bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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.

Sequence Specific Modeling of *E. coli* Cell-Free Protein Synthesis

Michael Vilkhovoy,† Nicholas Horvath,† Che-Hsiao Shih,‡ Joseph Wayman,† Kara Calhoun,¶ James Swartz,¶ and Jeffrey D. Varner*,†

†*Robert Frederick Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY 14853*

‡*Davidson School of Chemical Engineering, Purdue University, West Lafayette, IN 47907*

¶*School of Chemical Engineering, Stanford University, Stanford, CA 94305*

E-mail: jdv27@cornell.edu

Phone: +1 (607) 255-4258. Fax: +1 (607) 255-9166

Abstract

Cell-free protein synthesis (CFPS) has become a widely used research tool in systems and synthetic biology. In this study, we used sequence specific constraint based modeling to evaluate the performance of an *E. coli* cell-free protein synthesis system. A core *E. coli* metabolic model, describing glycolysis, the pentose phosphate pathway, energy metabolism, amino acid biosynthesis and degradation was augmented with sequence specific descriptions of transcription and translation with effective models of promoter function. Thus, sequence specific constraint based modeling explicitly couples transcription and translation processes and the regulation of gene expression with the availability of metabolic resources. We tested this approach by simulating the expression of two model proteins: chloramphenicol acetyltransferase and dual emission green fluorescent protein, for which we have training data sets; we then

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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.

expanded the simulations to a range of therapeutically relevant proteins. Protein expression simulations were consistent with measurements for a variety of cases. We then compared optimal and experimentally constrained CFPS reactions, which suggested the experimental system over-consumed glucose and had suboptimal oxidative phosphorylation activity. Lastly, global sensitivity analysis identified the key metabolic processes that controlled the CFPS productivity, energy efficiency, and carbon yield. In summary, sequence specific constraint based modeling of CFPS offered a novel means to *a priori* estimate the performance of a cell-free system, using only a limited number of of adjustable parameters. In this study we modeled the production of a single protein, however this approach could be extended to multi-protein synthetic circuits, RNA circuits or small molecule production.

Keywords

Synthetic biology, constraint based modeling, cell-free protein synthesis

1 Introduction

Cell-free protein expression has become a widely used research tool in systems and synthetic biology, and a promising technology for personalized protein production. Cell-free systems offer many advantages for the study, manipulation and modeling of metabolism compared to *in vivo* processes. Central amongst these is direct access to metabolites and the biosynthetic machinery without the interference of a cell wall or the complications associated with cell growth. This allows interrogation of the chemical environment while the biosynthetic machinery is operating, potentially at a fine time resolution. Cell-free protein synthesis (CFPS) systems are arguably the most prominent examples of cell-free systems used today (*1*). However, CFPS is not new; CFPS in crude *E. coli* extracts has been used since the 1960s to explore fundamental biological mechanisms (*2*, *3*). Today, cell-free

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

systems are used in a variety of applications ranging from therapeutic protein production (*4*) to synthetic biology (*5*). However, if CFPS is to become a mainstream technology for advanced applications such as point of care manufacturing (*6*), we must first understand the performance limits and costs of these systems (*1*). One tool to address these questions is constraint based modeling.

Stoichiometric reconstructions of microbial metabolism, popularized by constraint based approaches such as flux balance analysis (FBA), have become standard tools to interrogate metabolism (*7*). FBA and metabolic flux analysis (MFA) (*8*), as well as convex network decomposition approaches such as elementary modes (*9*) and extreme pathways (*10*), model intracellular metabolism using the biochemical stoichiometry and other constraints such as thermodynamical feasibility (*11*, *12*) under pseudo steady state conditions. Constraint based approaches have used linear programming (*13*) to predict productivity (*14*, *15*), yield (*14*), mutant behavior (*16*), and growth phenotypes (*17*) for biochemical networks of varying complexity, including genome scale networks. Since the first genome scale stoichiometric model of *E. coli* (*18*), stoichiometric reconstructions of hundreds of organisms, including industrially important prokaryotes such as *E. coli* (*19*) and *B. subtilis* (*20*), are now available (*21*). Stoichiometric reconstructions have been expanded to include the integration of metabolism with detailed descriptions of gene expression (ME-Model) (*17*, *22*) and protein structures (GEM-PRO) (*23*, *24*). These expansions have greatly increased the scope of questions these models can explore. Thus, constraint based methods are powerful tools to estimate the performance of metabolic networks with very few adjustable parameters. However, constraint based methods are typically used to model *in vivo* processes, and have not yet been applied to cell-free metabolism.

In this study, we used sequence specific constraint based modeling to evaluate the performance of *E. coli* cell-free protein synthesis. A core *E. coli* cell-free metabolic model describing glycolysis, pentose phosphate pathway, energy metabolism, amino acid biosynthesis and degradation was developed from literature (*19*); it was then augmented with

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

sequence specific descriptions of promoter function, transcription and translation processes. Thus, the sequence specific constraint based approach explicitly coupled transcription and translation processes with the availability of metabolic resources in the CFPS reaction. We tested this approach by simulating the cell-free production of two model proteins, and then investigated the productivity, energy efficiency, and carbon yield for eight additional therapeutically relevant proteins. Productivity and carbon yield were inversely proportional to carbon number, while energy efficiency was independent of protein size. Based upon these simulations, effective correlations for the productivity, carbon yield and energy efficiency as a function of protein length were developed. These correlation models were then independently validated with a protein not in the original data set. Further, global sensitivity analysis identified the key metabolic processes that controlled CFPS performance; oxidative phosphorylation was vital to energy efficiency and carbon yield, while the translation rate was the most important factor controlling productivity. Lastly, we compared theoretically optimal metabolic flux distributions with an experimentally constrained flux distribution; this comparison suggested CFPS retained an *in vivo* operational memory that led to the overconsumption of glucose which negatively influenced carbon yield and energy efficiency. Taken together, sequence specific constraint based modeling of CFPS offered a novel means to *a priori* estimate the performance of a cell-free system, using only a limited number of of adjustable parameters. In this study we modeled the production of a single protein, but this approach could be extended to multiple protein synthetic circuits, RNA circuits (*25*) or small molecule production.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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 Results and discussion

2.1 Model derivation and validation

The cell-free stoichiometric network was constructed by removing growth associated reactions from the *i*AF1260 reconstruction of K-12 MG1655 *E. coli* (*19*), and adding deletions associated with the specific cell-free system (see Materials and Methods). We then added the transcription and translation template reactions of Allen and Palsson for the specific proteins of interest (*22*). A schematic of the metabolic network, consisting of 264 reactions and 146 species, is shown in Fig. 1A. The network described the major carbon and energy pathways, as well as amino acid biosynthesis and degradation pathways. Using this network, in combination with effective promoter models taken from Moon et al. (*26*), and literature values for cell-free culture parameters (Table 2), we simulated the sequence specific production of two model proteins: chloramphenicol acetyltransferase (CAT) and dual emission green fluorescent protein (deGFP), using different *E. coli* cell-free extracts. We calculated the transcription rate using effective promoter models, then maximized the rate of translation within biologically realistic bounds. Transcription and translation rates were subject to resource constraints encoded by the metabolic network, and transcription and translation model parameters were largely derived from literature (Table 2). The cell-free metabolic network, along with all model code and parameters, can be downloaded under an MIT software license from the Varnerlab website (*27*).

Cell-free simulations of the time evolution of CAT and deGFP production were consistent with experimental measurements (Fig. 1B and C). CAT was produced under a T7 promoter in a glucose/NMP cell-free system (*28*) using glucose as a source of carbon and energy (Fig. 1B). Apart from the first 10-15 min, the prediction of CAT abundance was consistent with the measured cell-free values (coefficient of determination, $R^2 = 0.86$). Meanwhile, deGFP was produced under a P70a promoter in TXTL 2.0 *E. coli* extract for eight hours using maltose as a carbon and energy source (Fig. 1C). The cell-free simulation

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

predicted the overall deGFP abundance, but failed to capture saturation at the end of the CFPS culture (R^2 = 0.84). Uncertainty in experimental factors such as the concentration of RNA polymerase, ribosomes, transcription and translation elongation rates, as well as the upper bounds on oxygen and glucose consumption rates (modeled as being normally distributed around the parameter values shown in Table 2), did not qualitatively alter the performance of the model (blue region, 95% confidence estimate). Together, these simulations suggested the description of transcription and translation, and its integration with metabolism encoded in the cell-free model, were consistent with experimental measurements. However, these simulations were only conducted at a single plasmid concentration of 5 nM. Thus, it was unclear if the model could capture cell-free protein synthesis for a range of plasmid concentrations.

Simulations of the cell-free deGFP titer for a range of plasmid concentrations (with all other parameters fixed) were consistent with experimental measurements (Fig. 1D). The titer at each plasmid concentration was calculated by multiplying the deGFP synthesis flux by the active time of production, approximately 8 hours in TXTL 2.0 (*29*). The mean of the ensemble (calculated by sampling the uncertainty in the model parameters) captured the saturation of deGFP production as a function of plasmid concentration (R^2 = 0.97). However, while the mean and 95% confidence estimate of the ensemble were consistent with measured deGFP levels, the model under predicted the deGFP titer at the saturating plasmid concentration of 5 nM. These results, in combination with the time-dependent CAT and deGFP simulations, validated our modeling approach, which required very few adjustable parameters. It showed that the sequence specific template reactions, metabolic network and literature parameters were sufficient to predict protein production under different promoters. Next, we analyzed the theoretical performance limits of CFPS.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

2.2 Analysis of CFPS performance

To better understand the performance of CFPS reactions, we analyzed the productivity, energy efficiency and carbon yield for the cell-free production of eight proteins with and without amino acid supplementation (Fig. 2). The expression of each of these proteins was under a P70a promoter, with the exception of CAT which was expressed using a T7 promoter. In all cases, the CFPS reaction was supplied with glucose; however, we considered different scenarios for amino acid supplementation. In the first case, the CFPS reaction was supplied with glucose and amino acids, and was able to synthesize amino acids from glucose (AAs supplied and *de novo* synthesis). In the second case, the CFPS reaction was supplied with glucose and amino acids, but *de novo* amino acid biosynthesis was not allowed (AAs supplied w/o *de novo* synthesis). This scenario was consistent with common cell-free extract preparation protocols which often involve amino acid supplementation; thus, we expected the enzymes responsible for amino acid biosynthesis to be largely absent from the CFPS reaction. In the final case, the CFPS reaction was supplied with glucose but not amino acids, and was forced to synthesize them *de novo* from glucose (*de novo* synthesis only). Eight proteins, ranging in size, were selected to evaluate CFPS performance: bone morphogenetic protein 10 (BMP10), chloramphenicol acetyltransferase (CAT), caspase 9 (CASP9), dual emission green fluorescent protein (deGFP), prothrombin (FII), coagulation factor X (FX), fibroblast growth factor 21 (FGF21), and single chain variable fragment R4 (scFvR4). An additional case was considered for CAT, where central metabolic fluxes were constrained by experimental measurements of glucose, organic and amino acids (see Supporting Information). Using these model proteins, we developed effective correlation models that predicted the productivity, energy efficiency and carbon yield given the carbon number of the protein. Finally, we independently validated the correlations with a protein not in our original data set: maltose binding protein (MBP).

