Sequence Specific Modeling of 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

In this study, we used sequence specific constraints based modeling to evaluate the performance of synthetic circuits in an *E. coli* cell-free protein synthesis system. A core *E. coli* metabolic model, consisting of glycolysis, pentose phosphate pathway, amino acid biosynthesis and degradation and energy metabolism, was then augmented with sequence specific descriptions of genetic circuits which included mechanistic models of promoter function, transcription and translation. Thus, unlike other synthetic biology modeling efforts, sequence specific constraints based modeling explicitly couples the transcription and translation of circuit components with the availability of metabolic resources. Model parameters for transcription and translation were taken from literature, allowing a first principles prediction of circuit performance. We tested this approach by first simulating T7 induced chloramphenicol acetyltransferase production and σ_{70}-induced deGFP expression; we then expanded these studies for

a range of different proteins. First principles predictions of circuit performance were consistent with measurements for a variety of cases. Further, global sensitivity analysis identified the key metabolic processes that controlled circuit performance in terms of productivity, energy efficiency, and carbon yield. A sufficient energy supply with oxidative phosphorylation is instrumental for high energy efficiency and carbon yields; in addition, the translation rate could be optimized for higher productivity. Taken together, sequence specific constraints 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 and carbon yield for eight different proteins. From this, higher carbon number proteins typically had lower productivity rates 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 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 272 reactions and 145 species, is shown in Fig. 1A. Using this network, in combination with detailed promoter models, and literature values for cell free culture parameters (Table ZZ), 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 P70 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, and 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).

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 no 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 P70 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

