Improving Glycosylation Efficiency in *Escherichia coli* through Model-Guided Metabolic Engineering

Joseph A. Wayman[1], Thomas J. Mansell[2], Matthew P. DeLisa[3], and Jeffrey D. Varner[4*]

[1*]School of Applied and Engineering Physics, Cornell University, Ithaca, NY 14853

[2*]School of Chemical and Biological Engineering, Iowa State University, Ames, IA 50011

[3*]School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY 14853

[4*]School of Chemical Engineering, Purdue University, West Lafayette, IN 47907

Running Title: Model-guided glycoengineering in *E. coli*

To be submitted: *Metabolic Engineering*

[*]Corresponding author:

Jeffrey D. Varner,

Professor, School of Chemical Engineering,

Forney Hall, 480 Stadium Mall Drive, Purdue University, West Lafayette, IN 47907

Email: jdvarner@purdue.edu

Phone: (765) 496 - 0544

Fax: (765) 494 - 0805

Abstract

Abstract goes here ...

Introduction

Asparagine-linked (N-linked) glycosylation is the most common protein modification in eukaryotes, affecting over two-thirds of the proteome. N-linked glycans are complex, branched oligosaccharides, assembled on lipid carriers in the endoplasmic reticulum membrane and transferred to specific asparagine residues of acceptor proteins. Approximately 70% of therapeutic proteins require the attachment of complex branches of sugars to specific amino acid residues [1]. This common post-translational modification affects various protein properties including pharmacokinetic activity and immunogenicity [2]. Currently, eukaryotes possessing native glycosylation machinery serve as the preferred production host of therapeutic glycoproteins. Eukaryotic production hosts suffer from several limitations including slow growth and a susceptibility to viral and bacterial infection. Also, eukaryotes produce a variety of glycan structures and glycosylate proteins with a range of site occupancy, making purification of the desired glycoform difficult [1]. Though once thought only to occur in eukaryotes, protein glycosylation has been discovered in all other domains of life, including bacteria, spurring interest in the development of alternative glycoprotein expression platforms. The most well-characterized bacterial glycosylation system is that of the human pathogen *Campylobacter jejuni* [3]. The *C. jejuni* glycan has the form of a branched heptasaccharide Glc$_2$GalNAc$_5$Bac, where Glc is glucose, GalNAc is N-acetylgalactosamine, and Bac is bacillosamine. This glycan is assembled on the lipid carrier undecaprenyl pyrophosphate (UDCP) on the cytoplasmic face of the inner membrane by an enzyme pathway encoded by the *pgl* (protein glycosylation) genetic locus (Fig. 1). The fully assembled glycan is flipped across the membrane and transferred to asparagine residues on acceptor proteins by an oligosaccharyltransferase (OST) called PglB. PglB attaches the heptasaccharide to periplasm-localized proteins containing the consensus sequence D/E-X-N-X-S/T, where X is any residue except proline [4]. The functional transfer of this system into *E. coli* has spurred interest in producing non-native glycans in a more genetically tractable host [3, 5].

By now, production of a variety of periplasmic, extracellular, and secretory proteins has been demonstrated in glycosylation-competent *E. coli* [4]. Producing glycoprotein from prokaryotic hosts continues to suffer from several limitations including poor glycosylation efficiency and insufficient yield. The highest reported efficiency, percent of acceptor protein glycosylated, using the *C. jejuni* system has been 47% [6]. Synthesis of more complex, highly branched glycan, like those found in humans, has proved challenging. Recently, a key step in human-like glycan production was achieved in *E. coli* with the recombinant expression of a synthetic glycosylation pathway able to assemble the tri-mannose core glycan [7]. The pathway yielded approximately 50 μg/L of glycosylated protein with an efficiency of less than 1%. Though an important step in the development of *E. coli* as a viable glycoprotein production host, optimization of this system remains a key challenge. Factors affecting protein glycosylation in *E. coli* may include the expression and activity of glycosylation pathway enzymes, the availability of lipid carrier sites, and the availability of nucleotide-activated sugar substrates serving as glycan precursors [1, 8]. Wright and coworkers have applied genome-scale metabolic engineering techniques toward the improvement of glycosylation efficiency in *E. coli*. Using a high-throughput proteomic screening and probabilistic metabolic network analysis, they showed that upregulation of the glyoxylate cycle by overexpression of isocitrate lyase (*aceA*/*icl*) led to a three-fold increase in glycosylation efficiency of a prototypic protein [6]. Also, a genome-wide screening of gene overexpression identified targets that led to increased glycoprotein production as well as glycosylation efficiency [9]. Genes in pathways associated with glycan precursor synthesis (UDP-GlcNAc) as well as lipid carrier production (isoprenoid synthesis) were identified as bottlenecks. Improving expression of the OST PglB by codon optimization was also shown to improve glycosylation efficiency [10]. These studies demonstrate the complex interplay between recombinant protein production, glycan synthesis and assembly, and glycosylation efficiency. Increasing glycoprotein production in *E. coli* demands the simultaneous optimization of competing metabolic functions. Recombinant protein production

requires immense energy from catabolic processes while glycan precursor synthesis requires anabolic processes.

In this study, we use an adapted genome-scale constraint-based model of *E. coli* metabolism to design gene knockout strains that increase glycosylation efficiency by overproduction of glycan precursors. First, we incorporate reactions associated with *C. jejuni* glycan assembly into a genome-scale model of *E. coli* metabolism. We use a combination of constraint-based modeling and heuristic optimization in order to identify gene knockout strains that couple optimal growth to glycan synthesis. Simulations identify growth-coupled strains flux analysis unveils modes of metabolite imbalance that reroute flux toward glycan precursor synthesis. Experimentally, measuring fluorescently-labeled, cell surface-displayed glycan expression, we show that model-identified knockout strains increase glycan synthesis in *E. coli*. We also show an increase in glycosylation efficiency of a prototypical acceptor protein. This study demonstrates the promising role metabolic modeling can play in optimizing the performance of a next generation microbial glycosylation platform.

Results

Construction of a constraint-based model of glycosylation in *E. coli* We used constraint-based modeling of glycosylation in *E. coli* to identify genetic knockouts that promote biosynthesis of glycan precursors. We modified the existing genome-scale *E. coli* model iAF1260 from Palsson and coworkers [11] to include the reactions of the *C. jejuni* glycosylation pathway listed in Table 1. The adapted network consists of 2395 reactions, 1271 open reading frames, and 1986 metabolites segregated into cytoplasmic, periplasmic, and extracellular compartments. Added reactions include the biochemical transformations associated with glycan biosynthesis and flipping into the periplasm, as well as the conjugation of glycan to an acceptor protein. In addition, we incorporated the transcriptional regulatory network of Covert *et al.*, consisting of 101 transcription factors, regulating the state of the metabolic enzyme genes (see below) [12]. This network imparts constraints on metabolic fluxes in a Boolean fashion based on the boundary conditions that represent the nutrient environment.

Identification of growth-coupled gene knockout strains We used the constraint-based modeling technique known as flux balance analysis (FBA) to predict cellular phenotypes, i.e., the identification of all intracellular reaction rates, growth rate, byproduct formation, and rate of glycan synthesis [13]. FBA is largely a parameter-free method to identify optimal steady-state flux distributions in a stoichiometric model. Using our adapted genome-scale model of glycosylation in *E. coli*, we used a combination of heuristic optimization and FBA to identify genetic knockout strains that coupled optimal growth to target product flux, *C. jejuni* glycan synthesis (see Materials and Methods). Strains that route flux toward a desired product when growing optimally are called growth-coupled [14]. This phenotype is desirable for a variety of reasons. For example, growth-coupled strains create stoichiometric imbalances in metabolism such that flux is routed toward the desired product as a consequence of growth [15]. Also, because faster growth requires increasing product formation, further optimizing these strains through adaptive laboratory

evolution is made easy by selecting for growth through serial passage [14, 16]. Constraint-based methods have been developed to identify gene knockouts that couple production to growth. Most of these methods rely on an optimization approach known as OptKnock whereby a bi-level mixed integer optimization problem is solved to identify the optimal set of gene knockouts [15]. Though this method guarantees identification of the global optimum, it suffers from two limitations. First, search time for OptKnock-like algorithms scales exponentially with system size and number of gene knockouts, making them unable to handle large metabolic networks. Secondly, only linear engineering objectives (e.g., target production flux) can be searched over. In contrast, heuristic optimization is an effective approach for searching large networks for growth-coupled strains. Though identification of the global optimum is not guaranteed, desirable sub-optimal solutions can be found quickly [17, 18]. Also, heuristic optimization can search efficiently for gene knockouts rather than reaction knockouts. This is an important distinction because the mapping of genes to reactions is not one to one. Experimentally, many reactions may be difficult to knockout because they may be catalyzed by the products of many genes.

