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

Michael Vilkhovoy,[†] Nicholas Horvath,[†] Joseph Wayman,[†] Kara Calhoun,[‡] James Swartz,[‡] and Jeffrey D. Varner[*,†]

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

‡*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 expression has become a widely used research tool in systems and synthetic biology, and a promising technology for personalized medicine. 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, amino acid biosynthesis and degradation and energy metabolism, was augmented with sequence specific descriptions of transcription and translation processes, and 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 expanded the simulations

to a range of therapeutically relevant proteins. Protein expression simulations were consistent with measurements for a variety of cases. Further, global sensitivity analysis identified the key metabolic processes that controlled the productivity, energy efficiency, and carbon yield of the process. Taken together, sequence specific constraint based modeling offers a novel means to *a priori* estimate the performance of cell-free synthetic circuits.

Keywords

Synthetic biology, constraints 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 complications associated with cell growth. This allows us to interrogate 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 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 applications such as point of care manufacturing, we must first understand the performance limits of these systems. One tool to address this question is mathematical modeling.

Stoichiometric reconstructions of microbial metabolism, popularized by flux balance analysis (FBA), have become standard tools to interrogate metabolism (*6*). Since the first genome scale stoichiometric model of *E. coli* (*7*), stoichiometric reconstructions of hundreds of organisms, including industrially important prokaryotes such as *E. coli* (*8*) or *B. subtilis* (*9*), are now available (*10*). In this study, we used sequence specific constraints based modeling to evaluate the performance of *E. coli* cell-free protein synthesis (CFPS). A core *E. coli* cell-free metabolic model was developed from literature (*8*). This model, which described glycolysis, pentose phosphate pathway, amino acid biosynthesis and degradation and energy metabolism, was then augmented with sequence specific descriptions of promoter function, transcription and translation processes. Thus, sequence specific constraints based modeling explicitly coupled transcription and translation with the availability of metabolic resources. We tested this approach by simulating the production of two model proteins, and then investigated the productivity, energy efficiency, and carbon yield for eight different proteins. From this, higher carbon number proteins typically had lower productivity rates, energy efficiency, and carbon yields than that of the lower carbon number proteins. Further, global sensitivity analysis identified the key metabolic processes that controlled circuit performance, showing oxidative phosphorylation as instrumental for maintaining a high energy efficiency and carbon yield and the translation rate for productivity. Taken together, sequence specific constraints based modeling offers a novel means to *a priori* estimate the performance of cell-free synthetic circuits.

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* (*8*). We then added the transcription and translation template reactions of Allen and Palsson for the specific proteins of interest (*11*). A schematic of the metabolic network, which consisted of 264 reactions and 146 species, is shown in Fig. 1A. Using this network, in combination with detailed promoter models, 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 cell-free *E. coli* extracts. The cell-free metabolic network, model parameters, model code, and each protein sequence are available in the supplemental materials.

Cell-free simulations predicted CAT and deGFP production for the duration of the CFPS batch reactions (Fig. 1B and C). Chloramphenicol acetyltransferase (CAT), was produced under a T7 promoter in a glucose/NMP cell-free system (*12*) for 1 hour using glucose as a carbon and energy source (Fig. 1B). With the exception of the first 10-15 min, the cell-free prediction of CAT abundance was consistent with the measured values. On the other hand, deGFP was produced under a P70a promoter in TXTL 2.0 *E. coli* extract for 8 hours using maltose as a carbon and energy source (Fig. 1C). The cell-free simulation captured the overall trend of deGFP abundance, but was not able to capture saturation at the end of the CFPS culture. Uncertainty in experimental factors such as the concentration of RNA polymerase, ribosomes, transcription and translation elongation rates and the upper bounds for oxygen and glucose consumption rates did not alter the qualitative performance of the model. Thus, the metabolic network and molecular description of transcription and translation were consistent with experimental measurements.

Next, we predicted deGFP production as a function of plasmid concentration (Fig. 1D).