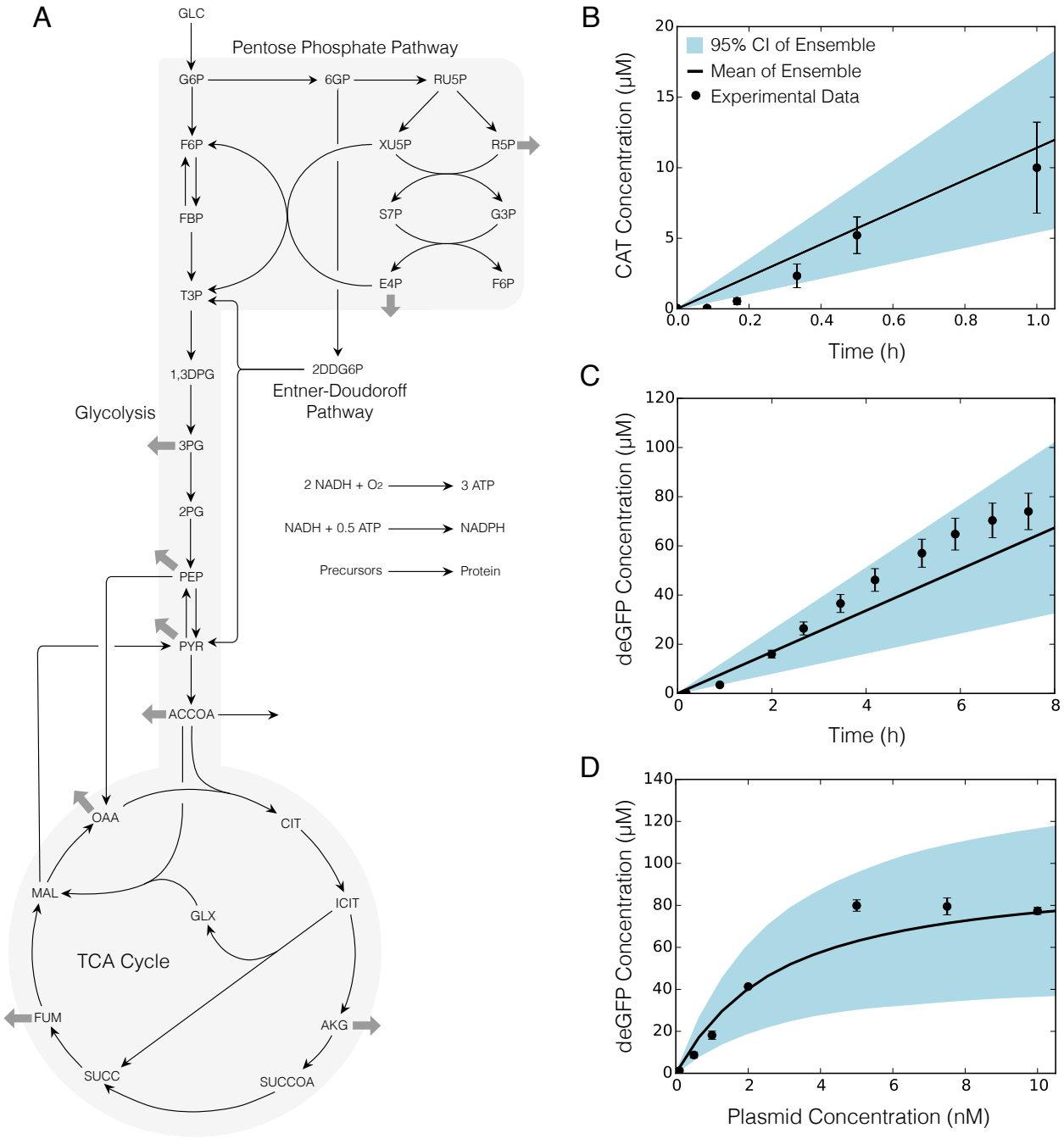

Figure 1: Sequence specific flux balance analysis. A. Core metabolic network with glycolysis, pentose phoshate pathway, TCA cycle, Entner-Doudoroff pathway. Thick gray arrows indicate withdrawal of precursors for amino acid synthesis. B. CAT production under a T7 promoter in CFPS *E. coli* extract for 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).

(FII), coagulation factor X (FX), fibroblast growth factor 21 (FGF21), and single chain variable fragment R4 (scFvR4). We used ssFBA to estimate the productivity for each of these proteins for each case (Fig. 2). An additional case was considered for CAT, since a comprehensive dataset is available (*12*); 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

We evaluated CFPS productivity for all eight proteins and for each case. 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 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 difference in the T7 and σ_{70} promoters did not alter the qualitative productivity performance of CAT. The higher productivity of CAT compared to all other proteins was most likely due to the lower transcription requirement of cytidine triphosphate which allowed a higher flux for translation. The drop in productivity in the

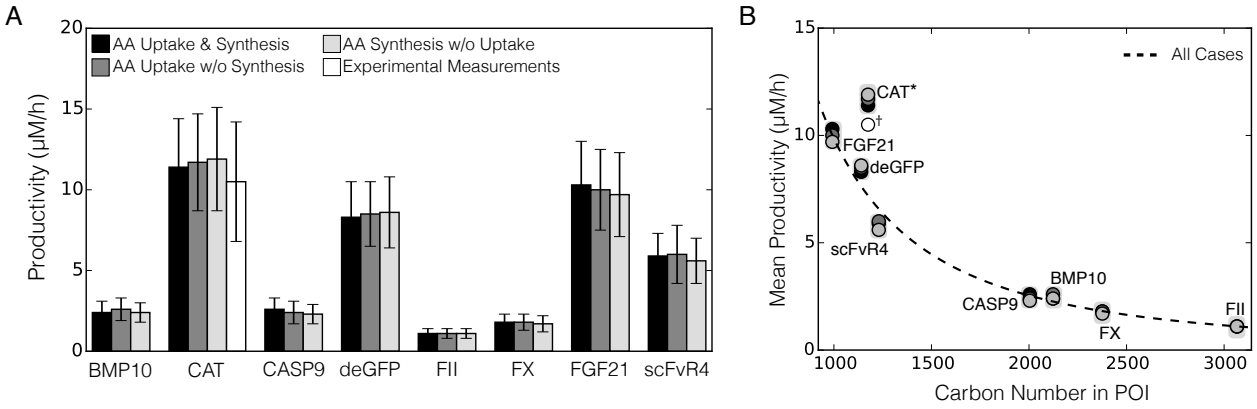

Figure 2: CFPS productivity of eight proteins for four cases. Amino acid uptake and synthesis (black), AA uptake without synthesis (dark grey), AA synthesis without uptake (light grey), and constrained by experimental measurements, for CAT only (white). A. Productivity across the ensemble (error bars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

third case is due to the fact that glucose is the only substrate supplied and must be used to provide the necessary energy requirements for transcription and translation, as well as synthesize each amino acid required for the production of the protein of interest.