Here, we employed simulated annealing to search over the states of metabolic enzyme and transcription factor (TF) genes included in the model in order to identify our desired phenotype [19]. Figure 2 illustrates our approach. The state of each gene in the model is represented as a binary array. A zero indicates a genetic knockout or regulatory repression. Using Boolean rules, nutrient conditions control the state of the TF genes, which, in turn, control the state of the metabolic enzyme genes. Once defined, the genetic state of the model modifies the flux constraints placed on each reaction. Again, Boolean rules map the state of each gene to the availability of each reaction. For example, the reaction governed by pyruvate dehydrogenase, a multi-component enzyme, relies on the assembly of three enzymes AceE, AceF, and Lpd. This reaction is encoded in the following way:

$$\text{Pyruvate} + \text{CoA} + \text{NAD} \longrightarrow \text{Acetyl-CoA} + \text{CO}_2 + \text{NADH: (}\ \textit{aceE}\ \text{) AND (}\ \textit{aceF}\ \text{) AND (}\ \textit{lpd}\ \text{)}$$

Thus, if any of the genes *aceE*, *aceF*, or *lpd* is knocked out or transcriptionally repressed,

the flux through this reaction is bound to zero. Gene–protein-reaction (GPR) associations are available in the SBML representation of the iAF1260 network [11]. The simulated annealing algorithm performs a random search of genetic knockouts, iteratively applying flux constraints based on the genetic state then performing an FBA simulation. Strains improving a fitness objective are selected for based on a Boltzmann factor criterion. In order to identify glycan-producing strains, we optimized a non-linear fitness objective known as shadow price which measures the degree of coupling between glycan flux and growth rate. The shadow price is equal to the change in growth rate for a specified increase in target flux. The shadow price for glycan synthesis is given by:

$$u_{glycan} = \frac{\Delta v_{growth}}{\Delta v_{glycan}} \tag{1}$$

where Δv_{growth} is the change in growth rate for a forced change in glycan flux Δv_{glycan}. v_{glycan} is the flux representing the fully assembled *C. jejuni* glycan transported into the periplasm. u_{glycan} was calculated for a particular knockout strain by first using FBA to calculate the optimal growth with the glycan flux constrained to zero. A second FBA simulation was performed with a forced incremental change in the glycan flux in order to obtain the difference in growth rate. The algorithm maximized this fitness objective until reaching a positive shadow price, indicating a growth-coupled phenotype.

We performed optimization simulations using boundary conditions representing minimal medium conditions (see [11]) with a single common carbon substrate. These substrates included 6-, 5-, and 3-carbon sources glucose, xylose, and glycerol, respectively. We performed simulations only for well-defined, minimal media, because this allowed for precise control over nutrient conditions experimentally. We feel that such well-defined conditions can be more accurately simulated, particularly for the transcriptional regulatory network. For each carbon substrate, we performed ten optimization simulations to identify growth coupled strains. In order to identify growth-coupled strains with four or few knockouts, those most likely to be experimentally viable, we had to restrict the formation

of extracellular byproducts to only acetate. Table 2 lists these strains organized by FBA simulated glycan yield. We were able to identify growth-coupled strains with four or fewer knockouts only for growth on glucose. Supplementary Figure 1 displays a representative example of a single optimization run. Increasing shadow price along with corresponding growth rate are shown over the course of the search. At selected points, production envelopes are plotted for the corresponding strain. The production envelope is calculated by iteratively increasing flux to glycan by adjusting its minimum flux bound and maximizing growth rate using FBA. The resulting plot shows the optimal growth rate at a given level of product flux. For the identified strain of type EcGM1 (*E. coli glycosylating mutant*), the strain with the highest simulated glycan yield, we see that its optimal growth rate occurs at a non-zero glycan flux. All identified strains contain a knockout of succinate dehydrogenase (*sdh*) and cut off pentose phosphate pathway (PPP) flux at either glucose 6-phosphate-1-dehydrogenase (*zwf*), 6-phosphogluconolactonase (*pgl*), or 6-phosphogluconate dehydrogenase (*gnd*).

Flux analysis reveals mechanisms of glycan production in growth-coupled strains

Figure 3 shows key differences in flux values between the wild-type strain and a four gene knockout glycan-producing strain of type EcGM1, Δ*sdh* Δ*gnd* Δ*pta* Δ*eutD*, as calculated by FBA. Normalizing all fluxes to glucose uptake rate, we see that EcGM1 displays greater flux through glycolysis by cutting off the PPP via knockout of NADPH-producing *gnd* (Fig. 3A). EcGM1 also showed relative decreased synthesis of every amino acid except for glutamine, indicating a source of stoichiometric imbalance that may be relieved by synthesis of the glycan precursor UDP-GlcNAc. The PEP-pyruvate node acts as a switch point in central carbon metabolism (Fig. 3B). Here, PEP and pyruvate, the products of glycolysis, enter the TCA cycle through decarboxylation of pyruvate to acetyl-CoA (ACCoA) and carboxylation of PEP to form oxaloacetate (OAA) [20]. The latter replenishes TCA cycle intermediates that exited TCA for anabolic processes. EcGM1, with a diminished anabolic capacity for cell growth, displayed lower flux through PEP carboxylase (*ppc*). As the result

of high glycolytic flux, EcGM1 increases flux through pyruvate dehydrogenase (*aceEF*), sending carbon into the oxidative branch of the TCA cycle. It is known that high glucose uptake rates result in excess acetyl-CoA, surpassing the capacity of the TCA cycle. Due to this, wild-type *E. coli* grown on glucose commonly displays acetate fermentation, even under aerobic conditions [21]. We observe much greater acetate secretion in EcGM1 simulations, but through a route differing from wild-type cells. The knockouts Δ*pta* and Δ*eutD* prevent ATP-generating acetate secretion. Instead flux is routed through the redox neutral reactions initiated by acetaldehyde dehydrogenase (*mhpF*). Some excess acetyl-CoA is utilized in the pathway generating UDP-GlcNAc. Acetyl-CoA also plays a key role in lipid synthesis. Maintaining high flux through glycolysis seems key to maintaining the EcGM1 phenotype. EcGM1 also displays a shifts in cofactor production (Fig. 3C). Higher flux through glycolysis naturally lead to NADH overproduction. Also, the primary source of NADPH shifted from PPP genes *zwf* and *gnd* to the membrane transhydrogenase *pnt*, capable of direct transfer of electrons from NADH to NADP. Sauer *et al.* identified *pnt* as a major source of NADPH in *E. coli* (35-45% of total) [22]. Thus, *pnt* is capable of carrying significant flux *in vivo*. Taken together, these results indicate that the model has identified strains that promote synthesis of glycan precursors, primarily UDP-GlcNAc, by creating a combination of metabolite and redox imbalance.