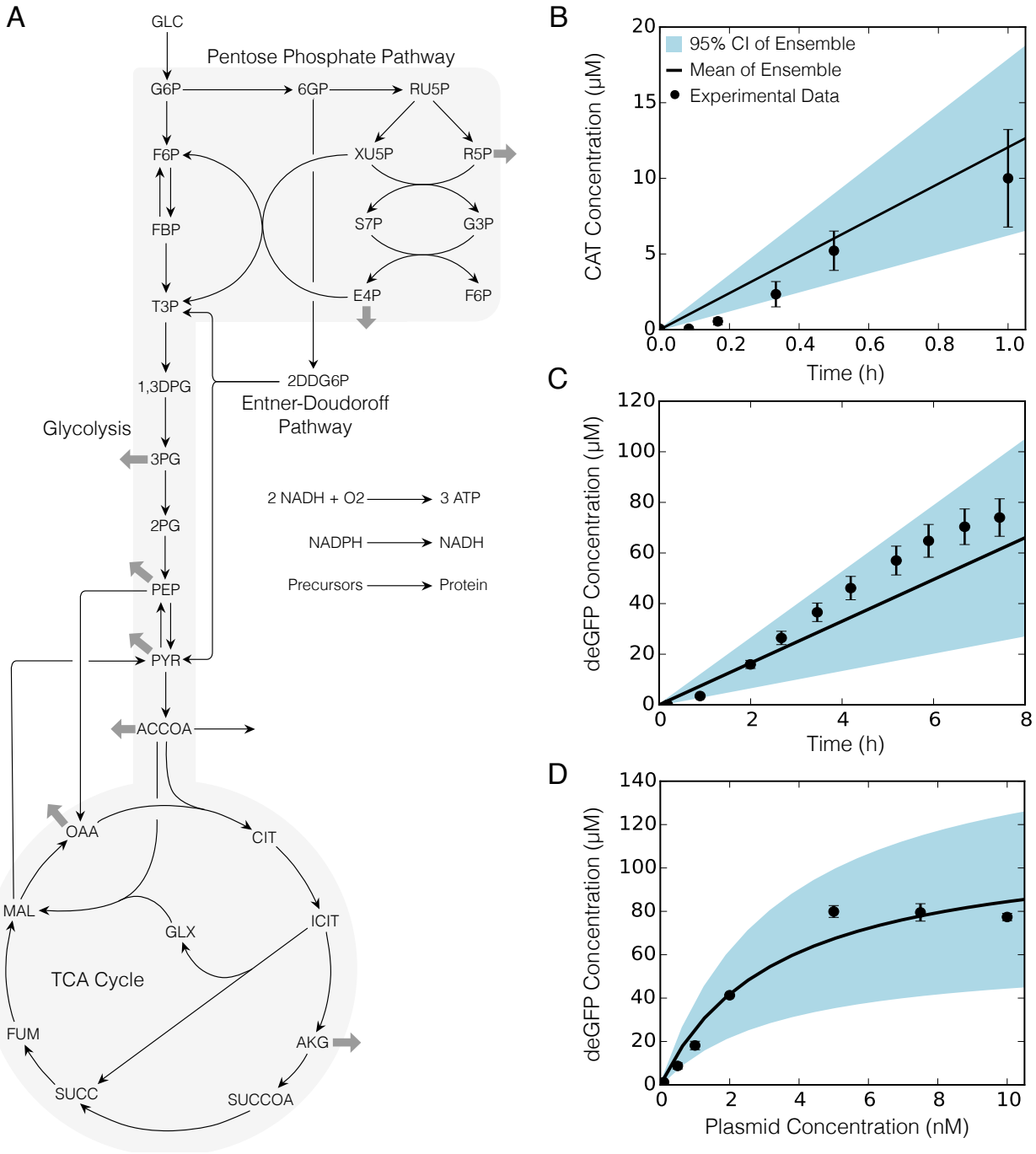

Figure 1: Sequence specific flux balance analysis. A. Core metabolic network with glycolysis, pentose phoshate pathway, TCA cycle, Entner-Doudoroff pathway. Thick gray arrows indicate withdrawal of precursors for amino acid synthesis. B. CAT production under a T7 promoter in CFPS *E. coli* extract for 1 h under glucose consumption. 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 under maltose consumption. Error bars denote a 10% coefficient of variation. D. Predicted deGFP concentration at different plasmid concentrations versus measurements of deGFP synthesized in TXTL 2.0. 95% CI (blue region) over the ensemble of 100 sets, mean of the ensemble (black line), and experimental measurements (dots).

Concentration of deGFP at each plasmid concentration was calculated by multiplying the flux of deGFP synthesis by the active time of production, approximately 8 hours in TXTL 2.0 (*13*). The mean of the ensemble shows a good prediction against the measured deGFP levels, even though it under predicted deGFP concentration at the saturating point of 5 nM of plasmid concentration. However, the ensemble and the mean of the ensemble captured the overall saturating dynamics of deGFP production as a function of plasmid concentration. These results validated our mathematical framework to model CFPS systems and predict the production of two proteins with very few adjustable parameters. It also showed that the sequence specific reactions were sufficient to predict the production of two different proteins under different promoters and cell-free systems. Since the model accurately predicted protein production, we used our mathematical framework to understand the performance limits of CFPS.

2.2 Analysis of CFPS performance

Our next goal was to examine the performance of CFPS for eight different proteins under three different cases. Each of the proteins was produced under a P70a promoter, except for CAT which was produced under a T7 promoter. In all cases, CFPS was supplied with glucose. In the first case, CFPS was supplied with amino acids, and the system was allowed to synthesize amino acids (AA uptake and synthesis). In the second case, CFPS was supplied with amino acids, but the amino acid synthesis reactions were turned off (AA uptake w/o synthesis). These amino acid synthesis reactions were blocked since during the cell-free extract preparation the cells are often supplied with amino acids; thus, the enzymes responsible for amino acid synthesis would not be present. In the third case, CFPS was not supplied with amino acids, but the system could synthesize them (AA synthesis w/o uptake). Eight different 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). We used ssFBA to estimate the productivity, energy efficiency, and carbon yield for each of these proteins for each case. An additional case was considered for CAT, since a comprehensive dataset is available (Supporting Information); in this case, fluxes were constrained to experimental measurements where avaliable, with the exception of CAT production which was determined by the transcription/translation parameters.

2.2.1 Productivity

The mean productivity was inversely proportional to the protein size and varied between 1 and 12 μM/h for the proteins sampled (Fig. 2). All cases had very similar performance for each protein and were within a standard deviation of each other (Fig. 2A). The model framework is setup to optimize for the production of each protein and is constrained by the translation rate. This shows the system had sufficient substrates and metabolic precursors to power CFPS and synthesize each protein of interest with the same productivity, regardless of the case. However, each individual protein had a different level of productivity. For instance, BMP10 had a productivity of about 2.5 μM/h whereas CAT had a productivity of about 12 μM/h. To examine this further, the mean productivity was plotted against the carbon number of each protein (Fig. 2B). The proteins with the highest productivity had the lowest carbon number, whereas proteins with low productivity had higher carbon numbers. This inverse trend was due to the fact that larger proteins require more amino acids and substrates to assemble them, resulting in lower productivity given the same resources. A single trendline for all cases shows the expected productivity in CFPS depending on the carbon number of the protein of interest (available in the Supporting Information). CAT was an outlier for the trendline, even though it was in the same order of magnitude as the trendline. The relative high productivity of CAT was due to its T7 promoter. CAT on a P70a promoter frollowed the same trendline as the other proteins and