2.2.2 Energy efficiency

Following the same outline as in examining the productivity, we calculated the energy efficiency of production for each protein (Fig. 3A). The first two cases, where amino acids were supplied in the media, had comparable performance including having the highest energy efficiencies. The third case (with no amino acid uptake) had the lowest energy efficiency; this was because glucose had to be utilized to synthesize the amino acids necessary for protein synthesis, in addition to being available for energy generation. We next investigated the effect of protein carbon number on energy efficiency (Fig. 3B). The same inverse trend was observed as for productivity, except that it was linear. The proteins with the lowest carbon number had the highest energy efficiency and the higher carbon number proteins had a lower energy efficiency for the first two cases. Proteins with a high carbon number have a higher transcription and translation cost then smaller proteins, leading to a lower energy efficiency of protein synthesis. The first two cases had the same

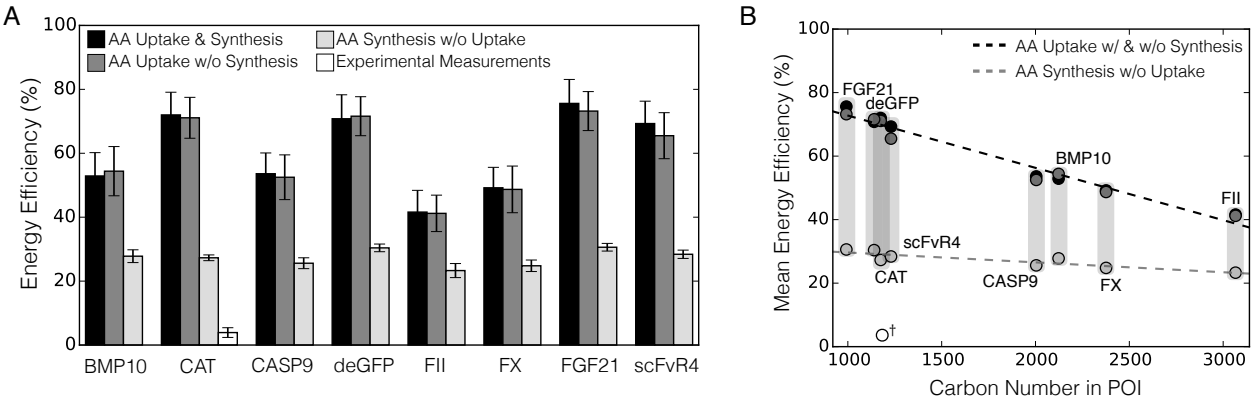

Figure 3: CFPS energy efficiency of eight proteins for four cases. Amino acid uptake and synthesis (black), AA uptake without synthesis (dark grey), AA synthesis without uptake (light grey), and constrained by experimental measurements, for CAT only (white). A. Energy efficiency (Error bars represent the 95% CI of the ensemble). B. Mean energy efficiency versus the carbon number for each corresponding protein. Trendline of energy efficiency versus carbon number (black dotted line) for first two cases (R^2 = 0.97) and trendline for AA synthesis without uptake (grey dotted line; R^2 = 0.79). Dagger: constrained by experimental measurements and excluded from trendline.

trendline whereas for the third case there was a significant drop in energy efficiency which required its own trendline (available in the Supporting Information). Interestingly, in the third case each protein had a similar energy efficiency of about 25% 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 3.9 $\pm$ 1.5% compared to the theoretical maximum of 72 $\pm$ 7.1%. This shows CFPS systems have a lot of room for improvement: first, the experimental setup still produced certain amino acids; these reactions could be turned off. 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

We also calculated the carbon yield for each protein (Fig. 4A). The same trends followed, where the cases with amino acids supplied in the media showed the highest carbon yield. The third case (with no amino acid uptake) had the lowest yields; this is most likely

because glucose is 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 56.9 $\pm$ 2.4% for CAT. Once amino acid synthesis was removed from the network (second case), each amino acid was utilized and the carbon yield increased to 57.0 $\pm$ 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%. Finally, for the third case (without amino acids supplemented), the carbon yield was reduced to 51.1 $\pm$ 1.8% for CAT, and the system used about twice the amount of glucose as in the first two cases. In this case, glucose was used to synthesize amino acids and provide the energy necessary to power transcription and translation; this trend was seen across all proteins. In the experimentally constrained case, CAT was produced with a carbon yield of 6% compared to the theoretical maximum of 57%. 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 setting protein production as the objective function. Therefore, carbon flux will travel through the network to optimize the flux through the protein synthesis reaction. In examining the experimental dataset, there is a high accumulation of organic acids, especially acetate. The experimental performance could be improved by diverting this carbon toward the protein of interest by knockouts during the cell-free extract preparation. Next we investigated the effect of the carbon number of each protein on the carbon yield (Fig. 4B). The same inverse qualitative trend was observed as for productivity. The proteins with the lowest carbon number had the highest yield and the higher carbon number proteins had a lower carbon yield within each case. A single trendline was formulated for the first two cases since they had similar performance, while another trendline was formulated for the third cases (available in Supporting Information).