Experimental validation of glycan-producing knockout strains Attempting to validate constraint-based model predictions, we measured both glycan production and glycosylation efficiency in mutant strains. Gene knockout strains were constructed using P1*vir* phage transduction and the Keio collection of single gene knockouts *E. coli* BW25113 [23]. Knockout strains were transformed with a plasmid constitutively expressing the *C. jejuni pgl* locus. In order to quantify glycan production, we take advantage of the crosstalk between the glycosylation pathway and native lipopolysaccharide (LPS) synthesis in *E. coli*. After the glycan is flipped into the periplasm, it may be transferred to the lipid carrier, lipid A, and shuttled to the outer membrane by LPS pathway enzymes. There, it is

displayed on the cell surface [1]. We used the amount of cell surface displayed glycan as a measure of glycan production. We labeled *C. jejuni* glycans for detection by flow cytometry with a fluorophore-conjugated carbohydrate-binding lectin protein called soybean agglutinin (SBA), specific to terminal galactose and GalNAc residues. Mutant strains were constructed containing single, double, and triple gene knockouts that appear in growth-couple strains identified by the constraint-based model. Also, we performed an FBA simulation of each single gene knockout in the model, maximizing glycan flux, to determine genes that prevent glycan synthesis. *galU*, a key enzyme in the synthesis of glycan precursor UDP-glucose, was the only non-lethal knockout predicted to do so. We grew strains in glucose minimal media and sampled cells from exponential growth phase, in order to most closely satisfy the pseudo-steady-state assumption of model predictions. Figure 4 shows the results of the glycan fluorescence detection experiments. As expected, $\triangle$*galU* shows no glycan production. PPP knockouts $\triangle$*zwf*, $\triangle$*pgl*, and $\triangle$*gnd* all display greater fluorescence than wild-type cells, with $\triangle$*gnd* being the most significant. Of the double knockouts containing PPP genes, only $\triangle$*sdhC* $\triangle$*gnd* maintained its fluorescence.

In order to determine if the observed increase in glycan production also enhanced protein glycosylation in *E. coli*, we performed Western blot analysis to determine glycosylation efficiency of a modified antibody fragment scFvR4. This protein was modified by the addition of four *C. jejuni* glycosylation sites along with a DsbA signal peptide that localizes the protein to the periplasm [4]. For this experiment, we used the *E. coli* strain CLM24 lacking the LPS ligase *waaL* so that glycan is not transported to the cell surface by the LPS pathway, making more available for protein conjugation by the *C. jejuni* OST *pglB*. Cells expressing the *C. jejuni pgl* locus were grown in glucose minimal media to exponential phase (OD_{600} $\sim$0.5), then scFvR4 expression from a pTrc99A plasmid was induced by addition of isopropyl β-D-1-thiogalactopyranoside (IPTG). Equal numbers of cells were sampled at time points ranging 2-10 hours. The periplasmic protein fraction was isolated and proteins were probed with anti-6$\times$-His antibody conjugated to

horseradish peroxidase (HRP). Figure 5 shows a time series Western blot analysis of the CLM24 $\triangle$*gnd* mutant. We observe a range of scFvR4 glycosylation, with up to four occupied sites. Comparing blot intensities of total and glycosylated protein, we see that *gnd* produced up to 60% more total acceptor protein while also displaying up to a 2.6-fold increase in glycosylation efficiency. This observed increase in glycosylation efficiency may be a combination of overexpression of *pgl* locus enzymes due to higher overall protein production as well as greater availability of glycan precursor substrates. Previous ^{13}C flux measurements of *E. coli*'s metabolic response to *gnd* knockout have shown a similar phenotype as the one predicted by our model simulations. Jiao *et al.* showed that $\triangle$*gnd* had a slightly higher glucose uptake rate and much greater acetate byproduct formation [24]. While significantly increasing flux through glycolysis, $\triangle$*gnd* also rerouted flux through the Entner-Doudoroff (ED) pathway. The ED pathway is an alternative to glycolysis that nets 1 ATP, 1 NADP, and 1 NADPH per glucose molecule, whereas glycolysis yields 2 ATP and 2 NADH per glucose molecule. Also, the *gnd* mutant displayed increased flux through the malic enzyme (maeB) of the anaplerotic pathway, upregulating phosphoenolpyruvate carboxylase (*ppc*) and downregulating phosphoenolpyruvate carboxykinase (*pck*). These results showed that *E. coli* compensates for an inability to produce NADPH through the oxidative PPP by upregulating flux through other NADPH-producing pathways. In this way, $\triangle$*gnd* maintained a growth rate and amino acid production comparable to wild-type cells. Compared to our model results and experimental validation, maintaining high glucose uptake and flux through glycolysis is key to simultaneous glycan and glycoprotein production. Taken together, experiments indicate that model predictions may have identified a novel approach to increase glycosylation efficiency in *E. coli*.

Discussion

In this study we adapted a genome-scale model of *E. coli* metabolism for the simulation of heterologous synthesis of glycans. We applied a heuristic optimization search with FBA that identified gene knockout strains that coupled *C. jejuni* glycan synthesis to growth. Simulations identified growth-coupled strains for growth on a single carbon substrate. Flux analysis of these strains revealed two modes of flux redistribution that promoted glycan synthesis. For growth on glucose, simulations showed that maintaining high glycolytic flux and producing excess glutamine for the amination of glycan precursor sugars led to a growth-coupled phenotype. Simulations identified the pentose phosphate pathway as a primary target. We validated model predictions by measuring cell surface-displayed glycans in *E. coli* mutants. In both growth conditions, a $Delta$*gnd* mutant outperformed the wild-type case in glycan synthesis. For growth on glucose, we showed an increase in glycosylation efficiency for a prototypical acceptor protein in the case of the $Delta$*gnd* mutant. Overall, our model-guided strategy shows promise toward rationally designing a better microbial glycosylation platform.

Many aspects of glycoprotein production in *E. coli* are amenable to investigation and engineering by metabolic modeling. This study focused on increasing the availability of glycan precursor metabolites through model-guided metabolic network manipulations. Other approaches have focused on optimizing expression of glycosylation pathway enzymes and identification of metabolic reaction targets through proteomic and genome engineering [6, 9, 10]. Despite these efforts, improving glycosylation efficiency in *E. coli* has proven to be a difficult task. A more comprehensive mathematical description of the cell, one that couples metabolism with gene expression and metabolic demand, may be required to more precisely model glycosylation in *E. coli*. Our approach does not take into account the metabolic burden associated with heterologous expression of glycosylation pathway enzymes nor the expression of the acceptor glycoprotein. Also, FBA lacks a description of enzyme kinetics and metabolite concentrations. It has indeed been shown

that single knockout mutants of genes in central metabolism of *E. coli* do little to change the relative flux distribution in the organism [25]. This also underscores the importance of serial passage of FBA-identified strains in order to achieve optimal growth flux distributions, maximizing the predictive capabilities of FBA. *E. coli* robustly controls metabolic flux using a plethora of allosteric, transcriptional regulatory, and post-translational modification systems [26, 27]. Predicting phenotypic changes to genetic perturbations is a primary challenge in model-guided metabolic engineering [28]. Glycoprotein production in *E. coli* is a unique challenge in that it requires optimization of two opposing cellular processes. Recombinant protein production of a desired glycoprotein along with glycosylation pathway enzymes requires immense energy from catabolic processes. In contrast, glycan precursor synthesis requires conservation of available sugars and anabolic processes. Addition of regulatory systems and gene expression to a stoichiometric model may be an effective strategy for optimizing these opposing processes. Other strategies that may be helpful for optimization of this system include the enhancement of glycan precursor pathways, such as hexosamine synthesis, as well as the removal of competing pathways.

Materials and Methods

Flux balance analysis and heuristic optimization Reactions encoding *C. jejuni* glycan formation (Table 1) were added to the genome-scale metabolic model of *E. coli* iAF1260 [11]. FBA requires two primary assumptions. First, the cell is assumed to operate at a pseudo-steady-state, where the rate of production of every intracellular metabolite is equal to its consumption. Second, we assume that the cell has evolved to operate optimally to achieve a cellular objective. Though many objectives have been proposed, we use the most common, namely, growth rate (i.e., biomass formation) maximization [29]. The determination of a flux distribution satisfying these assumptions can be formulated as the following linear optimization problem:

$$\max_{\boldsymbol{v}} \left(v_{growth} = \boldsymbol{c}^T \boldsymbol{v}\right)$$
$$\text{Subject to} : \boldsymbol{S}\ \boldsymbol{v} = 0$$
$$\alpha_i \leq v_i \leq \beta_i$$

where $\boldsymbol{v}$ is the steady-state flux vector and α_i and β_i are the lower and upper limits for the individual flux values, respectively. v_{growth} is the growth rate where $\boldsymbol{c}$ is a vector containing the stoichiometric contribution of each metabolic species to biomass. The stoichiometric matrix $\boldsymbol{S}$ encodes all reaction connectivity considered in the model. Each row of $\boldsymbol{S}$ describes a metabolite, while each column describes a particular reaction. The (i, j) element of $\boldsymbol{S}$, denoted by σ_{ij}, describes how species i participates in reaction j. If $\sigma_{ij} > 0$, species i is produced by reaction j. Conversely, if $\sigma_{ij} < 0$, then species i is consumed by reaction j. Lastly, if $\sigma_{ij} = 0$, then species i is not involved in reaction j. An optimization problem like this is readily solved by linear programming, even for large systems. Boundary conditions were set to allow for the unrestricted formation of acetate. All genes found to be essential for growth on Luria-Bertani (LB) medium according to [23] were excluded from the search. Maximum substrate uptake rates were set at 10 mmol/gDW/hr.