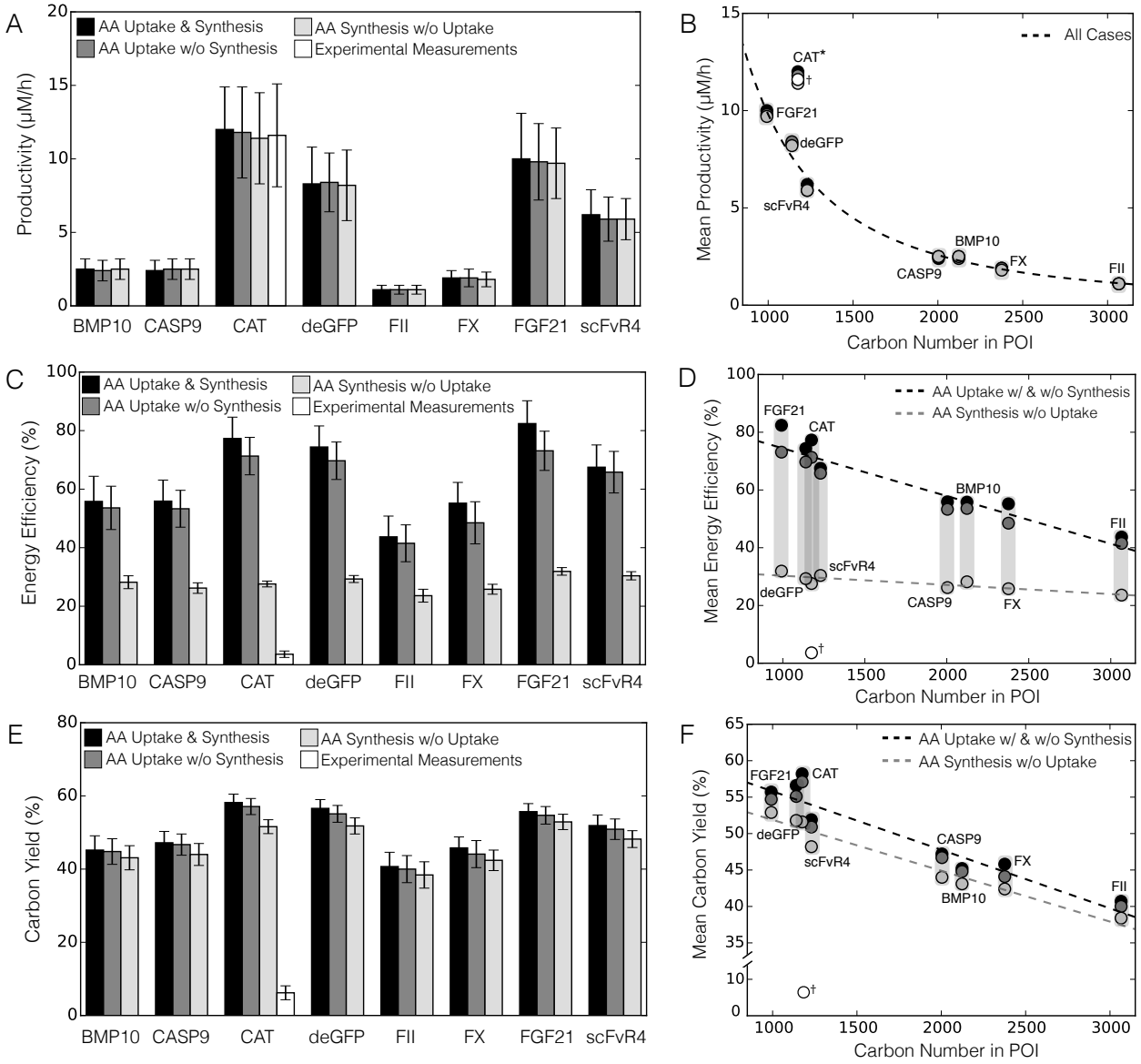

Figure 2: CFPS productivity of eight proteins for four cases. Amino acid uptake and synthesis (black), AA uptake without synthesis (dark grey), AA synthesis without uptake (light grey), and constrained by experimental measurements, for CAT only (white). A. Productivity across the ensemble (error bars represent 95% CI). B. Mean productivity versus carbon number. Single trendline (dotted line) calculated across all cases ($y = 6015004x^{-1.9}$; $R^2 = 0.99$). Asterisk: protein excluded from trendline; dagger: constrained by experimental measurements and excluded from trendline.

had a productivity of about 8.8 μM/h. Taken together, a single trendline showed good agreement ($R^2 = 0.99$) with estimating the expected productivity of a protein on a P70a promoter in CFPS, given the size of the protein.

2.2.2 Energy efficiency

The energy efficiency of protein production remained relatively high for when amino acids were supplied in the media, however energy efficiency dropped below 32% when amino acids were not available (Fig. **??**). Following the same outline as in examining the productivity, we calculated the energy efficiency of production for each protein. When amino acids were supplied in the media (first two cases), there was a comparable performance in having the highest energy efficiency, despite the network's ability to synthesis amino acids. For the case when amino acids were removed from the media, protein production resulted in a low energy efficiency; below 32% depending on the protein. This was because glucose had to be utilized to synthesize the amino acids necessary for protein synthesis and meet the necessary energy demands of CFPS. We next investigated the effect of protein carbon number on energy efficiency (Fig. **??**B). The same inverse trend was observed as for productivity, except that it was linear. Proteins with a high carbon number had a lower energy efficiency since they have a higher transcription and translation cost then smaller proteins. The cases supplemented with amino acids had the same trendline whereas when amino acids were not available there was a significant drop in energy efficiency. Interestingly, without supplemented amino acids (third case) each protein had a similar energy efficiency of about 28% regardless of carbon number. In this case, the energy burden of synthesizing each amino acid required for the assembly of the protein kept the energy efficiency saturated at a relatively low level. However, the experimentally constrained case of CAT production showed even a lower energy efficiency of $3.6 \pm 1.1\%$ compared to the theoretical maximum of $77.3 \pm 7.3\%$. This shows CFPS systems have a lot of room for improvement: first, the experimental measurements showed accumulation of certain amino acids; this carbon could potential could be diverted towards a protein of interest. Second, the system had a high accumulation of metabolic byproducts, specifically organic acids, which is a result of inefficient energy utilization.

2.2.3 Yield

The mean carbon yield was inversely proportional to protein size and varied between 40-57% for the proteins sampled (Fig. **??**). The same inverse qualitative trend was observed as for energy efficiency. There was a drop in carbon yield by about 7% once amino acids were not available; this is most likely because glucose was utilized to synthesize the necessary amino acids for each protein as well as power the system. For the first case, the system relied on a mixture of glucose and some amino acids for each protein with a carbon yield of 58.2 $\pm$ 2.3% for CAT. Once amino acid synthesis reactions were blocked in the network (second case), the carbon yield dropped to 57.1 $\pm$ 2.2% for CAT. Only the necessary amount of amino acids was used for the production of the protein of interest; thus, it may be hypothesized that all the glucose was used to power CFPS and did not contribute to the carbon yield. In that case, the carbon yield without glucose contribution would be 100%. In the experimentally constrained case, CAT was produced with a carbon yield of 6.2% compared to the theoretical maximum of 58.2%. This decrease in carbon yield suggests inefficiencies in CFPS that can potentially be improved. ssFBA assumes a psuedo steady state; thus, intermediate metabolites cannot accumulate within the cell-free extract. In addition, ssFBA is solved by maximizing the flux through the protein production reaction. Therefore, carbon flux will travel through the network to optimize the maximum flux through the protein synthesis reaction. In examining the experimental dataset, there is a high accumulation of organic acids, especially acetate, pyruvate and lactate. The experimental performance could be improved by diverting this carbon toward the protein of interest by knockouts during the cell-free extract preparation. A single trendline was formulated for the cases where amino acids were supplied in the media since they had similar performance, while another trendline was formulated for when amino acids were not available (Fig. **??**B). The trendlines showed good predictability for estimating carbon yield for a certain range of potein size (R^2 = 0.95 with amino acids and R^2 = 0.90 without amino acids), regardless of the type of promoters used. However, as the

protein size increased, the carbon yield decreased, suggesting that larger proteins may be less feasible for cell-free production. Thus, we examined the parameters that had the most significant effect on cell-free productivity, energy efficiency, and carbon yield in order to optimize CFPS performance.