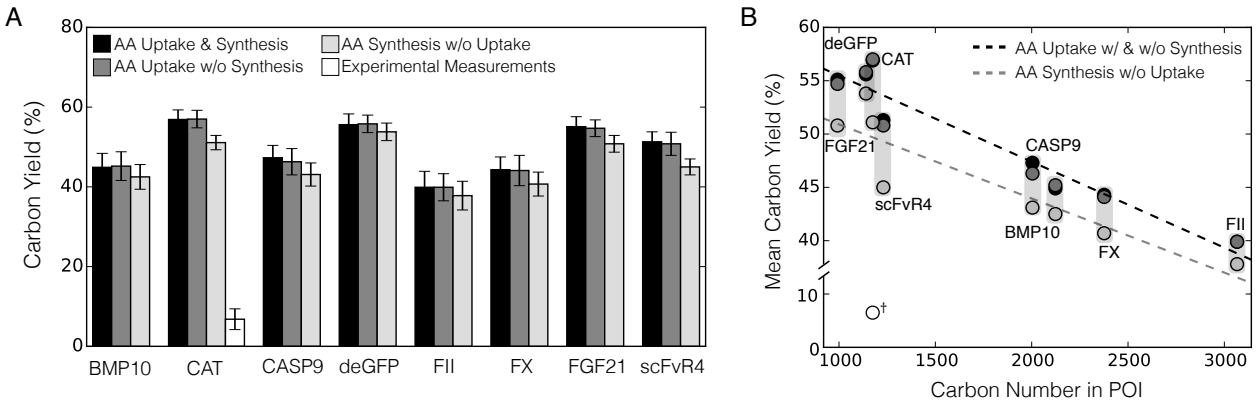

Figure 4: CFPS carbon yield of eight proteins for four cases. Amino acid uptake and synthesis (black), AA uptake without synthesis (dark grey), AA synthesis without uptake (light grey), and constrained by experimental measurements, for CAT only (white). A. Energy efficiency across the ensemble (error bars represent the 95% CI). B. Mean energy efficiency versus the carbon number for each corresponding protein. Trendline of energy efficiency versus carbon number (black dotted line) for first two cases (R^2 = 0.92) and trendline for AA synthesis without uptake (grey dotted line; R^2 = 0.83). Dagger: constrained by experimental measurements and excluded from trendline.

As the protein size increased, the carbon yield decreased, suggesting that large 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

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 deGFP, a representative protein (Fig. 5), as well as on the carbon yield (available in Supporting Information). In examining productivity performance (Fig. 5A), the significance of transcription/translation parameters was fairly constant across all three cases, with the rate of translation by ribosomes 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