The maximum oxygen uptake rate was set at 10 mmol/gDW/hr.

Use of the shadow prices objective was used in the FastPros algorithm developed by Ohno *et al.* [30]. Prior to optimization we removed genes associated with dead end reactions from consideration, since knocking those out would have no effect on the network. Also, we removed from consideration duplicate genes, i.e., those that produce identical effects when knocked out. Finally, we removed genes whose knockout resulted in zero growth. These pre-processing steps decreased the model's search space. The simulated annealing search optimization is similar to the OptGene algorithm [17]. The metabolic and transcriptional regulatory genes are represented by a binary array where 1 indicates the gene is expressed and 0 zero indicates it is knocked out. A random initial gene knockout array is generated. We allowed for a maximum of 20 knockouts during the search. New knockout arrays are generated through crossover and mutation operators (see [17]) that randomly introduce new knockouts. At each step, the fitness of the individual is computed using FBA. We use shadow price as our measure of fitness (Equation 1). When an individual with a higher fitness is encountered (greater shadow price), that individual is accepted. When an individual with a lower fitness is encountered, we accept it with a probability given by a Boltzmann factor:

$$P(accept) = e^{-\Delta u_{glycan}/T} \tag{2}$$

where Δu_{glycan} is the change in shadow price between the current solution and previous one and the temperature T decreases as the search goes on. The temperature decreases exponentially such that $T_{n+1} = \alpha T_n$ where α is a cooling rate. Here, we use a common cooling schedule consisting of initial and final temperatures as well as the cooling rate from [18] such that:

$$T_o = -\frac{\Delta u_{glycan,o}}{\log 0.5} \tag{3}$$

$$T_f = -\frac{\Delta u_{glycan,f}}{\log 0.5} \tag{4}$$

$$\alpha = \exp\left(\frac{\log T_f - \log T_o}{N_{max}/N_\alpha}\right) \tag{5}$$

where N_{max} is the maximum number of objective function evaluations to perform. N_α is the number of objective function evaluations to perform at each distinct temperature value. Here, we use $N_{max} = 10,000$ and $N_\alpha = 1$. $\Delta u_{glycan,o}$ is the difference in shadow price corresponding to an acceptance probability of worse solutions of 50% at the beginning of the search. $\Delta u_{glycan,f}$ is the shadow price difference giving a 50% probability of accepting a worse solution by the end of the search. These values were approximated using the typical shadow price values of random knockout arrays. We used $\Delta u_{glycan,o} = 0.005$ and $\Delta u_{glycan,f} = 0.0005$. Though, we sought to maximize glycan flux, we also wanted to identify experimentally viable strains. Thus, during an optimization search, we set a lower bound on the biomass reaction flux equal to 10% of the wild-type simulated growth rate. Strains that could not meet this constraint were ignored. The search is terminated once a positive shadow price is found. After optimization we processed growth-coupled knockout strains be iteratively knocking in each knockout gene to find knockouts that did not affect the phenotype. In this way we identified the smallest number of gene knockouts that produced glycan at optimal growth. Each optimization run required on the order of 6 hours on a single CPU Apple workstation (Apple, Cupertino, CA, USA; OS X v10.10).

Bacterial strains and media Plasmid pCP20 was used to excise KmR cassette [31]. For surface-labeled glycan fluorescence measurements, we used the *E. coli* strain BW25113 as our wild-type case [23]. BW25113 was used as the parent strain to construct gene knockout strains. *E. coli* strain CLM24 was used for glycoprotein expression experiments [32]. CLM24 lacks the LPS ligase *waaL*, allowing for *pgl* locus oligosaccharyltransferase PglB to attached glycans on the periplasmic side of the inner membrane to the acceptor protein. Minimal media consisted of 33.9 g/L Na_2HPO_4, 15.0 g/L KH_2PO_4, 5.0 g/L NH_4Cl, and 2.5 g/L NaCl. Media was supplemented with 0.4% glucose. Growth medium was supplemented by appropriate antibiotic at: 100 μg/mL ampicillin (Amp), 25 μg/mL chloramphenicol, and 50 μg/mL kanamycin (Kan). Growth was monitored by measuring optical

density at 600 nm (OD_{600}).

Flow cytometry BW25113-based knockout strains were transformed with plasmid pACYCpgl, constitutively expression the *C. jejuni pgl* locus. Cultures were inoculated from frozen stock in LB and grew for 3-6 hours. Cells were subcultured 1:100 in minimal media overnight and then transferred to fresh minimal media to an OD_{600} of 0.1. 300 μL cells were harvested during exponential growth phase ($OD_{600} \approx 0.6$). Cells were washed with PBS then incubated in the dark for 15 minutes at $37°$. Cells were resuspended in 5 μg/mL SBA-Alexa Fluor 488 (Invitrogen) and PBS. Cells were resuspended in 500 μL PBS and analyzed using a FACSCalibur (Becton Dickinson). Geometric mean fluorescence was determined from $100,000$ events.

Protein analysis For glycoprotein expression, CLM24 knockout strains were transformed with pACYCpgl and pTrc99a-ssDsbA-scFvR4-4xGT, expressing an IgG Fc domain containing four glycosylation tags and a DsbA signal peptide to localize it to the periplasm [4]. Cells from frozen stock were grown in LB for 3-6 hours then subcultured 1:100 in minimal media overnight. Cells were diluted to 0.1 OD_{600} in fresh minimal media and grown to OD_{600} 0.4. Cultures were then induced with IPTG and grown at 37$\deg$C. To isolate periplasmic localized glycoprotein, the periplasmic protein fraction was isolated. Equal numbers of cells were harvested and treated with 100 mM iodoacetamide as above, pelleted by centrifugation, and fractionated by the cold osmotic shock procedure as in [33]. Proteins were separated with SDS-polyacrylamide gels (Bio-Rad). Proteins were transferred onto polyvinylidene fluoride (PVDF) membranes probed with anti-6x-his antibodies conjugated to horseradish peroxidase (HRP).

Acknowledgements

The authors thank the anonymous reviewers for their helpful suggestions. We also acknowledge the gracious financial support to J.V. by the National Science Foundation CAREER (CBET-0846876) for the support of J.W.

Author Contributions

Author contributions go here ...

Conflict of Interest

The authors declare no conflicts of interest.

References

1. Merritt, J. H., Ollis, A. A., Fisher, A. C., and DeLisa, M. P. Glycans-by-design: engineering bacteria for the biosynthesis of complex glycans and glycoconjugates. *Biotechnol Bioeng* **110**(6), 1550–64, Jun (2013).

2. Solá, R. J. and Griebenow, K. Glycosylation of therapeutic proteins: an effective strategy to optimize efficacy. *BioDrugs* **24**(1), 9–21, Feb (2010).

3. Szymanski, C. M., Yao, R., Ewing, C. P., Trust, T. J., and Guerry, P. Evidence for a system of general protein glycosylation in campylobacter jejuni. *Mol Microbiol* **32**(5), 1022–30, Jun (1999).

