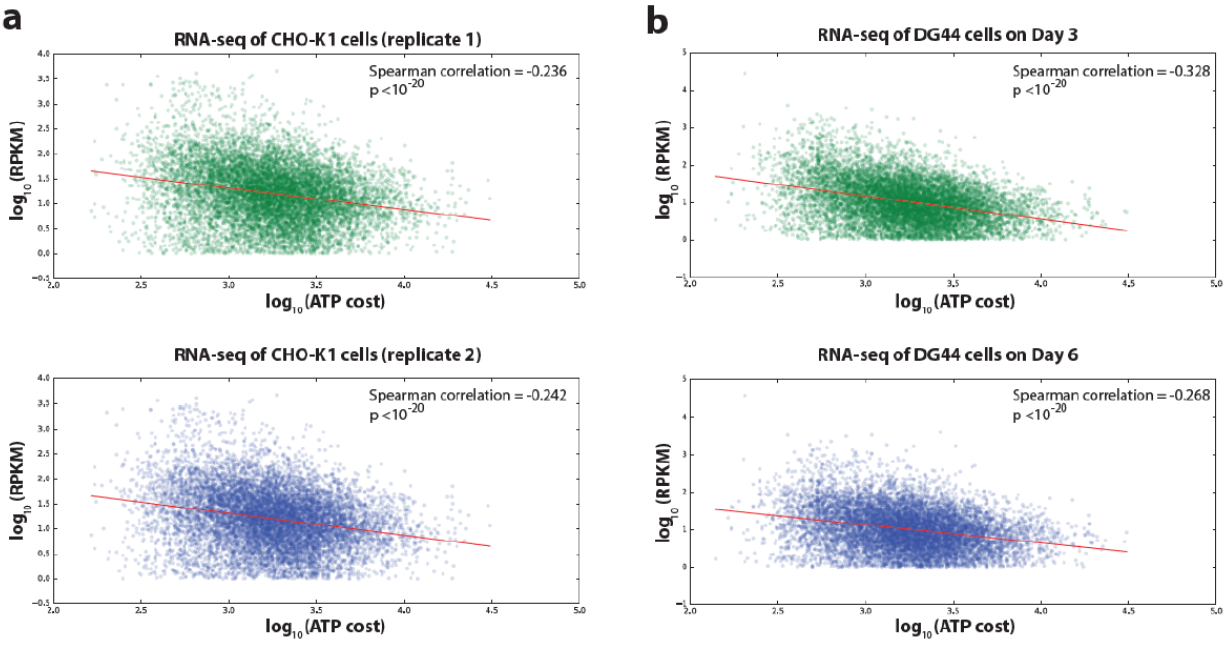
52. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* **10,** 291–305 (2012).

53. Martínez, V. S., Buchsteiner, M., Gray, P., Nielsen, L. K. & Quek, L.-E. Dynamic metabolic flux analysis using B-splines to study the effects of temperature shift on CHO cell metabolism. *Metab. Eng. Commun.* **2,** 46–57 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