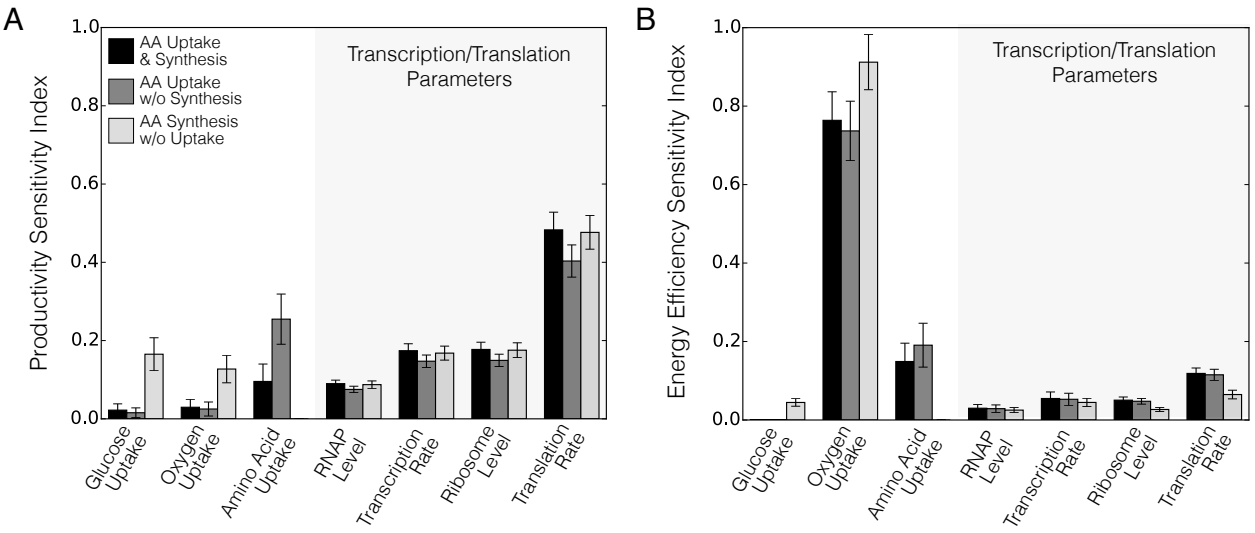

Figure 5: Total order sensitivity of deGFP productivity (A) and energy efficiency (B) to specific uptake rates and transcription/translation parameters for three cases: amino acid uptake and synthesis (black), amino acid uptake without synthesis (dark grey), and amino acid synthesis without uptake (light gray). Error bars represent a 95% CI on the sensitivity index.

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 first two cases, but its significance increased when amino acids were removed from CFPS. This makes sense, as amino acid synthesis from glucose became the only way to power protein synthesis in that case. In addition, oxygen uptake had more significance in the third case than in the first two cases, since in that case oxygen was required to synthesize amino acids and power the system. Also, oxygen determined how effectively glucose was utilized; therefore, if glucose uptake rate effects productivity, then oxygen could be expected to have a similar effect. On the other hand, amino acid uptake showed significance for the first two cases, and was even higher for the second case (without AA synthesis), as it was the only source of amino acids.

When considering energy efficiency performance (Fig. 5B), the oxygen uptake rates were the most important for all three cases while the sensitivity to transcription/translation

parameters decreased slightly. 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. Thus, in investigating CFPS performance in terms of productivity and energy efficiency, the translation rate was an important parameter to optimize, as has already been shown in literature (*14*, *15*). Oxygen uptake showed significant importance for energy efficiency, since it was responsible for oxidative phosphorylation, the most efficieny pathway for energy generation. Meanwhile, productivity was determined primarily by the rate of the most downstream steps, transcription and translation. Across all three cases, substrate utilization (amino acid uptake and glucose) was shown to be the next most important as these substrates contributed to the carbon yield of deGFP. In the first two cases where amino acids were supplied, amino acid uptake was significant, as without it energy was used to synthesize amino acids. Glucose uptake was only seen to be significant in the third case, 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*). We also investigated the sensitivity of carbon yield to network fluxes; it followed the same trends as the energy efficiency sensitivity analysis. It is unknown how active oxidative phosphorylation is compared to in *in vivo* systems. To investigate this further we compared deGFP carbon yield to oxidative phosphorylation flux (Fig. 6). Interestingly, oxidative phosphorylation has a strong effect on the carbon yield in all cases. The first two cases follow the same trend, ranging from a carbon yield of 20% to 55%, depending on the oxidative phosphorylation activity. The third case, followed the same trend; however, its carbon yield ranged from about 10% to 55%. The third case was expected to have a lower carbon yield for the same oxidative phosphorylation flux compared to the first two cases, since carbon must be utilized for energy generation and amino acid synthesis. In all three cases, whenever the