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

2.2.1 Productivity

The theoretical maximum productivity for proteins expressed using a P70a promoter (μM/h) was inversely proportional to the carbon number (C_{POI}) and varied between 1 and 12 μM/h for the proteins sampled (Fig. 2A-B). The theoretical maximum productivity with and without amino acid supplementation was within a standard deviation of each other for each protein, but varied significantly between proteins. Productivity varied non-linearly with protein length; for instance, BMP10 (424 aa) had a optimal productivity of approximately 2.5 μM/h, whereas the optimal productivity of deGFP (229 aa) was approximately 8.4 μM/h. To examine the influence of protein length, we plotted the mean optimal productivity against the carbon number of each protein (Fig. 2B). The optimal productivity and protein length were related by the power-law relationship $\alpha \times (C_{POI})^{\beta}$, where $\alpha = 6.02 \times 10^6$ μM/(h·carbon number) and $\beta = -1.93$ for a P70a promoter. Interestingly, CAT did not obey the P70a power-law relationship; the relatively high productivity of CAT was due to its T7 promoter. The higher transcription rate of the T7 promoter increased the steady state level of mRNA by 34%, resulting in a higher productivity. However, CAT expressed under a P70a promoter followed the P70a power-law correlation with a productivity of approximately 8.2 $\pm$ 2.2 μM/h (predicted to be 7.2 μM/h by the optimal productivity correlation). Taken together, these simulations suggested a promoter specific relationship between the productivity and protein length. However, it was unclear if the productivity correlation was predictive for proteins not considered in the original training set.

We independently validated the productivity correlation by calculating the optimal productivity of MBP (which was not in the original training set) using the full model and the effective correlation model (Fig. 2B). The prediction error was less than 8% for an *a priori* prediction of CFPS productivity using the effective correlation. Thus, the effective productivity correlation could be used as a parameter free method to estimate optimal productivity for cell-free protein production using a P70a promoter. For CFPS

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

using other promoters, a similar correlation model could be developed. For example, maximal transcription occurs when the promoter model coefficient $u(\kappa) = 1$; the theoretical maximum productivity correlation for maximum promoter activity also followed a power-law distribution ($\alpha = 1.39 \times 10^7$ μM/(h·carbon number) and $\beta = -1.99$) (Fig. 2B, gray). The CAT value under a T7 promoter was similar to the maximal productivity as $u_{T7}(\kappa) \simeq 0.95$ given the T7 promoter model parameters used in this study (Table 2). Taken together, the maximum optimal productivity of a cell-free reaction was found to be inversely proportional to protein size, following a power-law relationship for proteins expressed under a P70a promoter.

2.2.2 Energy efficiency

The optimal energy efficiency of protein synthesis was independent of protein length, with and without amino acid supplementation (Fig. 2C-D); it was approximately 86% for the model proteins sampled. The relationship was observed to be linear, but with negligible slopes: $m_Y \times (C_{POI}) + b_Y$, where $m_Y = -4.01 \times 10^{-4}$ energy efficiency (%)/carbon number for the case with supplementation, and $m_Y = 3.03 \times 10^{-3}$ energy efficiency (%)/carbon number for the case without supplementation. The energy efficiency (y-intercept) was calculated at $b_Y = 86.20$ (%) with supplementation, and $b_Y = 67.40$ (%) without supplementation. In the presence of amino acids, energy was utilized to power CFPS instead of synthesizing amino acids; thus, a constant energy efficiency was observed regardless of the protein size. In the absence of supplementation, the energy efficiency decreased to between 68% and 76%. In this case, glucose consumption more than doubled (64% increase for CAT) compared to cases supplemented with amino acids; meanwhile, the productivity was similar for each protein (Fig. 2D). Therefore, the energy burden required for synthesizing each amino acid and powering CFPS lowered the energy efficiency. Surprisingly, without amino acid supplementation, proteins with a higher carbon number had marginally higher energy efficiency; however, this linear trend was mostly independent of protein size (R^2

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

= 0.65). Lastly, MBP was well predicted by the linear efficiency model with and without amino acid supplementation. The estimated MBP energy efficiency had a maximum error of 5% without supplementation, and an error of 1% - 3% in the presence of amino acids.

Experimentally constrained CAT simulations showed suboptimal energy efficiency (Fig. 2D, dagger). CAT production was simulated using the constraint based model in combination with experimental measurements of glucose consumption, organic and amino acid consumption and production rates (Fig. 1B). The experimentally constrained energy efficiency was 13.3 $\pm$ 5.0% compared to the theoretical maximum of approximately 84 $\pm$ 0.1%. Given that the CAT productivity was similar between the simulated and measured systems, differences in the glucose consumption rate and the ATP yield per glucose were likely responsible for the difference between the optimal and experimental systems. The glucose consumption rate was approximately 30 - 40 mM/h in the experimental system (even in the presence of amino acids). On the other hand, the constraint based simulation suggested the optimal glucose consumption rate was significantly less than the observed rate, approximately 1 - 7 mM/h (depending upon amino acid supplementation). In the constraint based simulation, the CFPS reaction produced only acetate as a byproduct, but in the experimental system acetate, lactate, pyruvate, succinate and malate all accumulated during the first hour of production. Thus, the constraint based simulation was more carbon efficient. The energy produced per unit glucose was also different between the optimal and experimentally constrained cases. In the optimal simulation, 12 ATPs were produced per unit glucose (the theoretical maximum for this network was 21), while the experimentally constrained simulation produced only 4 ATPs per glucose. Thus, approximately 120 - 160 mM ATP/hr was produced in the experimental case, in contrast to 12 - 84 mM ATP/hr for the optimal case. We know from measurements that ATP did not accumulate in the experimental system; rather, it was consumed by a variety of pathways that were not active in the optimal simulation. Thus, CFPS retained an *in vivo* like operational memory that led to the overconsumption of glucose, and counter intuitively the over production

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

of ATP. Toward understanding CFPS in terms of carbon utilization, we next explored the carbon yield.

2.2.3 Carbon yield

The theoretical maximum carbon yield was inversely proportional to protein length and varied between 40% to 64% for the proteins sampled (Fig. 2E-F). The relationship between the optimal carbon yield and carbon number was linear; $m_Y \times (C_{POI}) + b_Y$ where $m_Y = -8.03 \times 10^{-3}$ carbon yield (%)/carbon number with supplementation, and $m_Y = -6.98 \times 10^{-3}$ carbon yield (%)/carbon number without supplementation. The carbon yield (y-intercept) was $b_Y = 63.83$ (%) with supplementation, and $b_Y = 58.86$ (%) without supplementation. The linear yield models predicted the optimal carbon yield for a range of carbon numbers ($R^2 = 0.95$ with amino acid supplementation and $R^2 = 0.90$ without supplementation), irrespective of the promoter used to control protein expression. MBP also showed good agreement with the constraint based model, with a prediction error between 3% and 4% with and without amino acid supplementation. The effective yield models were approximately parallel, with a drop in carbon yield of 7% without amino acids. The difference between the cases followed from increased glucose consumption; glucose consumption increased in the absence of amino acids but the productivity remained the same. Thus, the carbon yield decreased. Taken together, the optimal carbon yield calculations (given the current metabolic network) suggested CFPS was 40-60% efficient with respect to carbon, with the balance of carbon resident in byproduct or CO_2 formation. However, these calculations assumed optimality with respect to byproduct formation. To further explore the question of carbon yield in realistic conditions, we constrained the calculation with experimentally derived glucose, organic and amino acid formation and consumption rates.

The carbon yield for experimentally constrained CAT production was 6.2% compared to the theoretical maximum of 58% (Fig. 2F, dagger). Given that the CAT productivity

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

was similar between the simulated and measured systems, differences in the glucose consumption rate and byproduct accumulation were likely responsible for the difference between the optimal and experimental systems. The translation rate has been identified as the rate-limiting step for cell-free protein synthesis (*30*, *31*), which was confirmed by the translation rate flux always hitting the upper bound in the simulation, regardless of protein. Since the CAT translation rate could not be increased, increased glucose consumption accumulated as byproducts: pyruvate, acetate, lactate, succinate, malate and CO_2.