4. Fisher, A. C., Haitjema, C. H., Guarino, C., Çelik, E., Endicott, C. E., Reading, C. A., Merritt, J. H., Ptak, A. C., Zhang, S., and DeLisa, M. P. Production of secretory and extracellular n-linked glycoproteins in escherichia coli. *Appl Environ Microbiol* **77**(3), 871–81, Feb (2011).

5. Wacker, M., Linton, D., Hitchen, P. G., Nita-Lazar, M., Haslam, S. M., North, S. J., Panico, M., Morris, H. R., Dell, A., Wren, B. W., and Aebi, M. N-linked glycosylation in campylobacter jejuni and its functional transfer into e. coli. *Science* **298**(5599), 1790–3, Nov (2002).

6. Pandhal, J., Ow, S. Y., Noirel, J., and Wright, P. C. Improving n-glycosylation efficiency in escherichia coli using shotgun proteomics, metabolic network analysis, and selective reaction monitoring. *Biotechnol Bioeng* **108**(4), 902–12, Apr (2011).

7. Valderrama-Rincon, J. D., Fisher, A. C., Merritt, J. H., Fan, Y.-Y., Reading, C. A., Chhiba, K., Heiss, C., Azadi, P., Aebi, M., and DeLisa, M. P. An engineered eukaryotic protein glycosylation pathway in escherichia coli. *Nat Chem Biol* **8**(5), 434–6, May (2012).

8. Jaffé, S. R., Strutton, B., Levarski, Z., Pandhal, J., and Wright, P. C. Escherichia coli as a glycoprotein production host: recent developments and challenges. *Curr Opin Biotechnol* **30C**, 205–210, Aug (2014).

9. Pandhal, J., Woodruff, L. B. A., Jaffe, S., Desai, P., Ow, S. Y., Noirel, J., Gill, R. T., and Wright, P. C. Inverse metabolic engineering to improve escherichia coli as an n-glycosylation host. *Biotechnol Bioeng* **110**(9), 2482–93, Sep (2013).

10. Pandhal, J., Desai, P., Walpole, C., Doroudi, L., Malyshev, D., and Wright, P. C. Systematic metabolic engineering for improvement of glycosylation efficiency in escherichia coli. *Biochem Biophys Res Commun* **419**(3), 472–6, Mar (2012).

11. Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. Ø. A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Mol Syst Biol* **3**, 121 (2007).

12. Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., and Palsson, B. O. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**(6987), 92–6, May (2004).

13. Varma, A. and Palsson, B. O. Metabolic flux balancing: Basic concepts, scientific and practical use. *Nat Biotechnol* **12**, 994–998 (1994).

14. Feist, A. M., Zielinski, D. C., Orth, J. D., Schellenberger, J., Herrgard, M. J., and Palsson, B. Ø. Model-driven evaluation of the production potential for growth-coupled products of escherichia coli. *Metab Eng* **12**(3), 173–86, May (2010).

15. Burgard, A. P., Pharkya, P., and Maranas, C. D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* **84**(6), 647–57, Dec (2003).

16. Ibarra, R. U., Edwards, J. S., and Palsson, B. O. Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420**(6912), 186–9, Nov (2002).

17. Patil, K. R., Rocha, I., Förster, J., and Nielsen, J. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* **6**, 308 (2005).

18. Rocha, M., Maia, P., Mendes, R., Pinto, J. P., Ferreira, E. C., Nielsen, J., Patil, K. R., and Rocha, I. Natural computation meta-heuristics for the in silico optimization of microbial strains. *BMC Bioinformatics* **9**, 499 (2008).

19. Kirkpatrick, S., Gelatt, Jr, C. D., and Vecchi, M. P. Optimization by simulated annealing. *Science* **220**(4598), 671–80, May (1983).

20. Sauer, U. and Eikmanns, B. J. The pep-pyruvate-oxaloacetate node as the switch point for carbon flux distribution in bacteria. *FEMS Microbiol Rev* **29**(4), 765–94, Sep (2005).

21. Gosset, G. Improvement of escherichia coli production strains by modification of the phosphoenolpyruvate:sugar phosphotransferase system. *Microb Cell Fact* **4**(1), 14, May (2005).

22. Sauer, U., Canonaco, F., Heri, S., Perrenoud, A., and Fischer, E. The soluble and membrane-bound transhydrogenases udha and pntab have divergent functions in nadph metabolism of escherichia coli. *J Biol Chem* **279**(8), 6613–9, Feb (2004).

23. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., and Mori, H. Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection. *Mol Syst Biol* **2**, 2006.0008 (2006).

24. Jiao, Z., Baba, T., Mori, H., and Shimizu, K. Analysis of metabolic and physiological responses to gnd knockout in escherichia coli by using c-13 tracer experiment and enzyme activity measurement. *FEMS Microbiol Lett* **220**(2), 295–301, Mar (2003).

461 25. Sauer, U., Lasko, D. R., Fiaux, J., Hochuli, M., Glaser, R., Szyperski, T., Wüthrich, K.,
462 and Bailey, J. E. Metabolic flux ratio analysis of genetic and environmental modula-
463 tions of escherichia coli central carbon metabolism. *J Bacteriol* **181**(21), 6679–88,
464 Nov (1999).

465 26. Kremling, A., Bettenbrock, K., and Gilles, E. D. A feed-forward loop guarantees robust
466 behavior in escherichia coli carbohydrate uptake. *Bioinformatics* **24**(5), 704–10, Mar
467 (2008).

468 27. Link, H., Kochanowski, K., and Sauer, U. Systematic identification of allosteric protein-
469 metabolite interactions that control enzyme activity in vivo. *Nat Biotechnol* **31**(4),
470 357–61, Apr (2013).

471 28. Link, H., Christodoulou, D., and Sauer, U. Advancing metabolic models with kinetic
472 information. *Curr Opin Biotechnol* **29C**, 8–14, Oct (2014).

473 29. Schuetz, R., Kuepfer, L., and Sauer, U. Systematic evaluation of objective functions
474 for predicting intracellular fluxes in escherichia coli. *Mol Syst Biol* **3**, 119 (2007).

475 30. Ohno, S., Shimizu, H., and Furusawa, C. Fastpros: screening of reaction knockout
476 strategies for metabolic engineering. *Bioinformatics* **30**(7), 981–7, Apr (2014).

477 31. Cherepanov, P. P. and Wackernagel, W. Gene disruption in escherichia coli: Tcr and
478 kmr cassettes with the option of flp-catalyzed excision of the antibiotic-resistance
479 determinant. *Gene* **158**(1), 9–14, May (1995).

480 32. Feldman, M. F., Wacker, M., Hernandez, M., Hitchen, P. G., Marolda, C. L., Kowarik,
481 M., Morris, H. R., Dell, A., Valvano, M. A., and Aebi, M. Engineering n-linked protein
482 glycosylation with diverse o antigen lipopolysaccharide structures in escherichia coli.
483 *Proc Natl Acad Sci U S A* **102**(8), 3016–21, Feb (2005).

484 33. DeLisa, M. P., Tullman, D., and Georgiou, G. Folding quality control in the export of

485 proteins by the bacterial twin-arginine translocation pathway. *Proc Natl Acad Sci U S*
486 *A* **100**(10), 6115–20, May (2003).

487 34. Bernatchez, S., Szymanski, C. M., Ishiyama, N., Li, J., Jarrell, H. C., Lau, P. C.,
488 Berghuis, A. M., Young, N. M., and Wakarchuk, W. W. A single bifunctional udp-
489 glcnac/glc 4-epimerase supports the synthesis of three cell surface glycoconjugates
490 in campylobacter jejuni. *J Biol Chem* **280**(6), 4792–802, Feb (2005).

491 35. Schoenhofen, I. C., McNally, D. J., Vinogradov, E., Whitfield, D., Young, N. M., Dick,
492 S., Wakarchuk, W. W., Brisson, J.-R., and Logan, S. M. Functional characterization of
493 dehydratase/aminotransferase pairs from helicobacter and campylobacter: enzymes
494 distinguishing the pseudaminic acid and bacillosamine biosynthetic pathways. *J Biol*
495 *Chem* **281**(2), 723–32, Jan (2006).