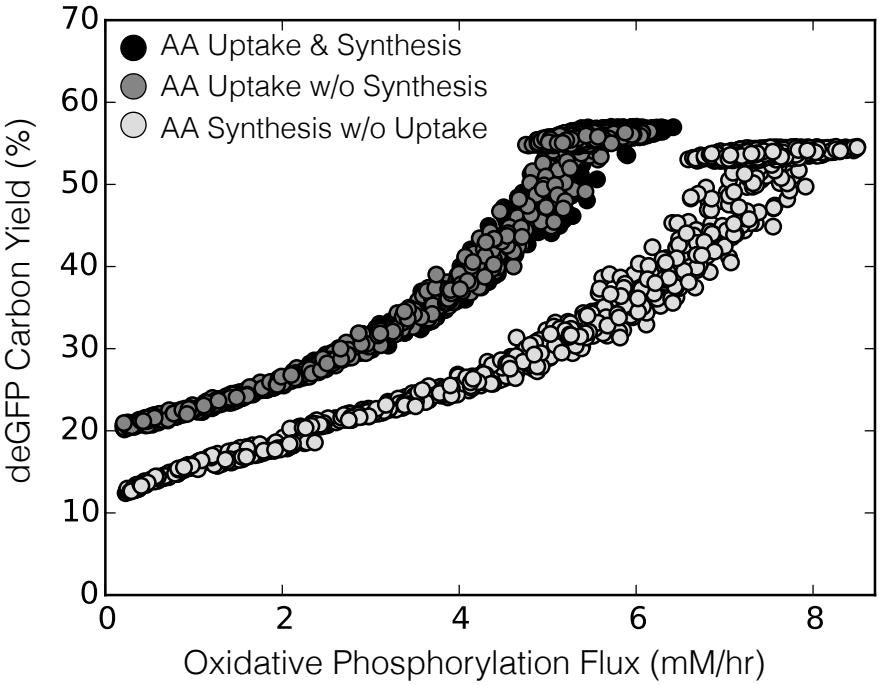

Figure 6: deGFP carbon yield versus oxidative phosphorylation flux, across an ensemble of 1000 ssFBA solutions, for three cases: amino acid uptake and synthesis (black), amino acid uptake without synthesis (dark grey), and amino acid synthesis without uptake (light gray).

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. It is unclear how active oxidative phosphorylation is in CFPS, since the reactions rely on electron transport from membrane vesicles. 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.

To investigate the differences between the three cases, we compared the flux distributions for a representative protein, deGFP, predicted by ssFBA simulations, as well as the case constrained by experimental measurements for CAT (Fig. 7). The first case, which was