2.3 Sensitivity analysis

The translation rate had the highest effect on protein productivity, whereas oxygen followed by substrate uptake had the highest effect on energy efficiency and carbon yield (Fig. 3). To better understand the effect of substrate utilization and the transcription/translation parameters on CFPS performance we performed global sensitivity analysis on the productivity and energy efficiency for CAT, a representative protein (Fig. 3), as well as on the carbon yield which followed the same trend of significance as energy efficiency (available in Supporting Information). In examining productivity performance (Fig. 3A), the significance of transcription/translation parameters was fairly constant across all three cases, with the rate of translation being the most significant. As expected, this showed that the translation rate was instrumental for productivity, and should be the first step investigated during optimization, prior to examining transcription parameters. Underwood and coworkers have also shown that an increase in ribosome levels did not significantly increase protein yields or rates; however, adding elongation factors increased yields by 23% at 30 minutes (*14*). In addition, Li et al. have increased productivity of firefly luciferase by 5-fold in CFPS systems by adjusting factors that affect transcription and translation such as elongation factors, ribosome recycling factor, release factors, chaperones, BSA, and tRNAs (*15*). In examining substrate utilization, glucose uptake was not seen to be very important for productivity in the cases with amino acid supplementation, but its significance increased when amino acids were not available. This makes sense, as amino acid synthesis from glucose became the only way to power protein synthesis. On the other hand, amino acid uptake only showed significance for the case where amino acids synthesis reactions

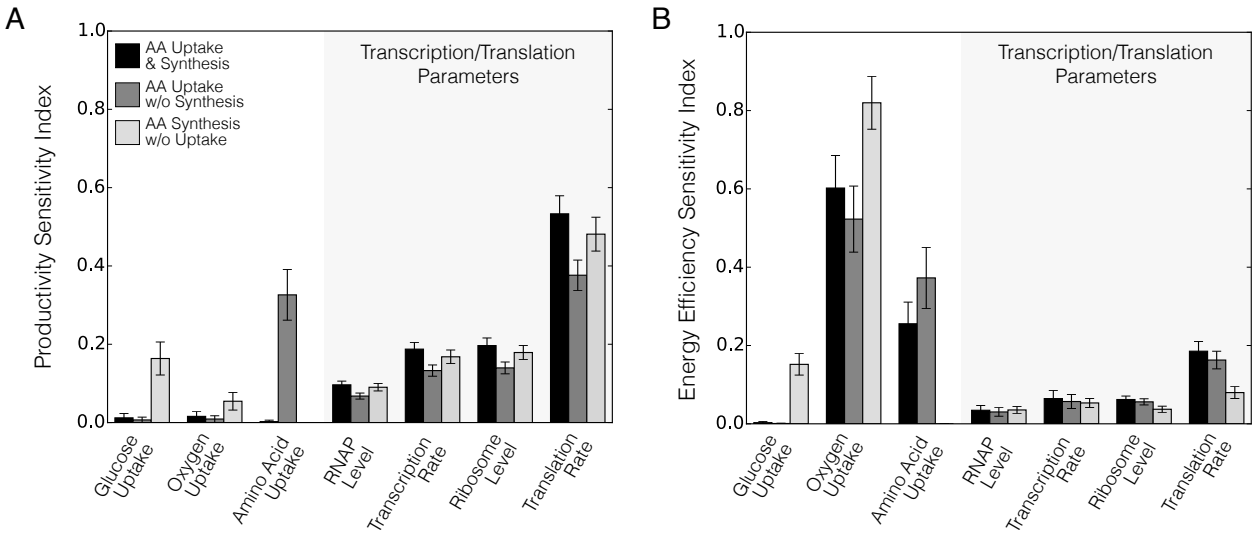

Figure 3: Total order sensitivity of deGFP productivity (A) and energy efficiency (B) to specific uptake rates and transcription/translation parameters for three cases: amino acid uptake and synthesis (black), amino acid uptake without synthesis (dark grey), and amino acid synthesis without uptake (light gray). Error bars represent a 95% CI on the sensitivity index.

were blocked, as it was the only source of amino acids for CAT synthesis.

When considering energy efficiency performance (Fig. 3B), the oxygen uptake rates were the most important for all cases while the sensitivity of the transcription/translation parameters decreased. The transcription/translation parameters had the same trend as for productivity, where the translation rate was the most sensitive compared to the other transcription/translation parameters and showed significance across all cases. Oxygen uptake was the most significant for energy efficiency, since it was responsible for oxidative phosphorylation, the most efficient pathway for energy generation. Across all cases, substrate utilization (amino acid uptake and glucose) was shown to be the next most important as these substrates contributed to the energy efficiency of CAT synthesis. In cases where amino acids were supplied, amino acid uptake was significant, as without it, energy was required to synthesize amino acids for the protein of interest. Glucose uptake was only important when amino acids were not available, since it was the only source of carbon for protein synthesis and energy generation. Jewett and coworkers have reported that oxidative phosphorylation still operated in cell-free systems, and that yield decreased from 1.5-fold to 4-fold when oxidative phosphorylation reactions were knocked