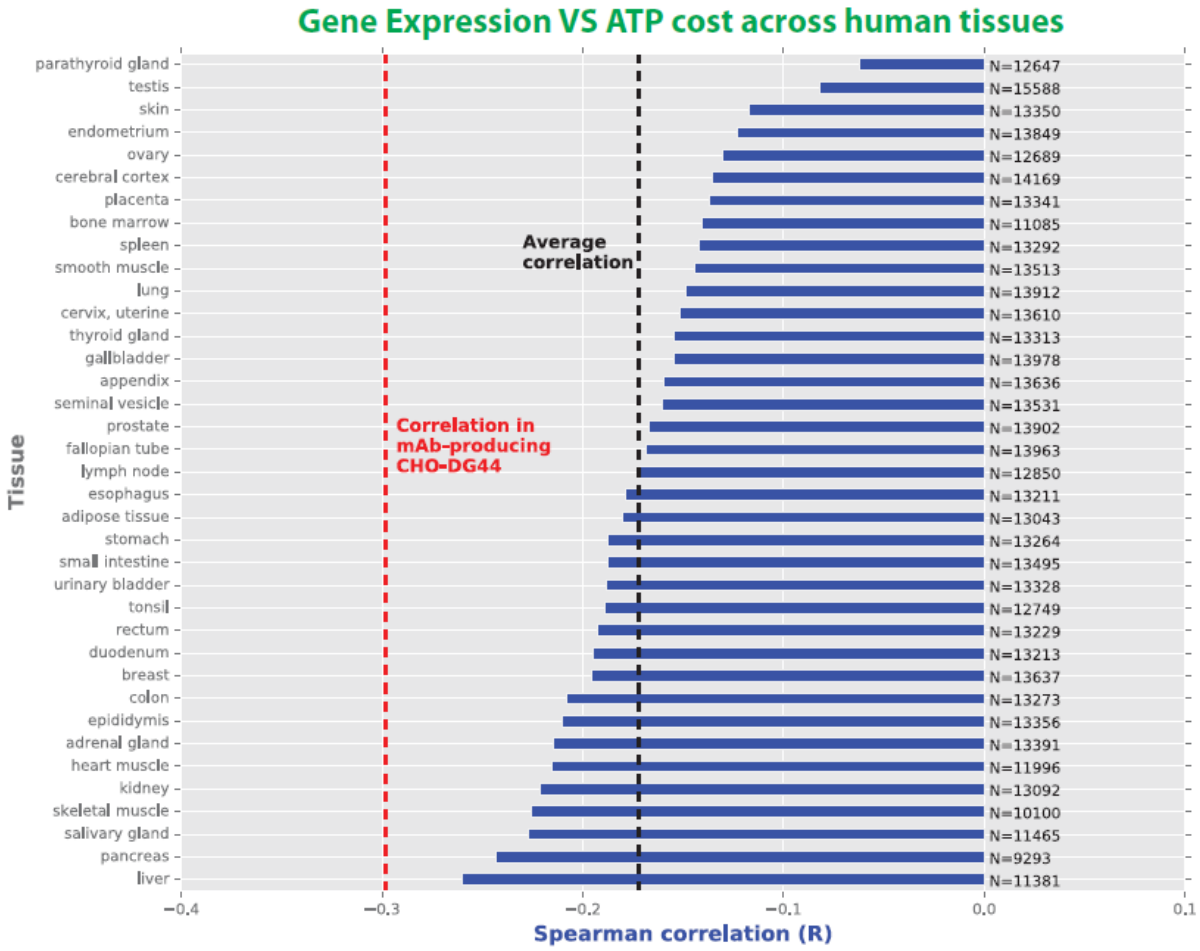


Supp. Figure 1 - Factors affecting iCHO2048s-predicted productivity with two different media compositions. The specific consumption rates are listed in Supp. File 3 as Kallehauge[19] (left panel) and Martinez[53] (right panel).



Supp. Figure 2 – Spearman correlation between ATP cost and gene expression levels in non-producing CHO-K1 and CHO-DG44 mAb-producing cells. Gene transcription levels from **(a)** van Wijk et al.[31] and **(b)** Kallehauge et al.[19] were compared against the ATP cost of producing the translated proteins.

bioRxiv preprint doi: https://doi.org/10.1101/351387; this version posted June 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supp. Figure 3 – Spearman correlations between ATP cost and gene expression levels across human tissues. Gene transcription levels from the Human Protein Atlas[1]were analyzed against the ATP cost of producing the translated proteins. The average of all Spearman correlations across samples in indicated by the black dashed line while the red dashed line indicates the correlation found using RNA-seq data in Kallehauge et al[19]. The number of genes used to compute the correlations in each sample is indicated next to the bars. All p-values associated to each correlation are less than $1x10^{-20}$