2.3 Global sensitivity analysis

To better understand the effect of substrate utilization and the transcription/translation parameters on CFPS performance, we performed global sensitivity analysis on the productivity, energy efficiency and carbon yield for the CAT protein (Fig. 3 and Fig. 1 in Supporting Information). Surprisingly, RNAP and ribosome abundance had only a modest effect; the translation elongation rate had the largest effect on protein productivity. On the other hand, oxygen and substrate consumption had the largest and second-largest influences on the energy efficiency and carbon yield, respectively. The significance of transcription/translation parameters was robust to amino acid supplementation, with the translation rate being the most sensitive across all cases (Fig. 3A). This suggested that the translation elongation rate, and not transcription parameters, controlled productivity. Underwood and coworkers showed that an increase in ribosome levels did not significantly increase protein yields or rates; however, adding elongation factors increased protein synthesis rates by 27% (*30*). In addition, Li et al. increased the productivity by 5-fold of firefly luciferase in PURE CFPS by first improving the rate-limiting step, translation, followed by transcription by adjusting elongation factors, ribosome recycling factor, release factors, chaperones, BSA, and tRNAs (*31*). In examining substrate utilization, glucose consumption was not important for productivity in the presence of amino acid supplementation. However, its importance increased significantly when amino acids were not

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

available. On the other hand, amino acid consumption was only sensitive when amino acids synthesis reactions were blocked, as it was the only source of amino acids for CAT synthesis. The oxygen consumption rate was the most important factor controlling the energy efficiency of cell-free protein synthesis (Fig. 3B). In the model, we assumed that ATP could be produced by both substrate level and oxidative phosphorylation. Jewett and coworkers reported that oxidative phosphorylation still operated in cell-free systems, and that the protein titer decreased from 1.5-fold to 4-fold when oxidative phosphorylation reactions were inhibited in pyruvate-powered CFPS (*1*). However, it is unknown how active oxidative phosphorylation is in a glucose-powered cell-free system. Moreover, the connection between oxidative phosphorylation activity and other performance metrics, such as carbon yield, is also unclear.

To investigate the connection between carbon yield and oxidative phosphorylation further, we calculated the optimal CAT carbon yield as a function of the oxidative phosphorylation flux (Fig. 4). We calculated yield across an ensemble of 1000 flux balance solutions by varying the oxygen uptake rate with transcription and translation parameters. Oxidative phosphorylation had a strong effect on the carbon yield, both with and without amino acid supplementation. In the presence of amino acid supplementation, the carbon yield ranged from 20% to approximately 60%, depending on the oxidative phosphorylation flux. However, without amino acid supplementation, the carbon yield dropped to approximately 10%, and reached a maximum of 50%. In the absence of supplementation, a lower carbon yield was expected for the same oxidative phosphorylation flux, as glucose was utilized for both energy generation and amino acid biosynthesis. In all cases, whenever the carbon yield was below its theoretical maximum, there was an accumulation of both acetate and lactate. The experimental dataset exhibited a mixture of acetate and lactate accumulation during CAT synthesis, which suggested the CFPS reaction was not operating with optimal oxidative phosphorylation activity. Oxidative phosphorylation is a membrane associated process, while CFPS has no cell membrane.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

Jewett et al. hypothesized that membrane vesicles present in the CFPS reaction carry out oxidative phosphorylation (*1*). Toward this hypothesis, they enhanced the CAT titer by 33% when the reaction was augmented with 10 mM phosphate; they suggested the additional phosphate either enhanced oxidative phosphorylation activity or inhibited phosphatase reactions. However, the number, size, protein loading, and lifetime of these vesicles remains an open area of study.

2.4 Optimal metabolic flux distribution

Amino acid supplementation altered the optimal metabolic flux distribution predicted for CAT production (Fig. 5). To investigate the influence of amino acid supplementation, we compared the simulated metabolic flux distributions for CAT production with and without external amino acids. In the presence of amino acid supplementation, and *de novo* amino acid synthesis, there was an incomplete TCA cycle, where a combination of glucose and amino acids powered protein expression (Fig. 5A). Glucose was consumed to produce acetyl-coenzyme A, and associated byproducts, while glutamate was converted to alpha-ketoglutarate which traveled to oxaloacetic acid and pyruvate for additional amino acid biosynthesis. In the presence of amino acid supplementation, but without *de novo* amino acid biosynthesis, there was no TCA cycle flux. In this case, ATP was produced by a combination of substrate level and oxidative phosphorylation, where ubiquinone was regenerated via *nuo* activity, without relying on succinate dehydrogenase in the TCA cycle (Fig. 5B). These first two cases where amino acids were available had similar performance, and their respective metabolic flux distributions had a 99% correlation for all proteins. In the absence of amino acid supplementation (where all amino acids were synthesized *de novo* from glucose), the energy efficiency and carbon yield decreased; in this case, the TCA cycle was largely complete and there was diversion of metabolic flux into the Entner-Doudoroff pathway to produce NADPH (Fig. 5C). However, these simulations represent the theoretical optimum. To more accurately describe the metabolic flux, we constrained

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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 feasible solution space with experimental measurements and estimated the optimal CAT flux distribution.

The experimentally constrained optimal flux distribution had a 52% correlation with the theoretically optimal flux distribution (Fig. 6). Metabolic fluxes were constrained by experimental measurements of glucose, amino and organic acid consumption and production rates for the first hour of the reaction (Supporting Information) and showed good agreement with the data with a coefficient of determination of $R^2 = 0.92$ (Fig. 6B). The low similarity suggested several differences between the experimentally constrained and optimal metabolic flux distributions. The largest discrepancy was in oxidative phosphorylation, where the experimental system relied heavily on *cyd* rather than *cyo*, which was less efficient for energy generation. In addition, the experimental system had a high flux through *zwf*, yielding an abundance of NADPH used for amino acid biosynthesis. The remaining NADPH was interconverted to NADH via the *pnt*1 reaction, which was consumed to convert pyruvate to lactate (or used in oxidative phosphorylation). In contrast, the optimal solutions with amino acid supplementation had low *zwf* and *pnt*1 activity. Folate, purine, and pyrimidine metabolism, along with amino acid biosynthesis, were inactive in the optimal system, but active in the experimental system. In particular, the experimental system had high alanine and glutamine biosynthetic flux (both accumulated in the media), while there was no accumulation of amino acids in the optimal simulations. Lastly, the optimal solution produced (or consumed) only the required amount of amino acids; meanwhile, alanine, glutamine, pyruvate, lactate, acetate, malate, and succinate all accumulated in the experimental system. This accumulation contributed to the difference in flux distribution.

2.4.1 Potential alternative metabolic optima

Optimal flux distributions predicted using constraint based approaches may not always be unique. Alternative optimal solutions have the same objective value, e.g., productivity,

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

but different metabolic flux distributions. Techniques such as flux variability analysis (FVA) (*32*, *33*) or mixed-integer approaches (*34*) can estimate alternative optima. In this study, we used group knockout analysis to estimate potential alternative optimal solutions for CAT production constrained by experimental measurements (Fig. 7). Groups of reactions were removed from the metabolic network, and the translation rate was maximized. The difference between the nominal and altered system was then calculated. Knockout analysis identified pathways required for CAT production; for example, deletion of the glycolysis/gluconoeogenesis or oxidative phosphorylation pathways resulted in no CAT production. Likewise, there were pathway knockouts that had no effect on productivity or the metabolic flux distribution, such as removal of isoleucine, leucine, histidine and valine biosynthesis. Globally, the constraint based simulation reached the same optimal CAT productivity for 40% of the pairwise knockouts, while 92% of these solutions had different flux distributions compared with the wild-type. For example, one of the features of the predicted optimal metabolic flux distribution was a high flux through the Entner-Douodoroff (ED) pathway. Removal of the ED pathway had no effect on the CAT productivity compared to the absence of knockouts (Fig. 7A). Pairwise knockouts of the ED pathway and other subgroups (i.e. pentose phosphate pathway, cofactors, folate metabolism, etc.) also resulted in the same optimal CAT productivity. However, there was a difference in the flux distribution with these knockouts (Fig. 7B); thus, alternative optimal metabolic flux distributions exist for CAT production, despite experimental constraints. In addition, knockouts of amino acid biosynthesis reactions had no effect on the productivity with the exception of alanine, aspartate, asparagine, glutamate and glutamine biosynthesis reactions, since amino acids were available in the media. Ultimately, to determine the metabolic flux distribution occurring in CFPS, we need to add additional constraints to the flux estimation calculation. For example, thermodynamic feasibility constraints may result in a better depiction of the flux distribution (*11*, *12*), and ^{13}C labeling in CFPS could provide significant insight. However, while ^{13}C labeling techniques are well established

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

for *in vivo* processes (*35*), application of these techniques to CFPS remains an active area of research.

2.5 Summary and conclusions

In this study, we developed a sequence specific constraint based modeling approach to predict the performance of cell-free protein synthesis reactions. First principle predictions of the cell-free production of deGFP and CAT were in agreement with experimental measurements, for two different promoters. While we considered only the P70a and T7 promoters here, we are expanding our library of possible promoters. These promoter models, in combination with the cell-free constraints based approach, could enable the *de novo* design of circuits for optimal functionality and performance. We also developed effective correlation models for the productivity, energy efficiency and carbon yield as a function of protein size that could be used to quickly prototype CFPS reactions. Further, global sensitivity analysis identified the key metabolic processes that controlled CFPS performance; oxidative phosphorylation was vital to energy efficiency and carbon yield, while the translation rate was the most important for productivity. While this first study was promising, there are several issues to consider in future work. First, a more detailed description of transcription and translation reactions has been utilized in genome scale ME models e.g., O'Brien et al (*17*). These template reactions could be adapted to a cell-free system. This would allow us to consider important facets of protein production, such as the role of chaperones in protein folding. We would also like to include post-translation modifications such as glycosylation that are important for the production of therapeutic proteins in the next generation of models. In conclusion, we modeled the cell-free production of a single protein, however sequence specific constraint based modeling could be extended to multi-protein synthetic circuits, RNA circuits or small molecule production.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

Materials and Methods

Glucose/NMP cell-free protein synthesis.

The glucose/NMP cell-free protein synthesis reaction was performed using the S30 extract in 1.5-mL Eppendorf tubes (working volume of 15 μL) and incubated in a humidified incubator. The S30 extract was prepared from *E. coli* strain KC6 (A19 ΔtonA ΔtnaA ΔspeA ΔendA ΔsdaA ΔsdaB ΔgshA met+). This K12-derivative has several gene deletions to stabilize amino acid concentrations during the cell-free reaction. The KC6 strain was grown to approximately 3.0 OD595 in a 10-L fermenter (B. Braun, Allentown PA) on defined media with glucose as the carbon source and with the addition of 13 amino acids (alanine, arginine, cysteine, serine, aspartate, glutamate, and glutamine were excluded) (*36*). Crude S30 extract was prepared as described previously (*37*). Plasmid pK7CAT was used as the DNA template for chloramphenical acetyl transferase (CAT) expression by placing the *cat* gene between the T7 promoter and the T7 terminator (*38*). The plasmid was isolated and purified using a Plasmid Maxi Kit (Qiagen, Valencia CA). Cell-free CAT synthesis was performed at 37 °C.