out in pyruvate-powered CFPS (*1*). It is unknown how active oxidative phosphorylation is compared to in *in vivo* systems. To investigate this further we compared CAT carbon yield to oxidative phosphorylation flux (Fig. 4). Interestingly, oxidative phosphorylation had a strong effect on the carbon yield in all cases. The cases where amino acids were supplied followed the same trend: ranging from a carbon yield of 20% to 58%, depending on the oxidative phosphorylation activity. Once amino acids were removed from the media, the carbon yield dropped to about 10%, and reached a maximum of 52%. When amino acids are not avialable in the media, a lower carbon yield was expected for the same oxidative phosphorylation flux, since carbon must be utilized for energy generation and amino acid biosynthesis. In all cases, whenever the carbon yield was below its theoretical maximum, there was an accumulation of acetate and lactate, resulting in the lower carbon yield. The experimental dataset exhibits a mixture of acetate and lactate accumulation during CAT synthesis, which shows that CFPS is not operating in a fully aerobic state. Glucose can not be fully oxidized and therefore fermentation pathways are used. Oxidative phosphorylation relies on electron transport from membrane vesicles, however CFPS has no cell membrane, thus it is expected CFPS has limited oxidative phosphorylation activity. The addition of phosphate showed an increase in CAT yield; however, it is unclear whether the addition of phosphate enhances oxidative phosphorylation, inhibits phosphatase reactions, or both (*1*). Interestingly, the addition of NADH did not increase the rate of protein synthesis, since the concentration of ATP was most likely saturated. An alternative strategy may be the inhibition of anaerobic processes in cell-free, in order to minimize unwanted byproducts such as acetate and lactate.

2.4 Flux distribution

To investigate the differences between the three cases, we compared the flux distributions for CAT production predicted by ssFBA simulations (Fig. 5). In the cases supplied with amino acids (Fig. 5A-B), glucose traveled to acetyl-coenzyme A that generated NADH for

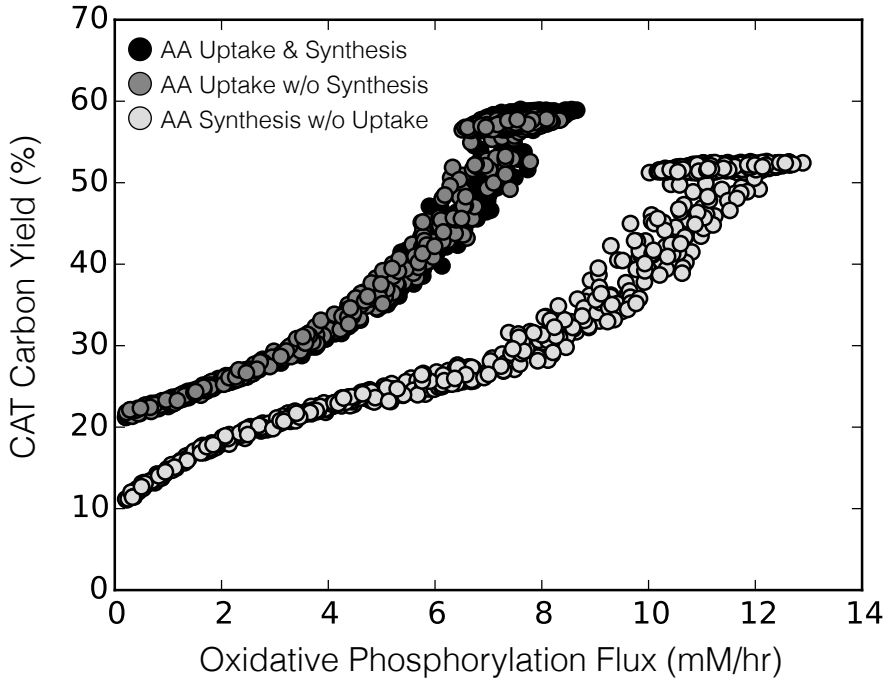

Figure 4: deGFP carbon yield versus oxidative phosphorylation flux, across an ensemble of 1000 ssFBA solutions, for three cases: amino acid uptake and synthesis (black), amino acid uptake without synthesis (dark grey), and amino acid synthesis without uptake (light gray).

oxidative phosphorylation to energize CFPS. Interestingly, in the first case with amino acid synthesis reactions, ssFBA relied on a combination of glucose and amino acids to power the system Amino acids rather than glucose powered the TCA cycle via glutamate to alpha-ketoglutarate which traveled to oxaloacetic acid and pyruvate for additional amino acid biosynthesis. In the case where amino acid synthesis reactions were blocked (Fig. 5B), ubiquinone was generated via *nuo* to power oxidative phosphorylation, instead of relying on the TCA cycle. Once the energy requirements for transcription and translation were met, amino acids were taken from the media to assemble the protein of interest. These first two cases where amino acids were available had similar performance with a correlation of 0.99 between their flux distributions and were comparable in terms of productivity, energy efficiency and carbon yield for all proteins. In the case where amino acids must be synthesized, there was a drop in the performance metrics of energy efficiency and carbon yield compared to the cases where amino acids are avialble. The flux distribution had a correlation of 0.90 when compared to both cases where amino acids were avaialbe. This