496 36. Olivier, N. B., Chen, M. M., Behr, J. R., and Imperiali, B. In vitro biosynthesis of udp-
497 n,n'-diacetylbacillosamine by enzymes of the campylobacter jejuni general protein
498 glycosylation system. *Biochemistry* **45**(45), 13659–69, Nov (2006).

499 37. Glover, K. J., Weerapana, E., Chen, M. M., and Imperiali, B. Direct biochemical
500 evidence for the utilization of udp-bacillosamine by pglc, an essential glycosyl-1-
501 phosphate transferase in the campylobacter jejuni n-linked glycosylation pathway.
502 *Biochemistry* **45**(16), 5343–50, Apr (2006).

503 38. Kelly, J., Jarrell, H., Millar, L., Tessier, L., Fiori, L. M., Lau, P. C., Allan, B., and Szy-
504 manski, C. M. Biosynthesis of the n-linked glycan in campylobacter jejuni and addition
505 onto protein through block transfer. *J Bacteriol* **188**(7), 2427–34, Apr (2006).

506 39. Linton, D., Dorrell, N., Hitchen, P. G., Amber, S., Karlyshev, A. V., Morris, H. R., Dell,
507 A., Valvano, M. A., Aebi, M., and Wren, B. W. Functional analysis of the campylobac-
508 ter jejuni n-linked protein glycosylation pathway. *Mol Microbiol* **55**(6), 1695–703, Mar
509 (2005).

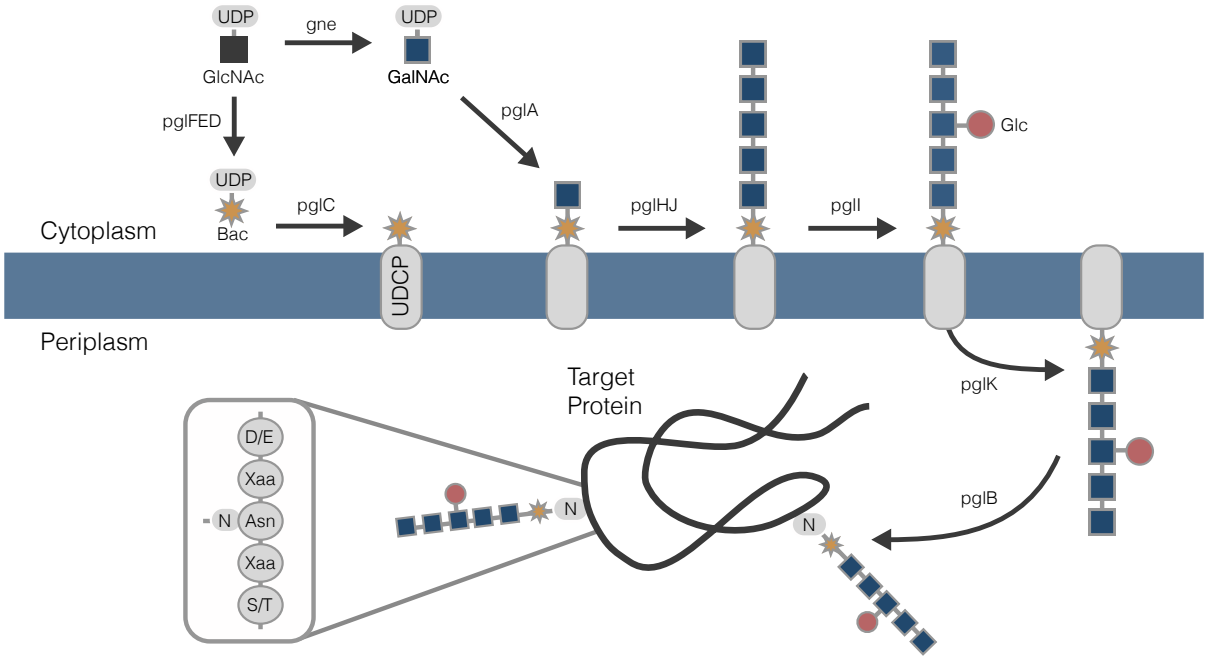

Fig. 1: Glycosylation pathway in *C. jejuni* and *E. coli*. Glycan assembly, facilitated by *pgl* locus enzymes, takes place on a lipid carrier, undecaprenyl pyrophosphate (UDCP), from cytoplasmic pools of nucleotide-activated sugars N-acetylglucosamine (GlcNAc), N-acetylgalactosamine (GalNAc), and glucose (Glc). The glycan is then flipped onto the periplasmic side of the inner membrane, where it is transferred to an asparagine residue on a glycoprotein acceptor motif.

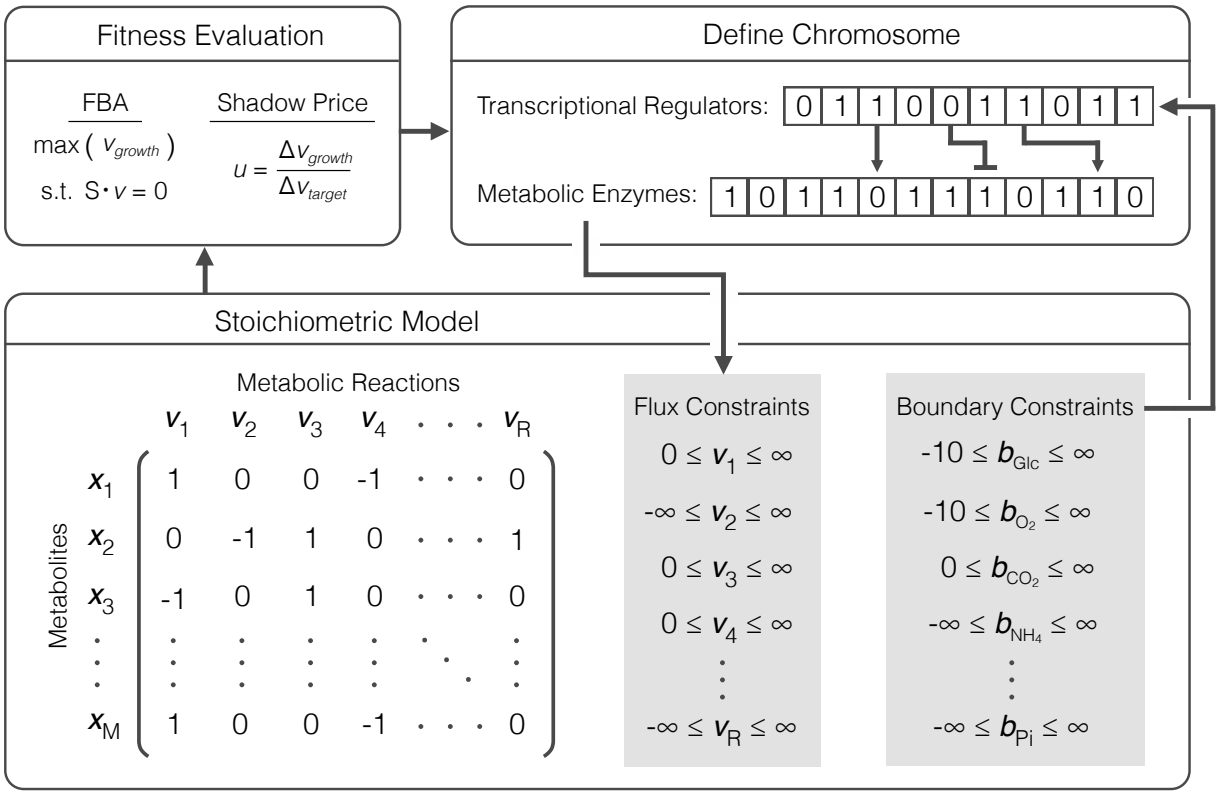

Fig. 2: Heuristic optimization approach used to identify strains coupling growth to glycan production. The chromosome is defined as two separate binary arrays, one defining the state of metabolic enzyme expression and another defining the state of transcriptional regulator activation. Gene repression and knockouts are designated by zeros. Nutrient conditions define the boundary constraints within the stoichiometric model which in turn affect the state of the metabolic enzyme chromosome. Gene repression and knockouts determine the constraints placed on fluxes in the stoichiometric model. Nutrients are mapped to the state of transcriptional regulators and genes are mapped to the state of flux constraints using Boolean rules as defined in [11, 12]. FBA is used to maximize growth rate under the constraints imposed by the mutant strain and transcriptional regulation and the fitness objective is calculated. Here, we use shadow price. The strain is accepted or rejected based on the change in fitness and a Boltzmann criterion. New mutant strains are randomly generated from accepted ones. The search continues until a positive shadow price is achieved.