The protein synthesis reaction was conducted using the PANOxSP protocol with slight modifications from that described previously (*39*). Unless otherwise noted, all reagents were purchased from Sigma (St. Louis, MO). The initial mixture included 1.2 mM ATP; 0.85 mM each of GTP, UTP, and CTP; 30 mM phosphoenolpyruvate (Roche, Indianapolis IN); 130 mM potassium glutamate; 10 mM ammonium glutamate; 16 mM magnesium glutamate; 50 mM HEPES-KOH buffer (pH 7.5); 1.5 mM spermidine; 1.0 mM putrescine; 34 μg/mL folinic acid; 170.6 μg/mL *E. coli* tRNA mixture (Roche, Indianapolis IN); 13.3 μg/mL pK7CAT plasmid; 100 μg/mL T7 RNA polymerase; 20 unlabeled amino acids at 2-3 mM each; 5 μM l-[U-^{14}C]-leucine (Amersham Pharmacia, Uppsala Sweden); 0.33 mM nicotinamide adenine dinucleotide (NAD); 0.26 mM coenzyme A (CoA); 2.7 mM sodium oxalate; and 0.24 volumes of *E. coli* S30 extract. This reaction was modified for the energy

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

source used such that glucose reactions have 30-40 mM glucose in place of PEP. Sodium oxalate was not added since it has a detrimental effect on protein synthesis and ATP concentrations when using glucose or other early glycolytic intermediate energy sources (*40*). The HEPES buffer (pKa $\sim$ 7.5) was replaced with Bis-Tris (pKa $\sim$ 6.5). In addition, the magnesium glutamate concentration was reduced to 8 mM for the glucose reaction since a lower magnesium optimum was found when using a nonphosphorylated energy source (*39*). Finally, 10 mM phosphate was added in the form of potassium phosphate dibasic adjusted to pH 7.2 with acetic acid.

Protein product and metabolite measurements.

Cell-free reaction samples were quenched at specific timepoints with equal volumes of ice-cold 150 mM sulfuric acid to precipitate proteins. Protein synthesis of CAT was determined from the total amount of ^{14}C-leucine-labeled product by trichloroacetic acid precipitation followed by scintillation counting as described previously (*28*). Samples were centrifuged for 10 min at 12,000g and 4°C. The supernatant was collected for high performance liquid chromatography (HPLC) analysis. HPLC analysis (Agilent 1100 HPLC, Palo Alto CA) was used to separate nucleotides and organic acids, including glucose. Compounds were identified and quantified by comparison to known standards for retention time and UV absorbance (260 nm for nucleotides and 210 nm for organic acids) as described previously (*28*). The standard compounds quantified with a refractive index detector included inorganic phosphate, glucose, and acetate. Pyruvate, malate, succinate, and lactate were quantified with the UV detector. The stability of the amino acids in the cell extract was determined using a Dionex Amino Acid Analysis (AAA) HPLC System (Sunnyvale, CA) that separates amino acids by gradient anion exchange (AminoPac PA10 column). Compounds were identified with pulsed amperometric electrochemical detection and by comparison to known standards.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

Formulation and solution of the model equations.

The sequence specific flux balance analysis problem was formulated as a linear program:

$$\begin{aligned} \max_{\boldsymbol{w}} \left(w_X = \boldsymbol{\theta}^T \boldsymbol{w} \right) \\ \text{Subject to : } \mathbf{Sw} = \mathbf{0} \\ \mathcal{L}_i \leq w_i \leq \mathcal{U}_i \qquad i = 1, 2, \ldots, \mathcal{R} \end{aligned} \tag{1}$$

where **S** denotes the stoichiometric matrix, **w** denotes the unknown flux vector, $\boldsymbol{\theta}$ denotes the objective cost vector and $\mathcal{L}_i$ and $\mathcal{U}_i$ denote the lower and upper bounds on flux w_i, respectively. The transcription (T) and translation (X) stoichiometry was modeled based upon the template reactions of Allen and Palsson (*22*) (Table 1). The objective of the sequence specific flux balance calculation was to maximize the rate of protein translation, w_X. The total glucose uptake rate was bounded by [0,40 mM/h] according to experimental data, while the amino acid uptake rates were bounded by [0,30 mM/h], but did not reach the maximum flux. Gene and protein sequences were taken from literature and are available in the Supporting Information. The sequence specific flux balance linear program was solved using the GNU Linear Programming Kit (GLPK) v4.55 (*41*). For all cases, amino acid degradation reactions were blocked as these enzymes were likely inactivated during the cell-free extract preparation (*28*, *29*). In the absence of *de novo* amino acid synthesis, all amino acid synthesis reactions were set to 0 mM/hr. In the experimentally constrained simulations, *E. coli* was grown in the presence of 13 amino acids (alanine, arginine, cysteine, serine, aspartate, glutamate, and glutamine were excluded) (*36*), thus the synthesis reactions responsible for those 13 amino acids were set to 0 mM/hr.

The bounds on the transcription rate ($\mathcal{L}_T = w_T = \mathcal{U}_T$) were modeled as:

$$w_T = V_T^{max} \left(\frac{G_{\mathcal{P}}}{K_T + G_{\mathcal{P}}} \right) \tag{2}$$

where $G_{\mathcal{P}}$ denotes the gene concentration of the protein of interest, and K_T denotes a

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

transcription saturation coefficient. The maximum transcription rate V_T^{max} was formulated as:

$$V_T^{max} \equiv \left[R_T \left(\frac{\dot{v}_T}{l_G} \right) u\left(\kappa\right) \right] \tag{3}$$

The term R_T denotes the RNA polymerase concentration (nM), $\dot{v}_T$ denotes the RNA polymerase elongation rate (nt/h), l_G denotes the gene length in nucleotides (nt). The term $u\left(\kappa\right)$ (dimensionless, $0 \leq u\left(\kappa\right) \leq 1$) is an effective model of promoter activity, where κ denotes promoter specific parameters. The general form for the promoter models was taken from Moon *et al.* (*26*). In this study, we considered two promoters: T7 and P70a. The promoter function for the T7 promoter, u_{T7}, was given by:

$$u_{T7} = \frac{K_{T7}}{1 + K_{T7}} \tag{4}$$

where K_{T7} denotes a T7 RNA polymerase binding constant. The P70a promoter function u_{P70a} (which was used for all other proteins) was formulated as:

$$u_{P70a} = \frac{K_1 + K_2 f_{\sigma_{70}}}{1 + K_1 + K_2 f_{\sigma_{70}}} \tag{5}$$

where K_1 denotes the weight of RNA polymerase binding alone, K_2 denotes the weight of RNAP-σ_{70} bound to the promoter, and f_{p70} denotes the fraction of the σ_{70} transcription factor bound to RNAP, modeled as a Hill function:

$$f_{\sigma_{70}} = \frac{\sigma_{70}^n}{K_D^n + \sigma_{70}^n} \tag{6}$$

where σ_{70} denotes the sigma-factor 70 concentration, K_D denotes the dissociation constant, and n denotes a cooperativity coefficient. The values for all promoter parameters are given in Table 2.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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 translation rate (w_X) was bounded by:

$$0 \leq w_X \leq V_X^{max} \left(\frac{mRNA^*}{K_X + mRNA^*} \right) \tag{7}$$

where mRNA* denotes the steady state mRNA abundance and K_X denotes a translation saturation constant. The maximum translation rate V_X^{max} was formulated as:

$$V_X^{max} \equiv \left[K_P R_X \left(\frac{\dot{v}_X}{l_P} \right) \right] \tag{8}$$

The term K_P denotes the polysome amplification constant, $\dot{v}_X$ denotes the ribosome elongation rate (amino acids per hour), l_P denotes the number of amino acids in the protein of interest. The steady state mRNA abundance mRNA* was estimated as:

$$mRNA^* \simeq \frac{w_T}{\lambda} \tag{9}$$

where λ denotes the rate constant controlling the mRNA degradation rate (hr^{-1}). All translation parameters are given in Table 2.

Calculation of energy efficiency.

Energy efficiency ($\mathcal{E}$) was calculated as the ratio of protein production to glucose consumption, written in terms of equivalent ATP molecules:

$$\mathcal{E} \equiv \frac{w_X \left(2 \cdot \mathrm{ATP_X} + \mathrm{GTP_X}\right) + w_T \left(2 \times \left(\mathrm{ATP_T} + \mathrm{CTP_T} + \mathrm{GTP_T} + \mathrm{UTP_T}\right)\right)}{\mathrm{q_{GLC}} \cdot \mathrm{ATP_{GLC}}} \tag{10}$$

where $\mathrm{ATP_T}$, $\mathrm{CTP_T}$, $\mathrm{GTP_T}$, $\mathrm{UTP_T}$ denote the stoichiometric coefficients of each energy species for the transcription of the protein of interest, $\mathrm{ATP_X}$, $\mathrm{GTP_X}$ denote the stoichiometric coefficients of ATP and GTP for the translation of the protein of interest, $\mathrm{q_{GLC}} = \mathrm{w_{GLC}}$ denotes the glucose uptake rate, and $\mathrm{ATP_{GLC}}$ denotes the equivalent ATP number pro-

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

duced per glucose. The energy species stoichiometric coefficients are available in the Supporting Information.

Calculation of the carbon yield.

The carbon yield (Y_C^{POI}) was calculated as the ratio of carbon produced as the protein of interest divided by the carbon consumed as reactants (glucose and amino acids):

$$Y_C^{POI} \equiv \frac{\mathrm{q}_{POI} \cdot C_{POI}}{\sum_{i=1}^{\mathcal{R}} \mathrm{q}_{m_i} \cdot C_{m_i}} \tag{11}$$

where q_{POI} denotes the flux of the protein of interest produced, C_{POI} denotes carbon number of the protein of interest, $\mathcal{R}$ denotes the number of reactants, q_{m_i} denotes the uptake flux of the ith reactant, and C_{m_i} denotes the carbon number of the ith reactant.