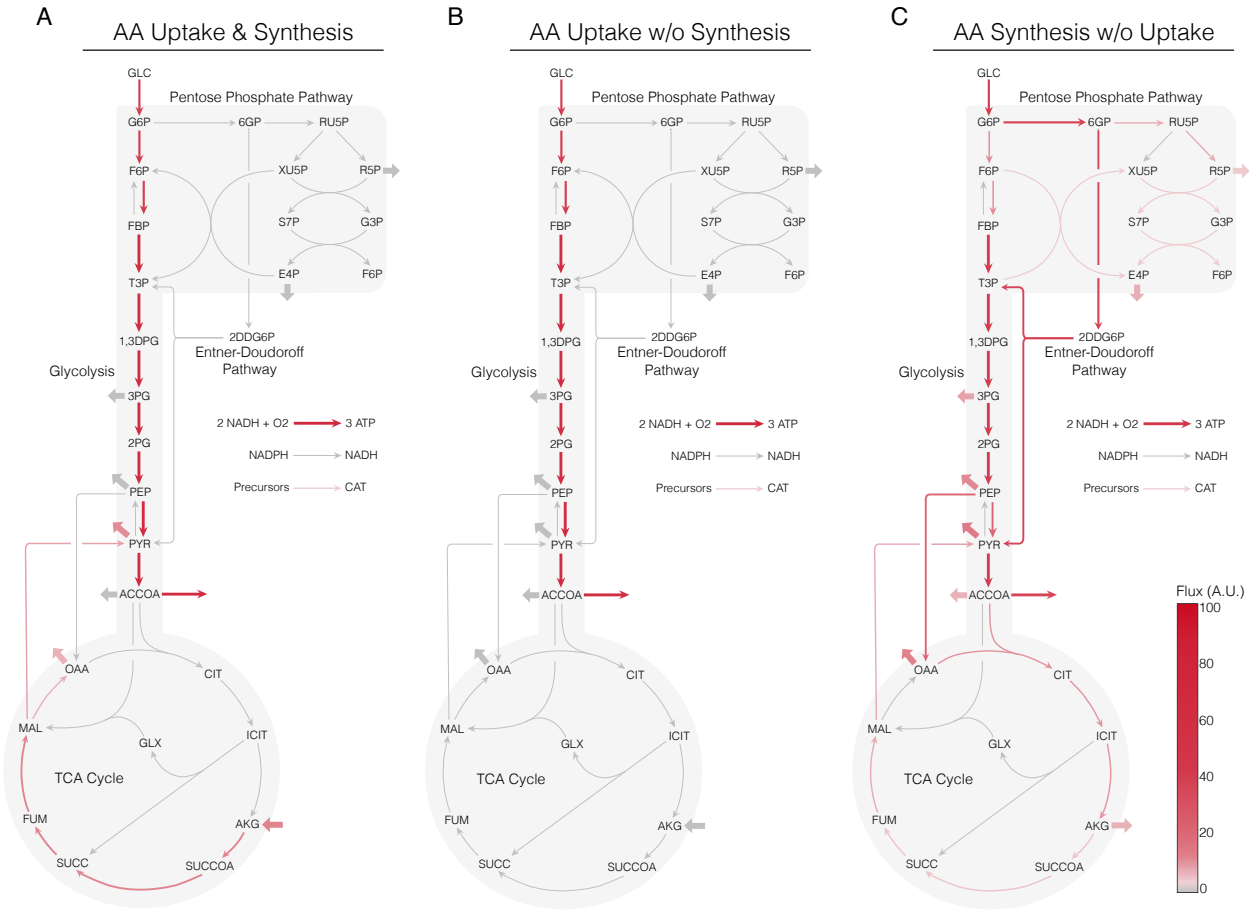

Figure 5: Flux profile for glycolysis, pentose phosphate pathway, Entner-Doudoroff pathway, TCA cycle, and oxidative phosphorylation, for three different cases: (A) amino acid uptake and synthesis, (B) amino acid uptake without synthesis, and (C) amino acid synthesis without uptake. Mean flux across the ensemble, normalized to glucose uptake flux. Thick arrows indicate flux to or from amino acids.

drop in performance is due to the burden of synthesizing amino acids, which require NADPH. This leads to the relatively high flux in the first step of the pentose phosphate pathway to generate NADPH, thus, less NADH is available for oxidative phosphorylation. The performance metrics and sensitivity analysis suggest that efficient energy generation via oxygen uptake is essential to higher energy efficiency and carbon yields.

The case constrained by experimental measurements (Fig. 6), had the highest correlation of 0.66 with the flux distribution from the case supplied with no amino acids and a correlation of 0.52 with the cases supplied with amino acids. Thus there are some differences in the flux distribution compared to the optimum solutions which may provide some insight to improve CFPS performance. Metabolic fluxes were constrained by

experimental measurements (available in Supporting Information) where available for the first hour which constrained the solution space of ssFBA to have a more realistic depiction of the flux distribution. The central carbon organic acids showed good agreement with the data (Fig. 6B). Only certain amino acid synthesis reactions were blocked since during the growth of *E. coli* not all amino acids were supplied (see Materials and Methods). During the cell-free reaction all amino acids were supplied, however glucose still traveled through all the major pathways, and the same metabolic precursors were still utilized for amino acid biosynthesis. In this case, it is unclear which substrate (glucose or amino acids) is used to power CFPS and may in fact be a combination of both. The optimum solutions only produced the required amount of amino acids necessary, however in examining the measurements there is an accumulation of alanine and glutamine which may explain some of the differences in the correlations of the flux distributions. Accumulation of pyruvate, lactate, acetate, and other organic acids can be seen (Fig. 6B) and shows an inefficiency of carbon utilization.

Despite the constraints by experimental measurements, it is difficult to calculate the physiological flux distribution of metabolism (Fig. 7) For example, there is a high flux through the Entner-Douodoroff pathway, but this is likely non-physiological, and simply an artifact of the optimal solution of ssFBA. A knockout of the Entner-Doudoroff pathway has no effect on the norm productivity of CAT compared to no knockouts (Fig. 7A). In addition, pairwise knockouts of Entner-Doudoroff and most subgroups in the network result in the same optimal solution of CAT productivity. However, there is a difference in the flux distribution with these knockouts (Fig. 7B), ssFBA will reroute the flux to optimize the objective function. Interestingly, a single group knockout of glycolysis/gluconoeogenesis, glutamate/glutamine biosynthesis, alanine/aspartate/asparagine biosynthesis was detrimental to CAT productivity. Flux balance analysis has been shown to have multiple alternative optimal solutions with flux variablitiy analysis and mixed-integer linear programming resulting in a poor depiction of the physiological flux distribution