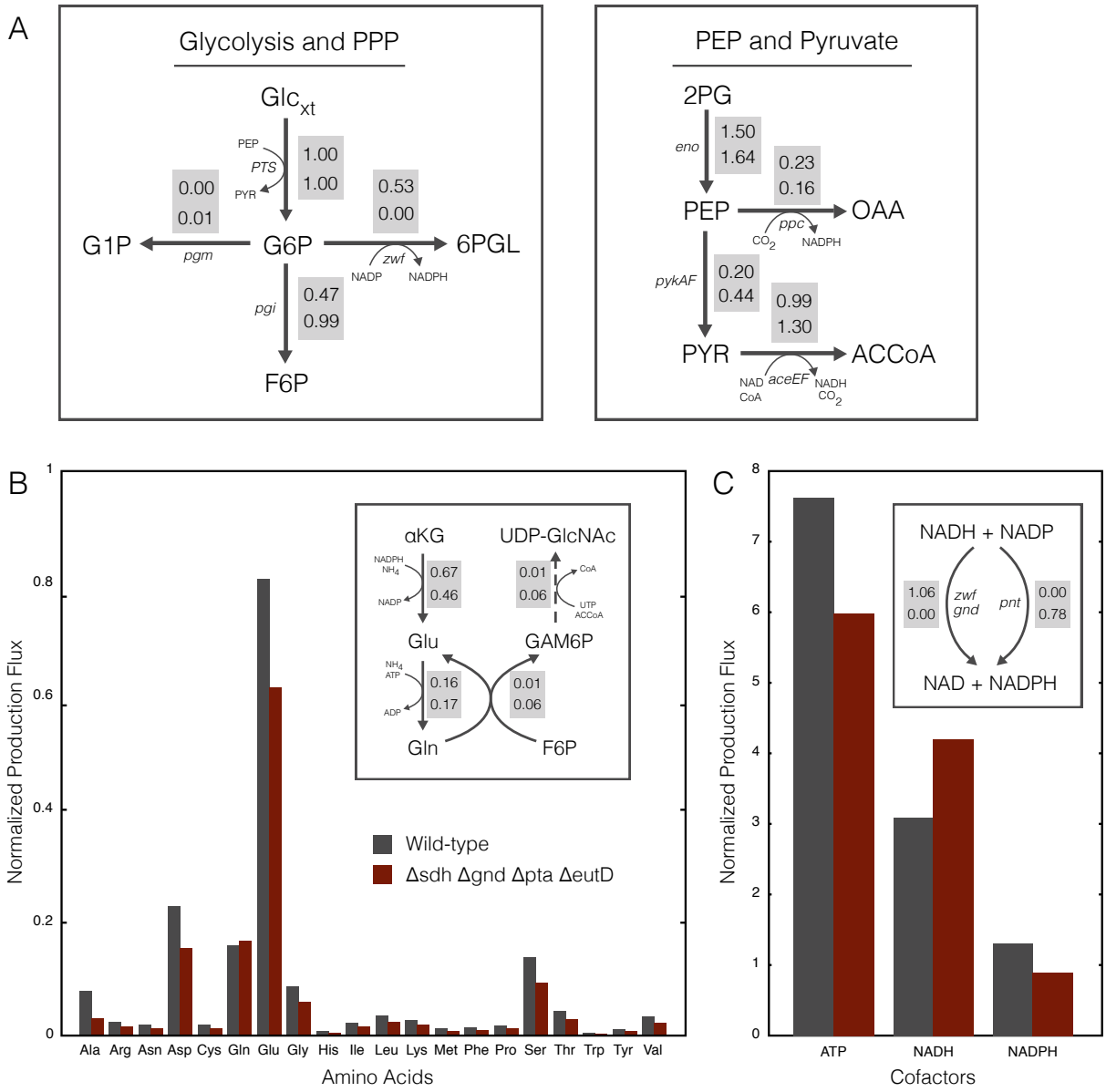

Fig. 3: Comparison of fluxes between the wild-type case and glycan-producing strain of type EcGM1 as calculated by FBA. **(A)** Fluxes through key nodes of metabolism. Top fluxes correspond to the wild-type case, bottom fluxes are for strain EcGM1. Fluxes are normalized by the glucose uptake rate. **(B)** Total flux into each amino acid, normalized to glucose uptake rate. Inset shows fluxes associated with glutamate and glutamine synthesis along with the pathway to glycan precursor UDP-GlcNAc. The dotted arrow represents a lumped pathway of multiple enzymes leading to the glycan precursor. **(C)** Total flux into selected cofactors, normalized to glucose uptake rate. Inset shows the primary modes of NADPH production in each strain. Abbreviations: Pentose phosphate pathway, PPP; Extracellular glucose, Glc$_{xt}$; Glucose-6-phosphate, G6P; Fructose 6-phosphate, F6P; 6-phospho D-glucono-1,5-lactone, 6PGL; Glucose 1-phosphate, G1P; Glycerate 2-phosphate, 2PG; Phosphoenolpyruvate, PEP; Pyruvate, PYR; Oxaloacetate, OAA; Acetyl-CoA, ACCoA; 2-Oxoglutarate, αKG; Glucosamine 6-phosphate, GAMP6P; UDP-N-acetyl-D-glucosamine, UDP-GlcNAc.

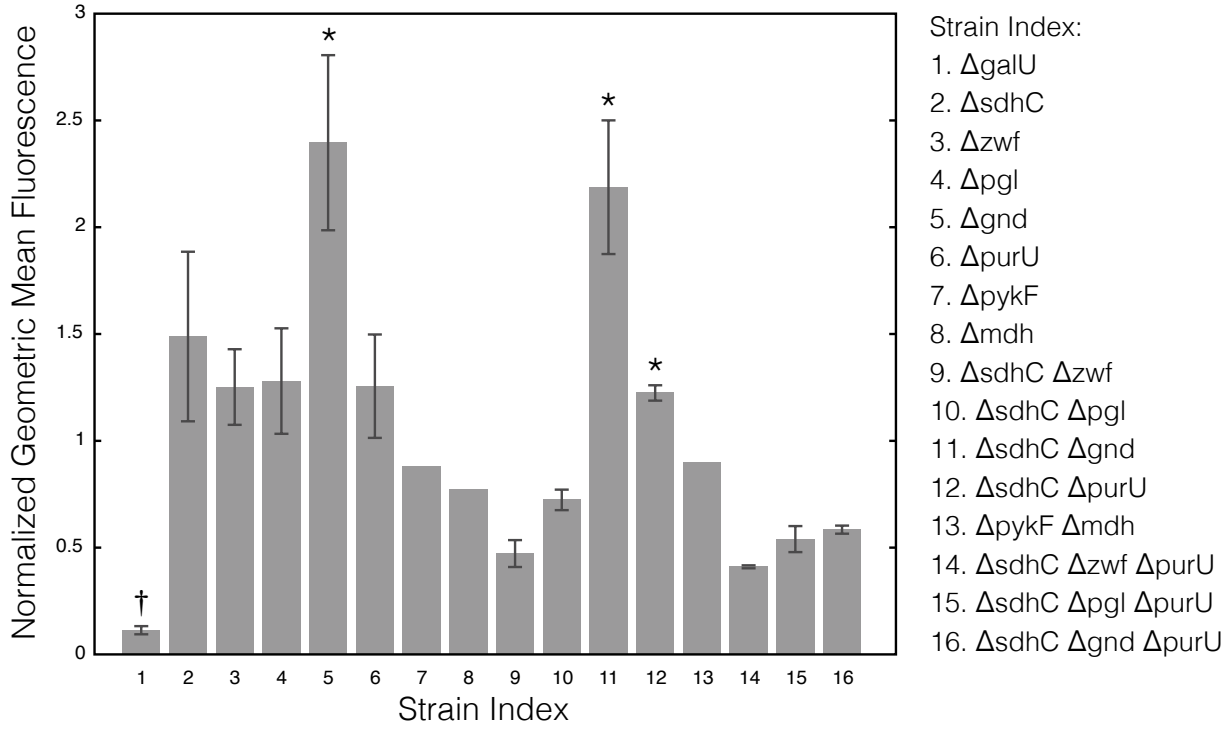

Fig. 4: Geometric mean fluorescence, normalized to the wild-type value, from gene knockout strains appearing in growth-coupled strains identified by the constraint-based model. † indicates a strain predicted to eliminate glycan flux. Stars indicate statistically significant increases in fluorescences according to a t-test ($p < 0.05$). Error bars indicate the average of at least three replicates.