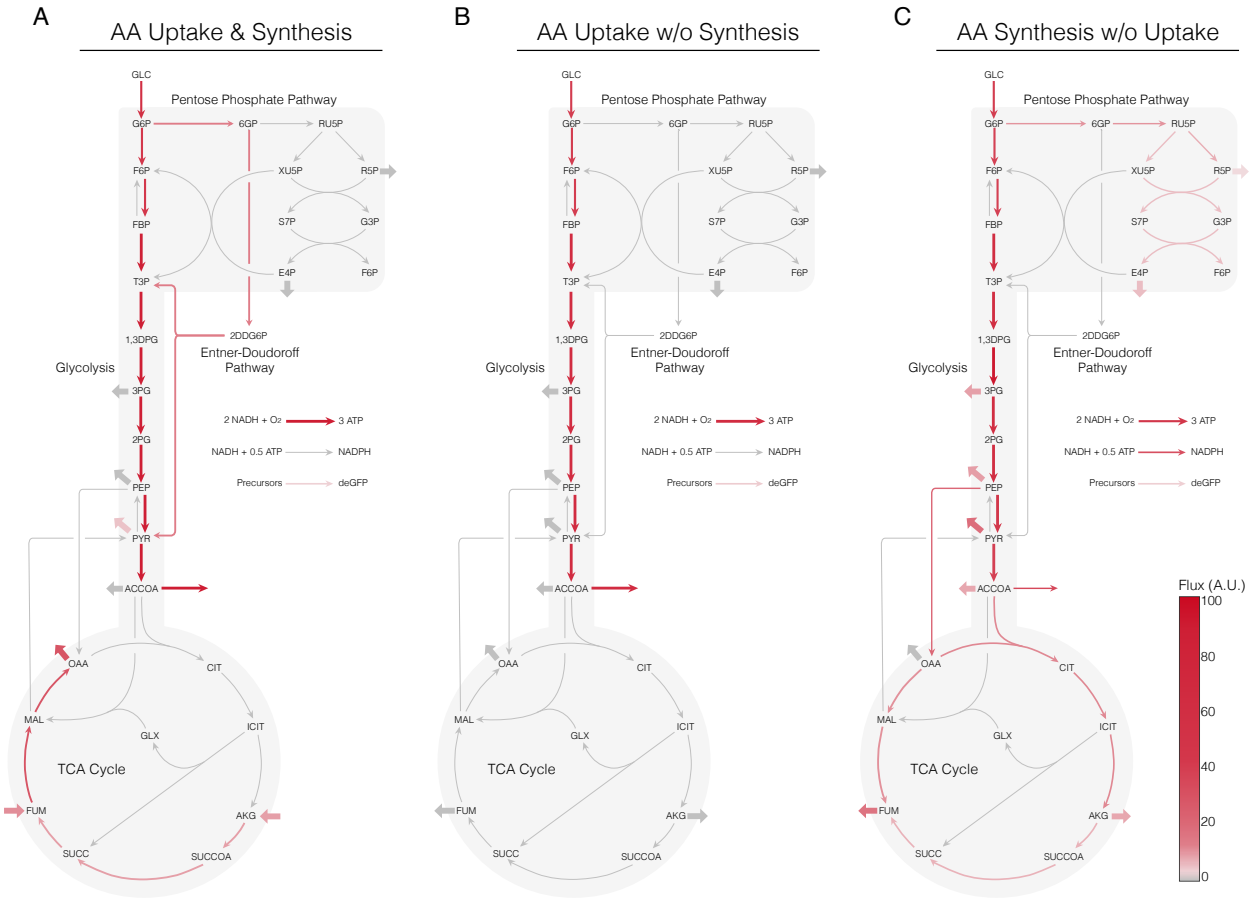

Figure 7: Flux profile for glycolysis, pentose phosphate pathway, Entner-Doudoroff pathway, TCA cycle, and oxidative phosphorylation, for three different cases: (A) amino acid uptake and synthesis, (B) amino acid uptake without synthesis, and (C) amino acid synthesis without uptake. Mean flux across the ensemble, normalized to glucose uptake flux. Thick arrows indicate flux to or from amino acids.

supplied with amino acids and could synthesize them in the network, relied on a combination of glucose and amino acids to power the system (Fig. 7A). Glucose traveled through glycolysis and the Entner-Doudoroff pathway to generate NADH for oxidative phosphorylation as well as utilize pyruvate for amino acid biosynthesis. Interestingly, amino acids rather than glucose powered the TCA cycle, via alpha-ketoglutarate and fumarate, and by utilizing oxaloacetic acid for additional amino acid biosynthesis. Thus, ssFBA found that the combined utilization of glucose and amino acids to energize CFPS was the optimal case for protein synthesis. In the second case, where amino acid synthesis was removed from the network, glucose was solely utilized to provide the necessary energy requirements via glycolysis which generated enough NADH for oxidative phosphorylation. 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. The first two cases had similar performance in terms of productivity, energy efficiency and carbon yield for all proteins. This is most likely due to the efficient utilization of carbon to energize CFPS without the burden of amino acid biosynthesis. In the third case, where amino acids must be synthesized, there is drop in all performance metrics (productivity, energy efficiency, and carbon yield) compared to the first two cases. This drop in performance is due to the burden of synthesizing amino acids, which require NADPH. This leads to the relatively high flux for the conversion of NADH to 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. Thus, removing anaerobic enzymes during the cell-free extract preparation could potentially improve CFPS performance and protein yield.

The fourth case, constrained by experimental measurements (Fig. 8), had a fairly similar flux distribution as the third case. The central carbon organic acids show good agreement with the data (Fig. 8B). 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. Unlike the second case, only certain amino acid synthesis reactions were blocked since during the growth of *E. coli* not all amino acids were supplied. 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. Accumulation of pyruvate, lactate, acetate, and other organic acids can be seen (Fig. 8B), implying an inefficiency of carbon utilization. In addition, 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. To determine which reactions occur