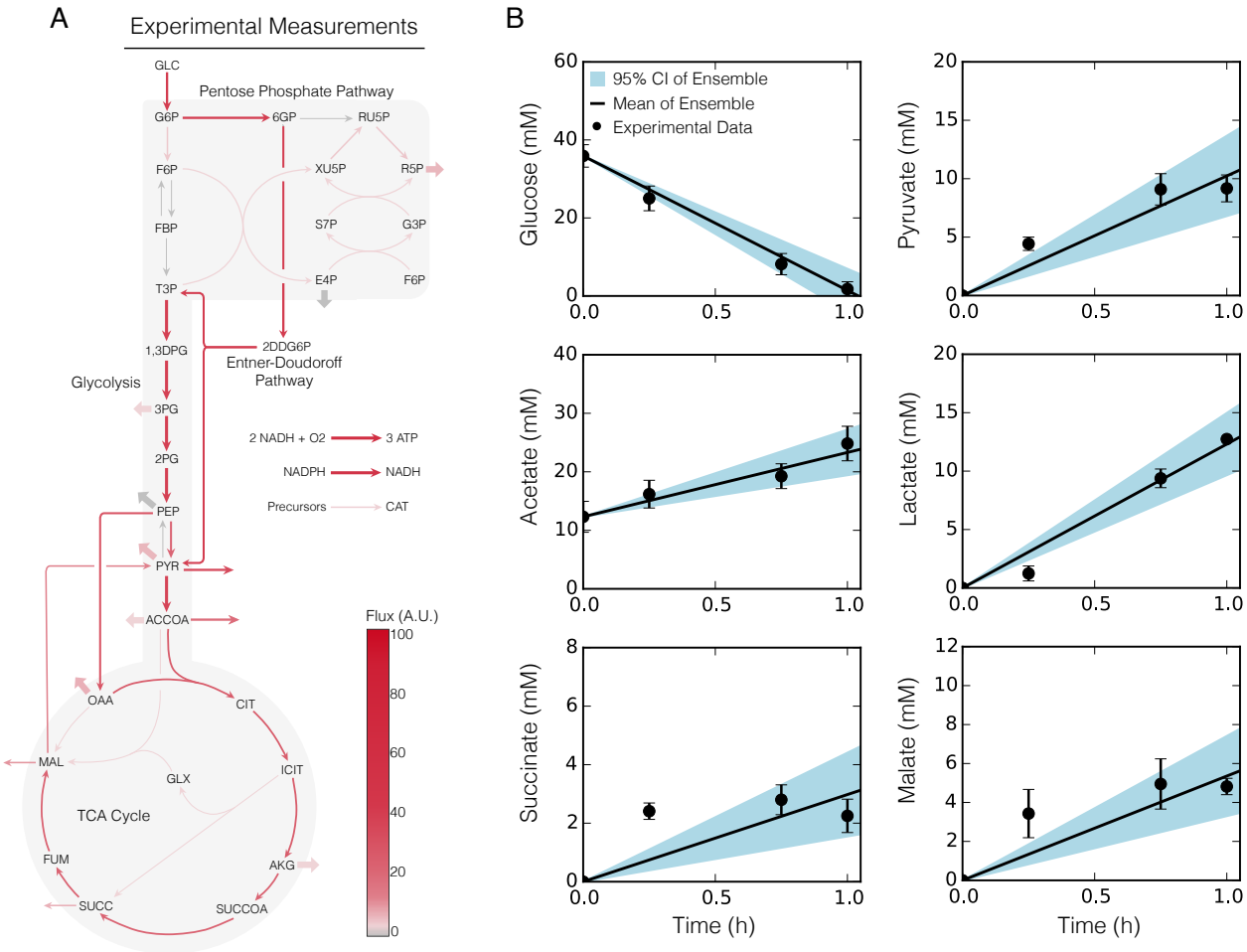

Figure 6: ssFBA simulation of CAT production for an experimentally constrained case. (A) Flux profile for glycolysis, pentose phosphate pathway, Entner-Doudoroff pathway, TCA cycle, and oxidative phosphorylation. Mean flux across the ensemble, normalized to glucose uptake flux. Thick arrows indicate flux to or from amino acids. (B) Central carbon metabolite measurements versus ssFBA simulations over a one hour time course.

(*16–18*). In our case, ssFBA reached the same optimal solution of CAT productivity for 73% of the pairwise group knockouts. To determine which reactions occur in CFPS, adding thermodynamic feasibility constraints to reactions may result in a better depiction of the flux distribution (*19*, *20*). It would also be interesting to track the carbon flux using C^{13} labeling in CFPS and constrain branch reactions in ssFBA to the resulting measurements, a method that has been shown to represent the flux distribution for *in vivo* processes well (*21*).

Taken together, we developed a sequence specific constraints based modeling approach

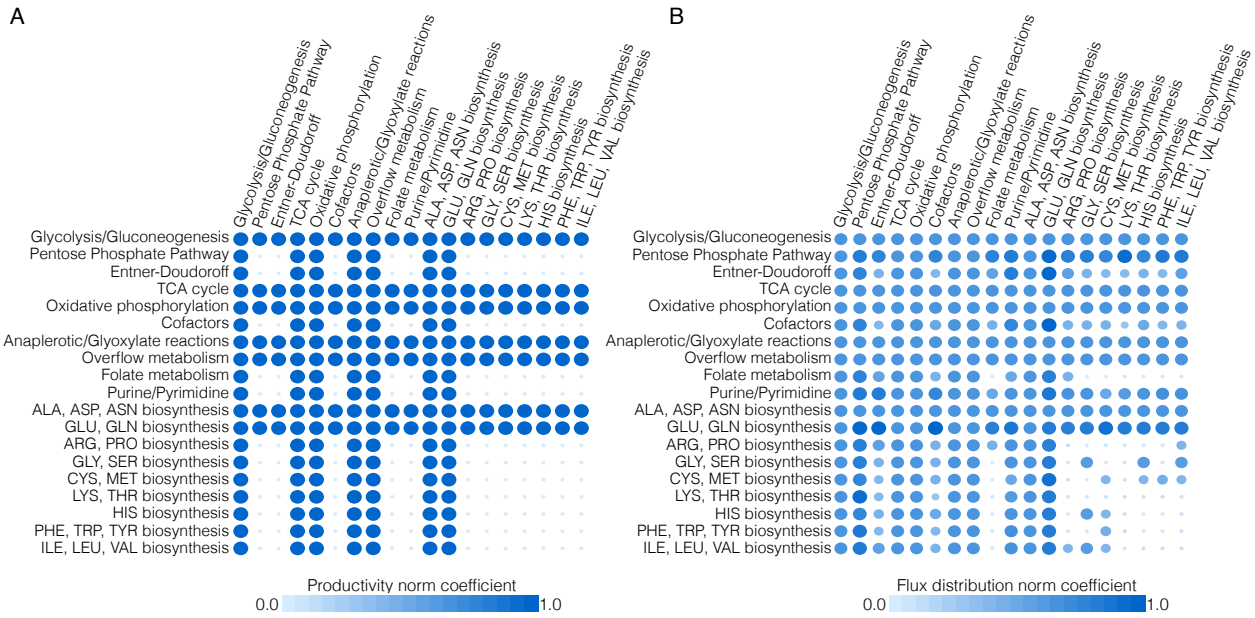

Figure 7: The norm effect of pairwise knockouts of subgroups in the cell-free network (A) CAT productivity norm compared to no knockouts in the experimentally constrained case. (B) Flux distribution norm compared to no knockouts in the experimentally constrained case.

to evaluate the performance of synthetic circuits in an *E. coli* CFPS system for a range of different proteins and three different cases. We have shown first principle predictions for protein production of deGFP and CAT in agreement with experimental measurements, under two different promoters and two different cell-free extract systems, with few adjustable parameters in the promoter models taken from literature. This modeling approach suggested trends for productivity, energy efficiency and carbon yield as a function of carbon number. Furthermore, global sensitivity analysis identified oxygen uptake as being instrumental for maintaining a high energy efficiency and carbon yield. The translation rate was identified as the rate limiting step for productivity. The model also suggested that cell-free systems can simultaneously operate aerobically and anaerobically, which can lead to inefficient production and should be addressed to optimize energy efficiency and carbon yield. In conclusion, sequence specific constraints based modeling offers a novel means to *a priori* estimate the performance of cell-free synthetic circuits.