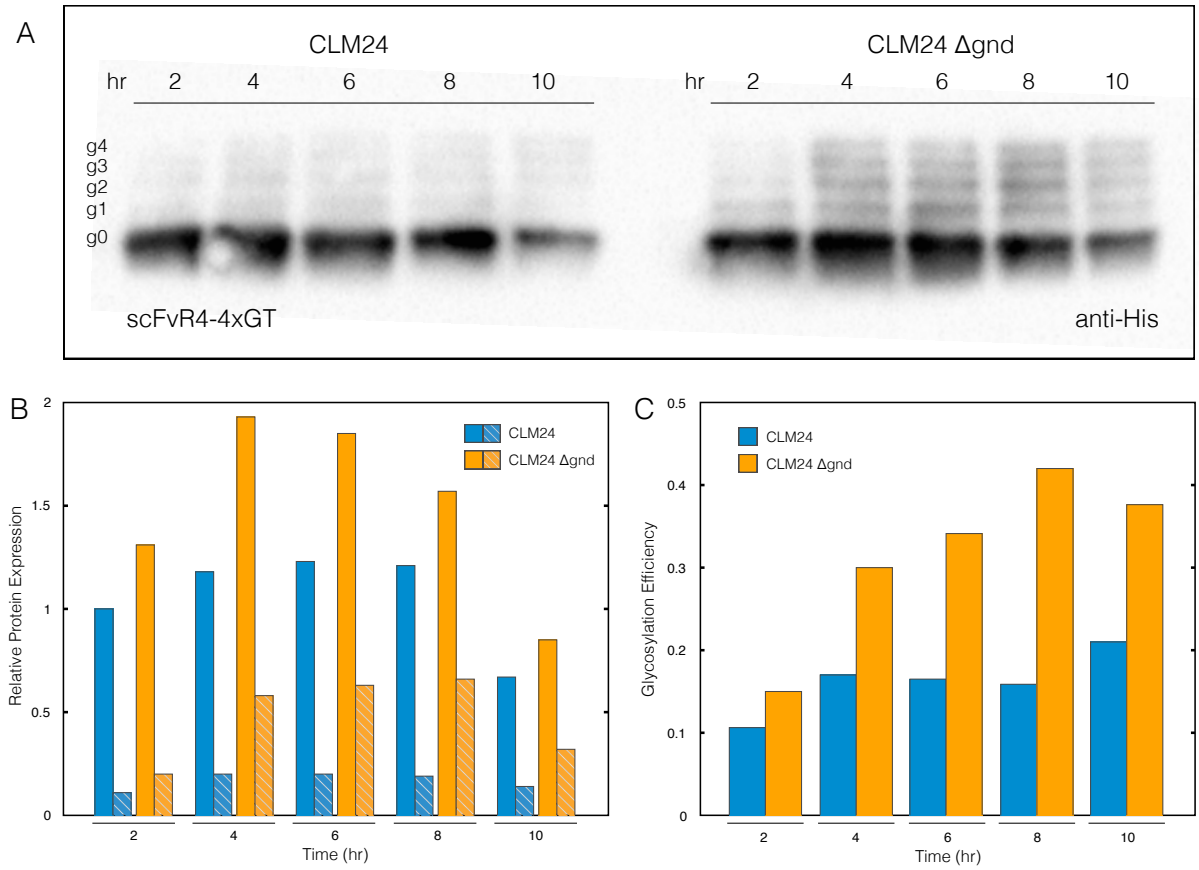

Fig. 5: Western blot analysis of glycosylation efficiency in *gnd* mutant. scFvR4 protein expression was induced during exponential growth phase. Time points indicate time after induction. *(A)* Relative protein expression over time. Striped bars indicate the portion of glycosylated protein. *(B)* Glycosylation efficiencies for each strain over time. Blot intensities were determined using image analysis software called ImageJ.

Table 1: Reactions added to the *E. coli* model iAF1260 [11] for biosynthesis of *C. jejuni* glycan. Species localized to the periplasm are denoted by (p), all others are cytoplasmic. Abbreviations: UDP-N-Acetyl-D-Glucosamine, UDP-GlcNAc; UDP-N-Acetyl-D-Galactosamine, UDP-GalNAc; UDP-2-acetamido-2,6-dideoxy-α-D-xylo-4-hexulose, KetoBac; L-Glutamate, Glu; UDP-N-Acetylbacillosamine, AminoBac; α-ketoglutarate, αKG; Acetyl-CoA, ACCoA; UDP-N,N'-diacetylbacillosamine, uBac; Coenzyme A, CoA; Undecaprenyl phosphate, Udcpp; *C. jejuni* glycan intermediates, UdcCjGlycan1, UdcCjGlycan6; Uridine monophosphate, UMP; Uridine diphosphate, UDP; UDP-Glucose, UDP-Glc; Lipid-linked *C. jejuni* glycan, UdcCjGlycan; Acceptor protein, AcceptorProt; GlycoProt, Glycoprotein; Undecaprenyl diphosphate, Udcpdp.

Gene	Enzyme Name	Reaction	Reference
gne	UDP-GlcNAc epimerase	UDP-GlcNAc $\rightarrow$ UDP-GalNAc	[34]
pglF	UDP-GlcNAc dehydratase	UDP-GlcNAc $\rightarrow$ KetoBac + H_2O	[35]
pglE	Aminotransferase	KetoBac + Glu $\leftrightarrow$ AminoBac + αKG	[35]
pglD	Acetyltransferase	AminoBac + ACCoA $\rightarrow$ uBac + CoA + H^+	[36]
pglC	Bacillosamine transferase	Udcpp + uBac $\rightarrow$ UdcCjGlycan1 + UMP	[37]
pglAHJ	GalNAc transferases	UdcCjGlycan1 + 5*UDP-GalNAc $\rightarrow$ UdcCjGlycan6 + 5*UDP + 5*H^+	[**?**]
pglI	Glucosyl transferase	UdcCjGlycan6 + UDP-Glc $\rightarrow$ UdcCjGlycan + UDP + H^+	[38]
pglK	ATP-driven flippase	UdcCjGlycan + ATP + H_2O $\rightarrow$ UdcCjGlycan(p) + ADP + H^+ + Pi	[38]
pglB	Oligosyltransferase	UdcCjGlycan(p) + AcceptorProt(p) $\rightarrow$ GlycoProt + Udcpdp(p)	[39]

Table 2: Growth-coupled strains producing *C. jejuni* glycan identified by FBA and heuristic optimization using single carbon substrate. Knockouts listing multiple genes indicate that knockout of any one of those genes produces the same phenotype in the model. Abbreviations: D-Glucose, Glc; *E. coli* glycosylating mutant, EcGM

Strain Name	Substrate	Genotype	Growth Rate (/hr)	Glycan Flux (mmol/gDW/hr)	Yield (mmol/gDW)
EcGM1	Glc	Δsdh Δ(zwf/pgl/gnd) Δpta ΔeutD	0.53	0.098	0.185
EcGM2	Glc	Δsdh Δ(zwf/pgl/gnd) ΔpykAF Δmdh	0.64	0.016	0.025
EcGM3	Glc	Δsdh Δ(zwf/pgl/gnd)	0.65	0.012	0.018