Quantification of uncertainty.

Experimental factors taken from literature, for example macromolecular concentrations or elongation rates, are uncertain. To quantify the influence of this uncertainty on model performance, we randomly sampled the expected physiological ranges for these parameters as determined from literature. An ensemble of flux distributions was calculated for the three different cases we considered: control (with amino acid synthesis and uptake), amino acid uptake without synthesis, and amino acid synthesis without uptake. The flux ensemble was calculated by randomly sampling the maximum glucose consumption rate within a range of 0 to 30 mM/h, (determined from experimental data) and randomly sampling RNA polymerase levels, ribosome levels, and elongation rates in a physiological range determined from literature. RNA polymerase levels were sampled between 60 and 80 nM, ribosome levels between 12 and 18 µM, the RNA polymerase elongation rate between 20 and 30 nt/sec, and the ribosome elongation rate between 1.5 and 3 aa/s (*29*, *30*).

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

Global sensitivity analysis.

We conducted a global sensitivity analysis using the variance-based method of Sobol to estimate which parameters controlled the performance of the cell-free protein synthesis reaction (*42*). We computed the total sensitivity index of each parameter relative to three performance objectives: productivity of the protein of interest, energy efficiency and carbon yield. We established the sampling bounds for each parameter from literature. We used the sampling method of Saltelli *et al.* (*43*) to compute a family of $N\,(2d+2)$ parameter sets which obeyed our parameter ranges, where N was a parameter proportional to the desired number of model evaluations and d was the number of parameters in the model. In our case, $N = 1000$ and $d = 7$, so the total sensitivity indices were computed from 16,000 model evaluations. The variance-based sensitivity analysis was conducted using the SALib module encoded in the Python programming language (*44*).

Potential alternative optimal metabolic flux solutions.

We identified potential alternative optimal flux distributions by performing single and pairwise reaction group knockout simulations. Reaction group knockouts were simulated by setting the flux bounds for all the reactions involved in a group to zero and then maximizing the translation rate. We grouped reactions in the cell-free network into 19 subgroups (available in Supporting Information). We computed the difference (l^2-norm) for CAT productivity in the presence and absence of pairwise reaction knockouts. Simultaneously, we computed the difference in the flux distribution (l^2-norm) for each pairwise reaction knockout compared to the flux distribution with no knockouts. Those solutions with the same or similar productivity but large changes in the metabolic flux distribution represent alternative optimal solutions.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

Acknowledgement

This study was supported by an award from the US Army and Systems Biology of Trauma Induced Coagulopathy (W911NF-10-1-0376) to J.V. for the support of M.V. The work was also supported by the Center on the Physics of Cancer Metabolism through Award Number 1U54CA210184-01 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

Supporting Information Available

The following files are available free of charge.

- Protein Sequences: DNA and protein sequences of each protein of interest.
- Supporting Information: Performance trendlines as a function of carbon number, transcription/translation stoichiometric coefficients of energy species, and experimental measurements of CAT production.
- Carbon Yield Sensitivity Analysis: Global sensitivity analysis on deGFP carbon yield.
- Metabolites and reactions of the cell-free stoichiometric network.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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 1: Transcription and translation template reactions for protein production. The symbol $G_{\mathcal{P}}$ denotes the gene encoding protein product $\mathcal{P}$, R_T denotes the concentration of RNA polymerase, $G^*_{\mathcal{P}}$ denotes the gene bounded by the RNA polymerase (open complex), η_i and α_j denote the stoichiometric coefficients for nucleotide and amino acid, respectively, Pi denotes inorganic phosphate, R_X denotes the ribosome concentration, R^*_X denotes bound ribosome, and AA_j denotes j^{th} amino acid.

Description	**Template reaction**		
Transcription initiation	$G_{\mathcal{P}} + R_T$	$\longrightarrow$	$G^*_{\mathcal{P}}$
Transcription (w_T)	$G^*_{\mathcal{P}} + \sum_{k\in\{A,C,G,U\}} \eta_k \cdot (\{k\}\, TP + H_2O)$	$\longrightarrow$	$mRNA + G_{\mathcal{P}} + R_T + \sum_{k\in\{A,C,G,U\}} \eta_k \cdot PPi$
mRNA degradation	$mRNA$	$\longrightarrow$	$\sum_{k\in\{A,C,G,U\}} \eta_k \cdot \{k\}\, MP$
Translation initiation	$mRNA + R_X$	$\longrightarrow$	R^*_X
tRNA charging	$\alpha_j \cdot \left(AA_j + tRNA + ATP + H_2O\right)$	$\longrightarrow$	$\alpha_j \cdot \left(AA_j\text{-}tRNA_j + AMP + PPi\right)$ $j = 1, 2, \ldots, 20$
Translation (w_X)	$R^*_X + \sum_{j\in\{AA\}} \alpha_j \cdot \left(AA_j\text{-}tRNA_j + 2GTP + 2H_2O\right)$	$\longrightarrow$	$\mathcal{P} + R_X + mRNA + \sum_{j\in\{AA\}} \alpha_j \cdot \left(tRNA + 2GDP + 2Pi\right)$

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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: Parameters for sequence specific flux balance analysis

Description	Parameter	Value	Units	Reference
RNA polymerase concentration	R_T	75	nM	(*29*)
Ribosome concentration	R_X	1.6	μM	(*29*, *30*)
Transcription elongation rate	$\dot{v}_T$	25	nt/sec	(*29*)
Translation elongation rate	$\dot{v}_X$	2	aa/sec	(*29*, *30*)
Transcription saturation coefficient	K_T	3.5	nM	estimated
Translation saturation coefficient	K_X	45.0	μM	estimated
Polysome amplification constant	K_P	10	constant	estimated
mRNA degradation rate	λ	5.2	hr^{-1}	(*29*)
T7 promoter	K_{T7}	10	constant	estimated
Weight RNA polymerase binding alone P70a	K_1	0.014	constant	estimated
Weight bound RNAP-σ_{70} P70a	K_2	10	constant	estimated
σ_{70} concentration	σ_{70}	35	nM	(*29*)
σ_{70} dissociation constant	K_D	130	nM	(*45*)
σ_{70} hill coefficient	n	1	constant	(*45*)
Gene concentration	$G_{\mathcal{P}}$	5	nM	(*29*)

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

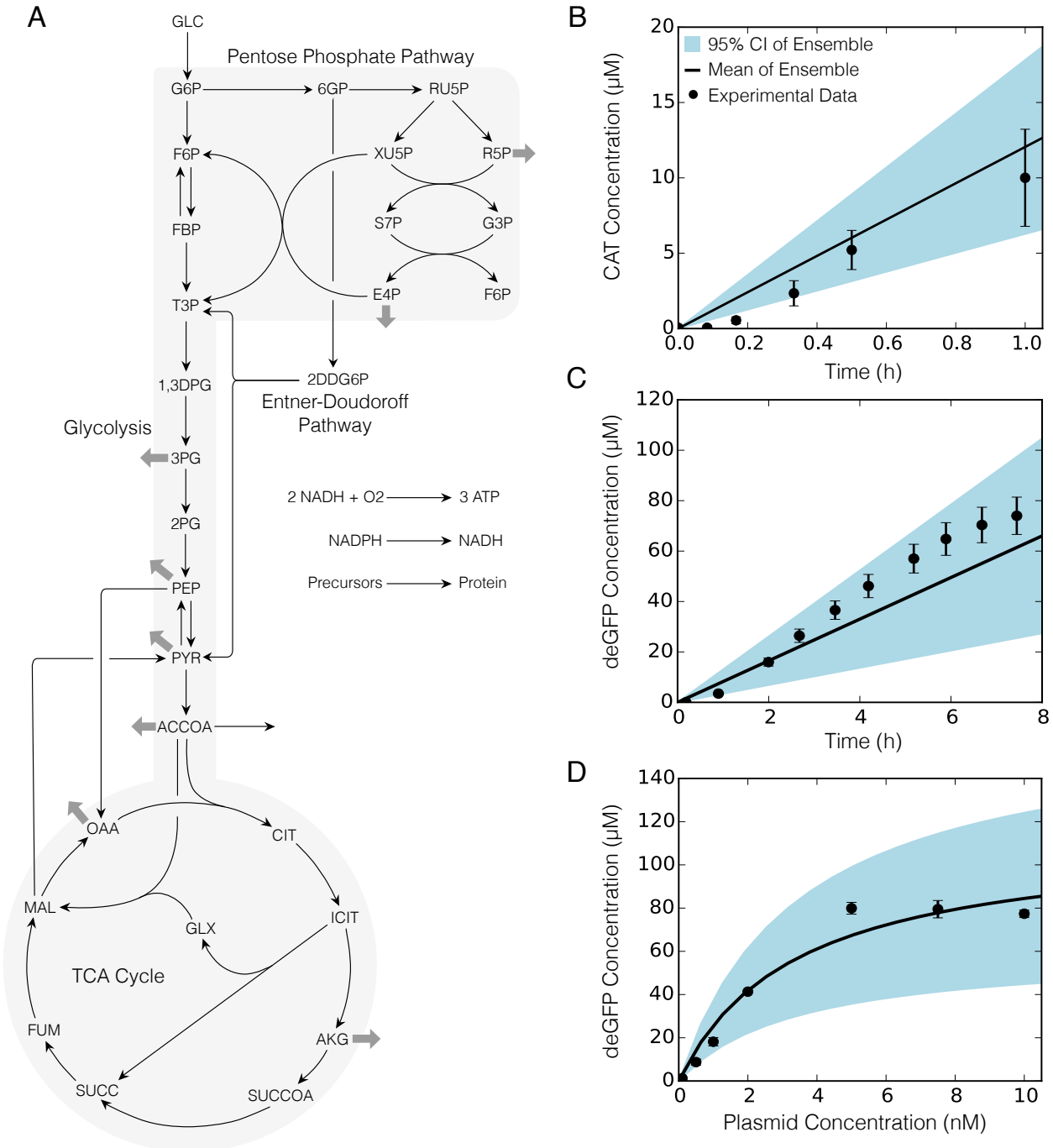

Figure 1: Sequence specific flux balance analysis. A. Schematic of the core metabolic network describing glycolysis, pentose phosphate pathway, the TCA cycle and the Entner-Doudoroff pathway. Thick gray arrows indicate withdrawal of precursors for amino acid synthesis. B. CAT production under a T7 promoter in CFPS *E. coli* extract for 1 h using glucose as a carbon and energy source. Error bars denote the standard deviation of experimental measurements. C. deGFP production under a P70 promoter in TXTL 2.0 *E. coli* extract for 8 h using glucose as a carbon and energy source. Error bars denote a 10% coefficient of variation. D. Predicted versus measured deGFP concentration as a function of plasmid concentration in TXTL 2.0. The blue region denotes the 95% CI over an ensemble of N = 100 sets, the black line denotes the mean of the ensemble, and dots denote experimental measurements.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