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) (*22*). Crude S30 extract was prepared as described previously (*23*). 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 (*24*). 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 (*25*). 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

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 (*26*). 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 (*25*). 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 (*12*). 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 (*12*). 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.

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 (*11*):

Table 1: Transcription and translation template reactions for protein production.

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)	$R^*_X + \sum_{j \in \{AA\}} \alpha_j \cdot \left(AA_j\text{-}tRNA_j + 2GTP + 2H_2O \right)$	$\longrightarrow$	$\mathcal{P} + R_X + mRNA + \sum_{j \in \{AA\}} \alpha_j \cdot \left(tRNA + 2GDP + 2Pi \right)$

where $G_{\mathcal{P}}$ 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.

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 (*27*). For all cases, amino acid degradation reactions were blocked as these enzymes were knocked out during the cell-free extract preparation (*12*, *13*). In the second case, all amino acid synthesis reactions were set to 0 mM/hr since *E. coli* was grown in the presence of amino acids, thus these enzymes would not be present in the cell-free extract media. In the third case, amino acid uptake reactions were set to 0 mM/hr. In the experimental constrained case, *E. coli* was grown in the presence of 13 amino acids (alanine, arginine, cysteine, serine, aspartate, glutamate, and glutamine were excluded) (*22*), thus the synthesis reactions responsible for those 13 amino acid were set to 0 mM/hr.

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

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

where G denotes the gene concentration and K_T denotes a 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.* (*28*). 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.

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.

Table 2: Parameters for sequence specific flux balance analysis

Description	Parameter	Value	Units	Reference
RNA polymerase concentration	R_T	75	nM	(*13*)
Ribosome concentration	R_X	1.6	μM	(*13*, *14*)
Transcription elongation rate	$\dot{v}_{TX}$	25	nt/sec	(*13*)
Translation elongation rate	$\dot{v}_{TL}$	2	aa/sec	(*13*, *14*)
Transcription saturation coefficient	K_{TX}	3.5	nM	estimated
Translation saturation coefficient	K_{TL}	45.0	μM	estimated
Polysome amplification constant	K_P	10	constant	estimated
mRNA degradation rate	λ	5.2	hr^{-1}	(*13*)
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	(*13*)
σ_{70} dissociation constant	K_D	130	nM	estimated
σ_{70} hill coefficient	n	1	constant	estimated
Gene concentration	G	5	nM	(*13*)
Gene length of CAT	l_G	683	nt	(*24*)
Gene length of deGFP	l_G	660	nt	(*13*)
Protein length of CAT	l_P	229	aa	(*24*)
Protein length of deGFP	l_P	219	aa	(*13*)

Calculation of energy efficiency.

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

$$\mathcal{E} = \frac{\mathtt{q}_{POI} \cdot \left(2\left(\mathrm{ATP_{TX}} + \mathrm{CTP_{TX}} + \mathrm{GTP_{TX}} + \mathrm{UTP_{TX}}\right) + 2 \cdot \mathrm{ATP_{TL}} + \mathrm{GTP_{TL}}\right)}{\mathtt{q}_{GLC} \cdot \mathtt{ATP}_{GLC}} \tag{10}$$

where $\mathtt{q}_{POI} = w_{TX}$ denotes the production rate for the protein of interest, $\mathrm{ATP_{TX}}$, $\mathrm{CTP_{TX}}$, $\mathrm{GTP_{TX}}$, $\mathrm{UTP_{TX}}$ denote the stoichiometric coefficients of each energy species for the transcription of the protein of interest, $\mathrm{ATP_{TL}}$, $\mathrm{GTP_{TL}}$ denote the stoichiometric coefficients of ATP and GTP for the translation of the protein of interest, $\mathtt{q}_{GLC} = w_{GLC}$ denotes the glucose uptake rate, and $\mathrm{ATP_{GLC}}$ denotes the equivalent ATP number for 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):

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

where $\mathtt{q}_{POI}$ 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, $\mathtt{q}_{m_i}$ denotes the uptake flux of the i^{th} reactant, and C_{m_i} denotes the carbon number of the i^{th} reactant.

Quantification of uncertainty.

Experimental factors taken from literature, for example macromolecular concentrations or elongation rates, have uncertainty associated with their values. 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 N = 100 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 (*13*, *14*).

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 (*29*). 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.* (*30*) 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 (*31*).

Pairwise group knockouts.

Pairwise and single group knouckouts were simulated in ssFBA by setting the flux bounds for the all the reactions in a group to zero. We grouped reactions in the cell-free network into 19 subgroups (available in Supporting Information). We computed the norm of the productivity of CAT for each pairwise knockout compared to the productivity of CAT with no knockouts. We also computed the norm of the flux distribution for each pairwise knockout compared to the flux distribution with no knockouts.

Acknowledgement

Please use "The authors thank ..." rather than "The authors would like to thank ...".

The author thanks Mats Dahlgren for version one of achemso, and Donald Arseneau for the code taken from cite to move citations after punctuation. Many users have provided feedback on the class, which is reflected in all of the different demonstrations shown in this document.

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/`.

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. 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.

7. 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.

8. 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.

9. 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.

10. 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.

11. 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.

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

13. 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.

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

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

16. 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. *Computers & Chemical Engineering 24*, 711 – 716.

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

18. Schuetz, R., Kuepfer, L., and Sauer, U. (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Molecular Systems Biology 3*.

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

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

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

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

23. 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.

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

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

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

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

28. 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.

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

30. 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. *Computer Physics Communications 181*, 259–70.

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