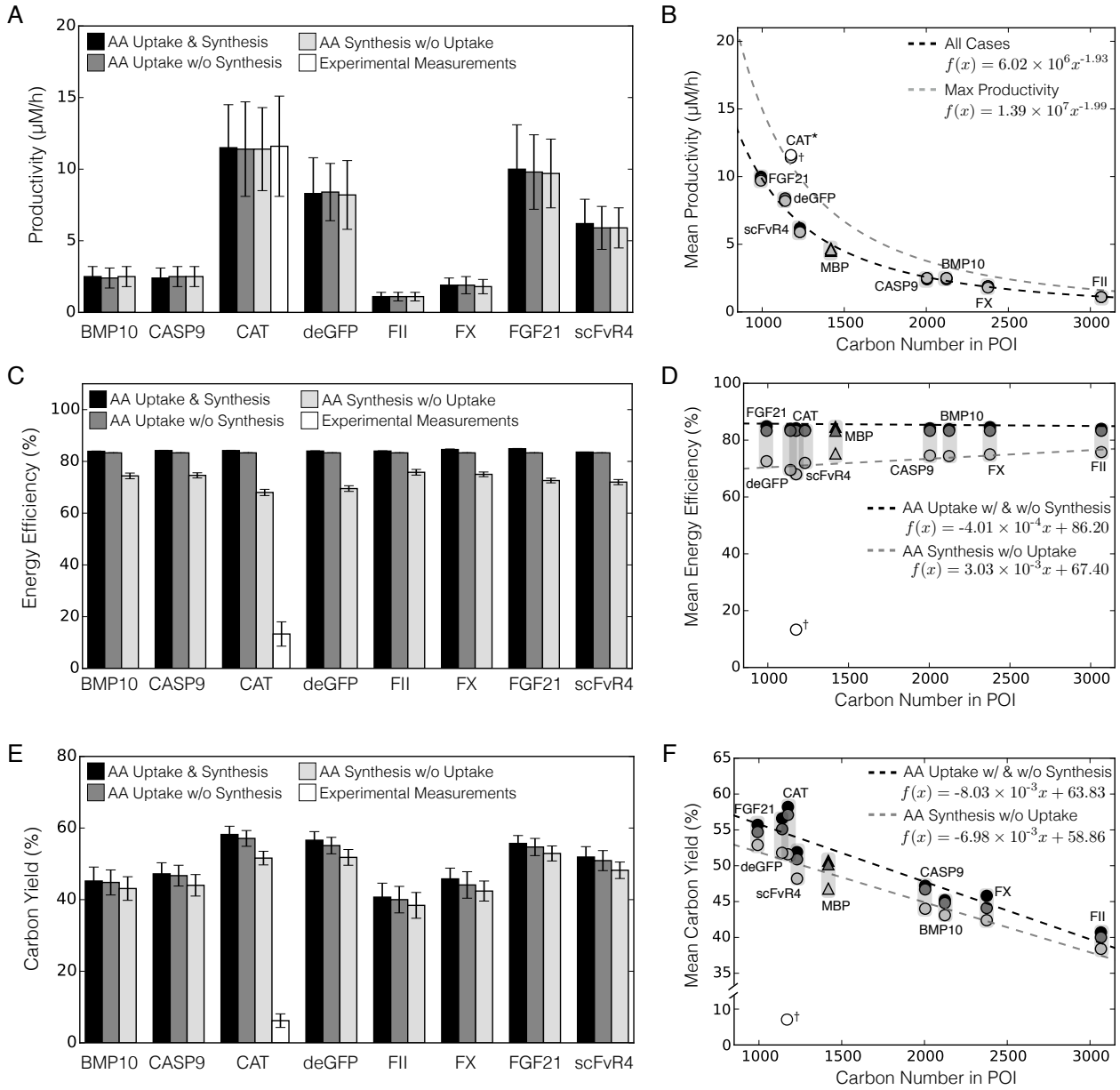

Figure 2: The CFPS performance for eight model proteins with and without amino acid supplementation. A. Mean CFPS productivity for a panel of model proteins with and without amino acid supplementation. B. Mean CFPS productivity versus carbon number for a panel of model proteins with and without amino acid supplementation. Trendline (black dotted line) was calculated across all cases for a P70a promoter (R^2 = 0.99) and maximum productivity trendline assumed $u(\kappa) = 1$ (grey dotted line; R^2 = 0.99). C. Mean CFPS energy efficiency for a panel of model proteins with and without amino acid supplementation. D. Mean CFPS energy efficiency versus carbon number for a panel of model proteins with and without amino acid supplementation. Trendline for cases with amino acids (black dotted line) and trendline for without amino acids (grey dotted line; R^2 = 0.65). E. Mean CFPS carbon yield for a panel of model proteins with and without amino acid supplementation. F. Mean CFPS carbon yield versus carbon number. Trendline for cases with amino acids (black dotted line; R^2 = 0.95) and trendline for without amino acids (grey dotted line; R^2 = 0.90). Error bars: 95% CI calculated by sampling; asterisk: protein excluded from trendline; dagger: constrained by experimental measurements and excluded from trendline; triangles: first principle prediction and excluded from formulating trendline.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

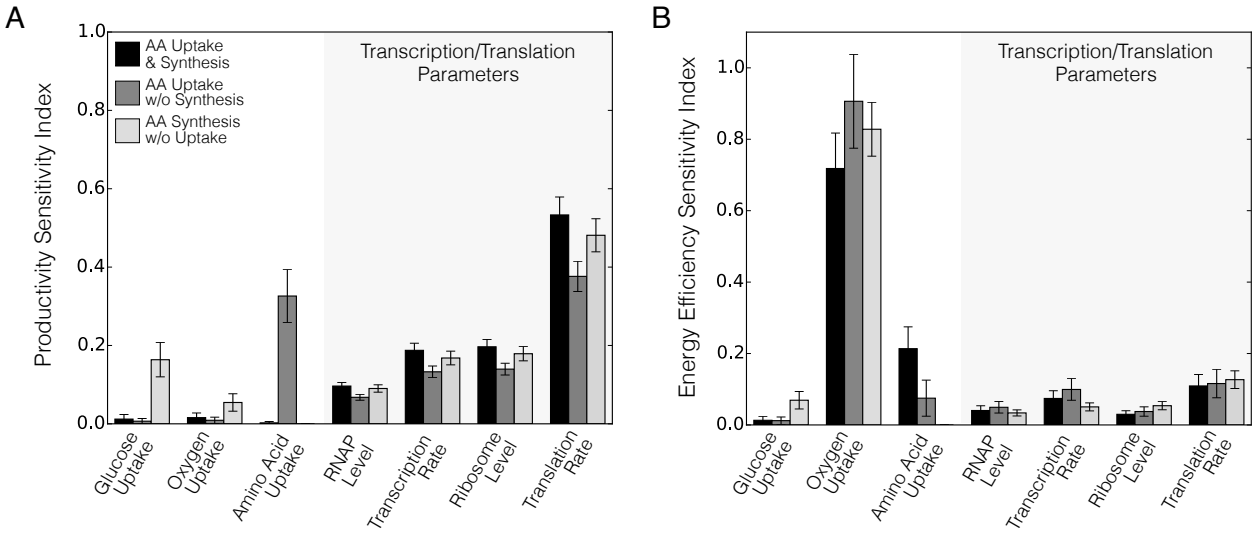

Figure 3: Total order sensitivity analysis of the cell free production of CAT. A. Total order sensitivity of the optimal CAT productivity with respect to metabolic and transcription/translation parameters. B. Total order sensitivity of the optimal CAT energy efficiency. Metabolic and transcription/translation parameters were varied for amino acid supplementation and synthesis (black), amino acid supplementation without synthesis (dark grey) and amino acid synthesis without supplementation (light gray). Error bars represent the 95% CI of the total order sensitivity index.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

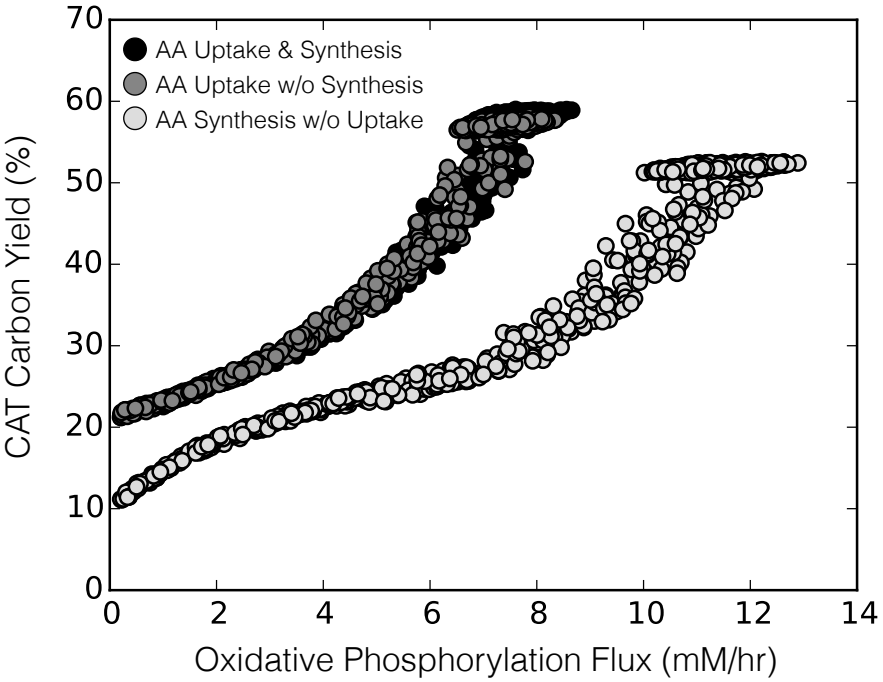

Figure 4: Optimal CAT carbon yield versus oxidative phosphorylation flux calculated across an ensemble (N = 1000) of flux balance solutions (points). Carbon yield versus oxidative phosphorylation flux for amino acid supplementation and *de novo* synthesis (black), amino acid supplementation without *de novo* synthesis (dark grey), and *de novo* amino acid synthesis without supplementation (light gray). The ensemble was generated by randomly varying the oxygen consumption rate from 0.1 to 10 mM/h and randomly sampling the transcription and translation parameters within 10% of their literature values. Each point represents one solution of the ensemble.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

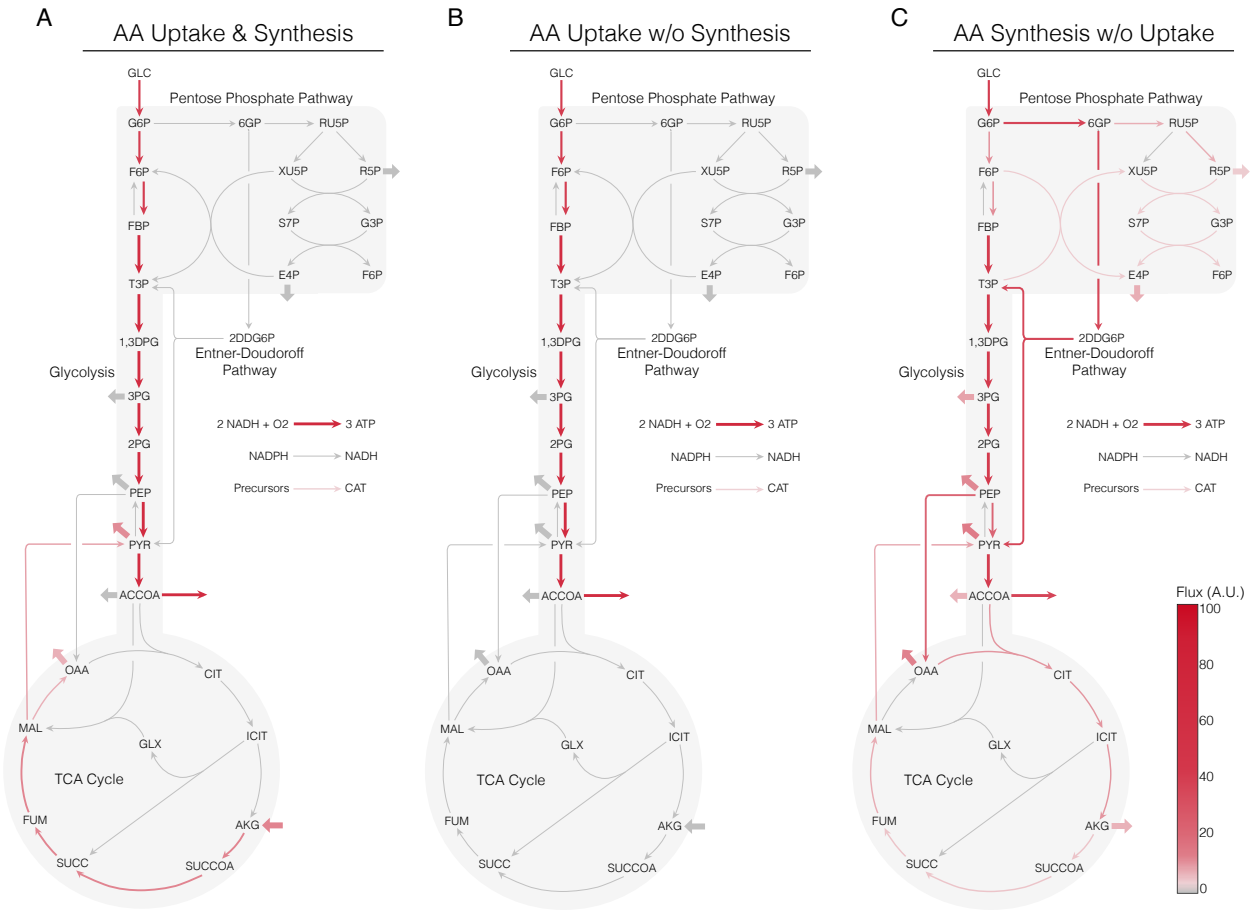

Figure 5: Optimal metabolic flux distribution for CAT production. A. Optimal flux distribution in the presence of amino acid supplementation and *de novo* synthesis. B. Optimal flux distribution in the presence of amino acid supplementation without *de novo* synthesis. C. Optimal flux distribution with *de novo* amino acid synthesis in the absence of supplementation. Mean flux across the ensemble (N = 100), normalized to glucose uptake flux. Thick arrows indicate flux to or from amino acid biosynthesis pathways.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

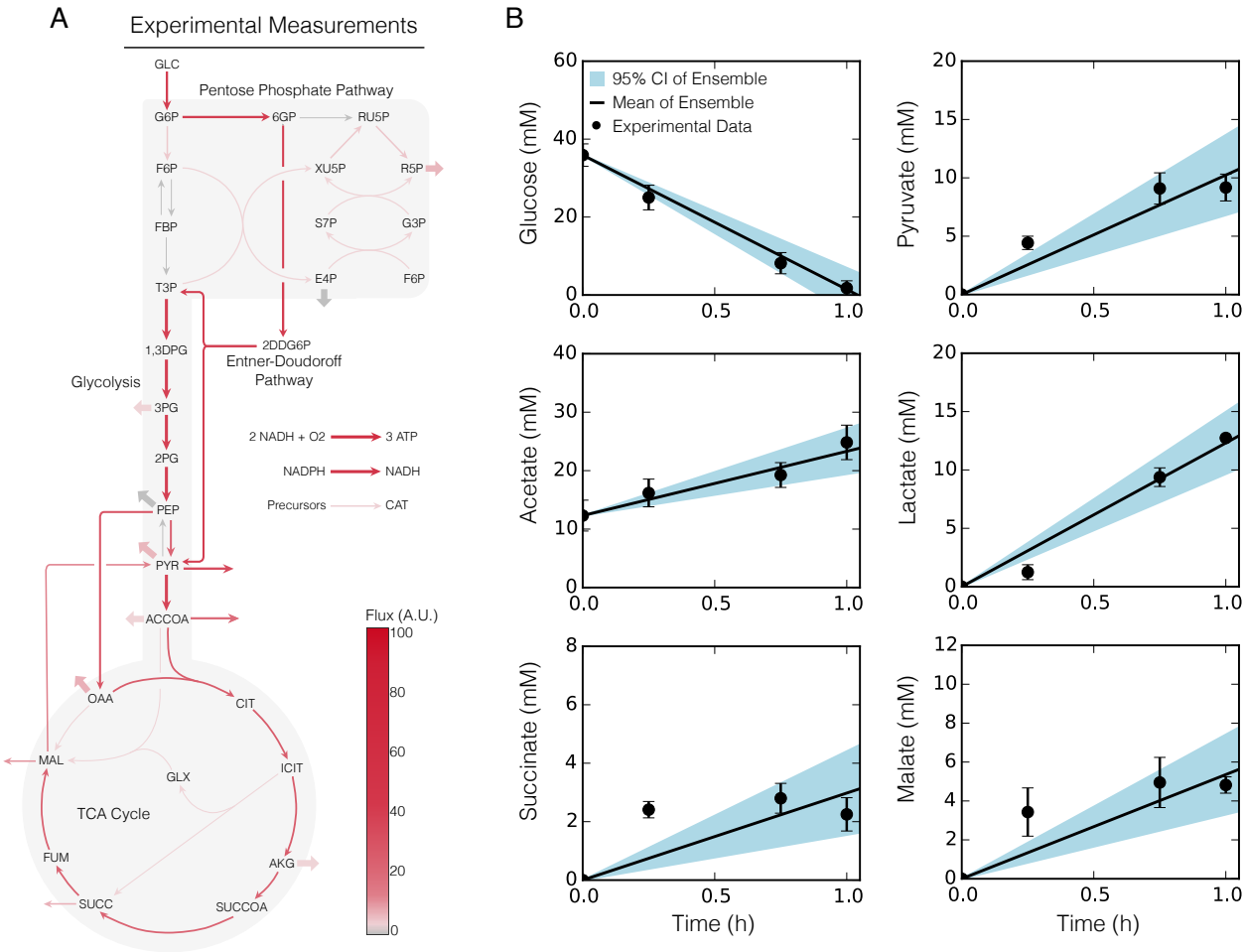

Figure 6: Experimentally constraint based simulation of CAT production. A. Metabolic flux distribution for CAT production in the presence of experimental constraints for glucose, organic acid and amino acid consumption and production rates. Mean flux across the ensemble, normalized to glucose uptake flux. Thick arrows indicate flux to or from amino acids. B. Central carbon metabolite measurements versus simulations over a 1 hour time course. The blue region denotes the 95% CI over an ensemble of N = 100 sets, the black line denotes the mean of the ensemble, and dots denote experimental measurements.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

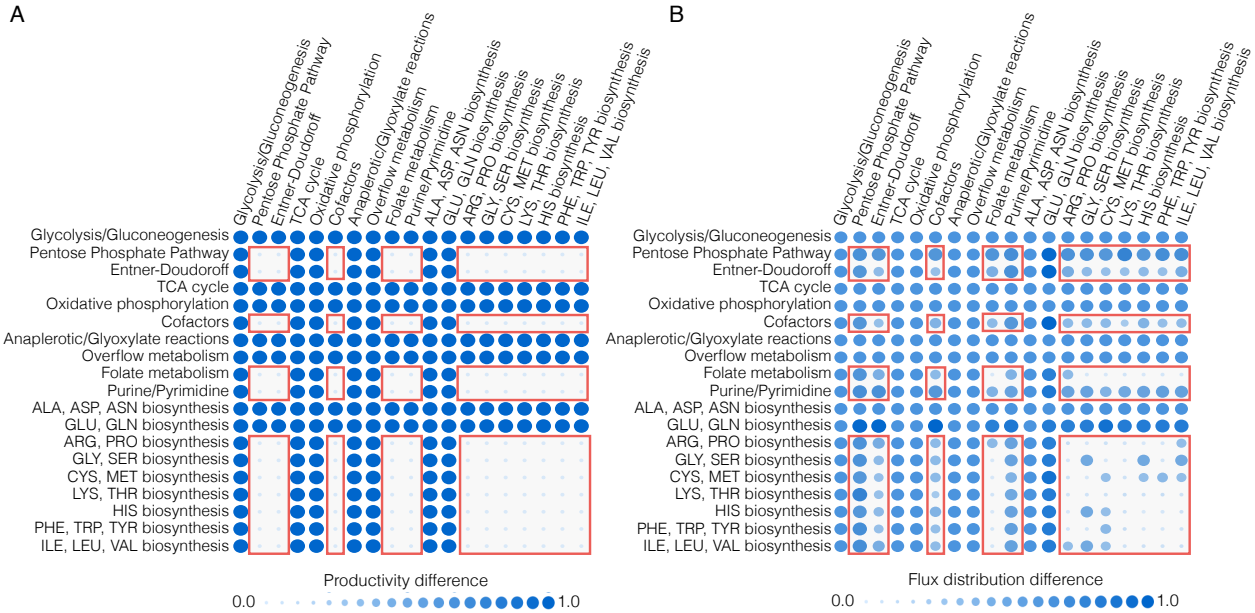

Figure 7: Pairwise knockouts of reaction subgroups in the cell-free network. A. Difference in the CAT productivity in the presence of reaction knockouts compared with no knockouts for experimentally constrained CAT production. B. Difference in the optimal flux distribution in the presence of reaction knockouts compared with no knockouts for experimentally constrained CAT production. The difference between perturbed and wild-type productivity and flux distributions was quantified by the l^2 norm, and then normalized so the maximum change was 1.0. Red boxes indicate potential alternative optimal flux distributions.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

References

1. Jewett, M. C., Calhoun, K. A., Voloshin, A., Wuu, J. J., and Swartz, J. R. (2008) An integrated cell-free metabolic platform for protein production and synthetic biology. *Mol Syst Biol 4*, 220.

2. Matthaei, J. H., and Nirenberg, M. W. (1961) Characteristics and stabilization of DNAase-sensitive protein synthesis in E. coli extracts. *Proc Natl Acad Sci U S A 47*, 1580–8.

3. Nirenberg, M. W., and Matthaei, J. H. (1961) The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U S A 47*, 1588–602.

4. Lu, Y., Welsh, J. P., and Swartz, J. R. (2014) Production and stabilization of the trimeric influenza hemagglutinin stem domain for potentially broadly protective influenza vaccines. *Proc Natl Acad Sci U S A 111*, 125–30.

5. Hodgman, C. E., and Jewett, M. C. (2012) Cell-free synthetic biology: thinking outside the cell. *Metab Eng 14*, 261–9.

6. Pardee, K., Slomovic, S., Nguyen, P. Q., Lee, J. W., Donghia, N., Burrill, D., Ferrante, T., McSorley, F. R., Furuta, Y., Vernet, A., Lewandowski, M., Boddy, C. N., Joshi, N. S., and Collins, J. J. (2016) Portable, On-Demand Biomolecular Manufacturing. *Cell 167*, 248–59.e12.

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

8. Wiechert, W. (2001) 13C Metabolic Flux Analysis. *Metabol. Eng. 3*, 195 – 206.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

9. Schuster, S., Fell, D. A., and Dandekar, T. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol 18*, 326–32.

10. Schilling, C. H., Letscher, D., and Palsson, B. O. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol 203*, 229–48.

11. Henry, C. S., Broadbelt, L. J., and Hatzimanikatis, V. (2006) Thermodynamics-Based Metabolic Flux Analysis. *Biophys. J 92*, 192–1805.

12. Hamilton, J. J., Dwivedi, V., and Reed, J. L. (2013) Quantitative Assessment of Thermodynamic Constraints on the Solution Space of Genome-Scale Metabolic Models. *Biophys J 105*, 512–522.

13. Covert, M. W., Knight, E. M., Reed, J. L., Herrgård, M. J., and Palsson, B. Ø. (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature 429*, 92–6.

14. Varma, A., and Palsson, B. O. (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Appl. Environ. Microbiol. 60*, 3724–3731.

15. Sánchez, C., Quintero, J. C., and Ochoa, S. (2015) Flux balance analysis in the production of clavulanic acid by Streptomyces clavuligerus. *Biotechnol Prog 31*, 1226–1236.

16. Edwards, J. S., and Palsson, B. O. (2000) Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions. *BMC Bioinformatics 1*, 1.

17. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R., and Palsson, B. Ø. (2013) Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Sys. Biol. 9*, 693.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

18. Edwards, J. S., and Palsson, B. Ø. (2000) The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A 97*, 5528–33.

19. 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. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol 3*, 121.

20. Oh, Y.-K., Palsson, B. Ø., Park, S. M., Schilling, C. H., and Mahadevan, R. (2007) Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J Biol Chem 282*, 28791–9.

21. Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., and Palsson, B. Ø. (2009) Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol 7*, 129–43.

22. Allen, T. E., and Palsson, B. Ø. (2003) Sequence-based analysis of metabolic demands for protein synthesis in prokaryotes. *J Theor Biol 220*, 1–18.

23. Zhang, Y., Thiele, I., Weekes, D., Li, Z., Jaroszewski, L., Ginalski, K., Deacon, A. M., Wooley, J., Lesley, S. A., Wilson, I. A., Palsson, B., Osterman, A., and Godzik, A. (2009) Three-Dimensional Structural View of the Central Metabolic Network of Thermotoga maritima. *Science 325*, 1544–1549.

24. Chang, R. L., Andrews, K., Kim, D., Li, Z., Godzik, A., and Palsson, B. O. (2013) Structural Systems Biology Evaluation of Metabolic Thermotolerance in Escherichia coli. *Science 340*, 1220–1223.

25. Hu, C. Y., Varner, J. D., and Lucks, J. B. (2015) Generating Effective Models and Parameters for RNA Genetic Circuits. *ACS Synth Biol 4*, 914–26.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

26. Moon, T. S., Lou, C., Tamsir, A., Stanton, B. C., and Voigt, C. A. (2012) Genetic programs constructed from layered logic gates in single cells. *Nature 491*, 249–53.

27. Varnerlab, http://www.varnerlab.org/downloads/.

28. Calhoun, K. A., and Swartz, J. R. (2005) An Economical Method for Cell-Free Protein Synthesis using Glucose and Nucleoside Monophosphates. *Biotechnol Prog 21*, 1146–53.

29. Garamella, J., Marshall, R., Rustad, M., and Noireaux, V. (2016) The All E. coli TX-TL Toolbox 2.0: A Platform for Cell-Free Synthetic Biology. *ACS Synth Biol 5*, 344–55.

30. Underwood, K. A., Swartz, J. R., and Puglisi, J. D. (2005) Quantitative polysome analysis identifies limitations in bacterial cell-free protein synthesis. *Biotech. Bioeng. 91*, 425–35.

31. Li, J., Gu, L., Aach, J., and Church, G. M. (2014) Improved Cell-Free RNA and Protein Synthesis System. *PLoS ONE 9*, 1–11.

32. Mahadevan, R., and Schilling, C. (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng 5*, 264 – 276.

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

34. Lee, S., Phalakornkule, C., Domach, M. M., and Grossmann, I. E. (2000) Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comput. Chem. Eng. 24*, 711 – 716.

35. Zamboni, N., Fendt, S.-M., and Sauer, U. (2009) ^{13}C-based metabolic flux analysis. *Nature Protocols 4*, 878–92.

36. Zawada, J., Richter, B., Huang, E., Lodes, E., Shah, A., and Swartz, J. R. *Fermentation Biotechnology*; Chapter 9, pp 142–156.

bioRxiv preprint doi: https://doi.org/10.1101/139774; this version posted May 23, 2017. 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.

37. Jewett, M., Voloshin, A., and Swartz, J. In *Gene cloning and expression technologies*; Weiner, M., and Lu, Q., Eds.; Eaton Publishing: Westborough, MA, 2002; pp 391–411.

38. Kigawa, T., Muto, Y., and Yokoyama, S. (1995) Cell-free synthesis and amino acid-selective stable isotope labeling of proteins for NMR analysis. *J Biomolec NMR 6*, 129–134.

39. Jewett, M. C., and Swartz, J. R. (2004) Mimicking the Escherichia coli cytoplasmic environment activates long-lived and efficient cell-free protein synthesis. *Biotech. Bioeng. 86*, 19–26.

40. Kim, D.-M., and Swartz, J. R. (2001) Regeneration of adenosine triphosphate from glycolytic intermediates for cell-free protein synthesis. *Biotech. Bioeng. 74*, 309–316.

41. GNU Linear Programming Kit, Version 4.52. 2016; `http://www.gnu.org/software/glpk/glpk.html`.

42. Sobol, I. (2001) Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation 55*, 271–80.

43. Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. (2010) Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Comp Phys Comm 181*, 259–70.

44. Herman, J. D. http://jdherman.github.io/SALib/.

45. Mauri, M., and Klumpp, S. (2014) A model for sigma factor competition in bacterial cells. *PLoS Comput Biol 10*, e1003845.