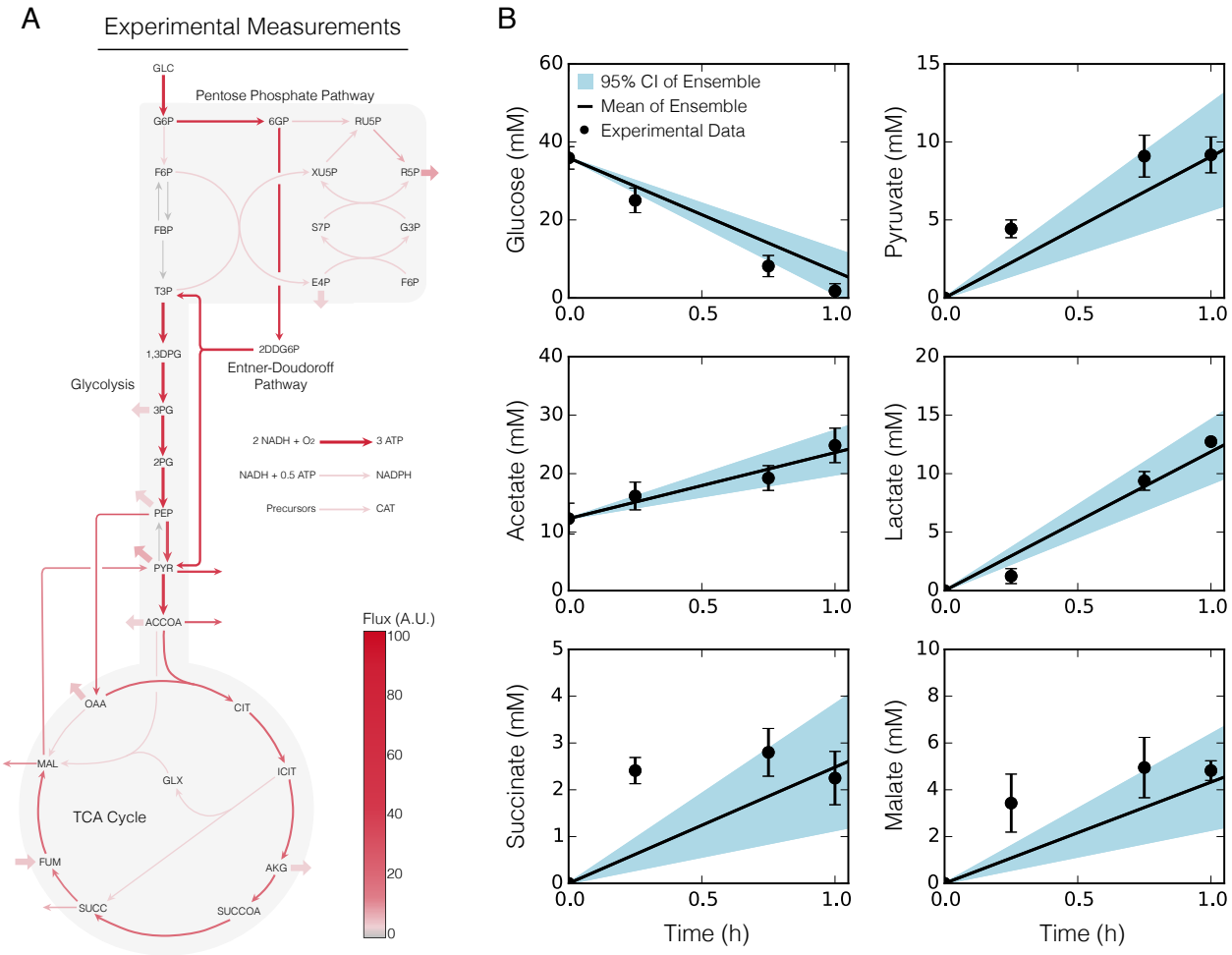

Figure 8: ssFBA simulation of CAT production for an experimentally constrained case. (A) Flux profile for glycolysis, pentose phosphate pathway, Entner-Doudoroff pathway, TCA cycle, and oxidative phosphorylation. Mean flux across the ensemble, normalized to glucose uptake flux. Thick arrows indicate flux to or from amino acids. (B) Central carbon metabolite measurements versus ssFBA simulations over a one hour time course.

in CFPS, adding thermodynamic feasibility constraints to reactions may result in a better depiction of the intracellular flux distribution (*16*, *17*). 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. Thus, it would 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 (*18*).

Taken together, we developed a sequence specific constraints based modeling approach 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 reactions were performed at a volume of 15 μL in 1.5-mL eppendorf tubes and incubated in a humidified incubator. 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 (*19*). The plasmid was isolated and purified using a Plasmid Maxi Kit (Qiagen, Valencia CA). Cell-free protein synthesis of CAT protein was performed at 37 °C for 3 h.

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 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) (*20*). Crude S30 extract was made as previously described (*21*).

The standard protein synthesis reaction was conducted according to the PANOxSP protocol with slight modifications from that described previously (*22*). Unless otherwise noted, all reagents were purchased from Sigma (St. Louis, MO). The initial mixture includes 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 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 mM glucose in place of PEP, and glutamate reactions do not have a secondary energy source added since glutamate is already present in high concentrations from the reaction salts. Sodium oxalate was not added since it has a detrimental effect on protein synthesis and ATP concentrations when using glucose or an early glycolytic intermediate as the energy source (*23*). 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 and glutamate reactions since a lower magnesium optimum is found when using a nonphosphorylated energy source (*22*). Finally, 10 mM phosphate was added to some reactions as described below in the form of potassium phosphate dibasic adjusted to pH 7.2 with acetic acid.

Measurements of protein product and metabolites.

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*). Calibration standards were routinely run for improved quantification of compounds. The standard compounds quantified with a refractive index detector included inorganic phosphate, glucose, and acetate. The compounds quantified with the UV detector included pyruvate, malate, succinate, and lactate. 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.

We estimated the theoretical maximum performance of the cell-free protein synthesis system using sequence specific flux balance analysis (ssFBA) (*11*). The sequence specific flux balance analysis problem was formulated as a linear program:

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

where $\mathbf{S}$ denotes the stoichiometric matrix, $\mathbf{w}$ denotes the unknown flux vector, $\boldsymbol{\theta}$ denotes the objective selection vector and α_i and β_i denote the lower and upper bounds on flux w_i, respectively. The stoichiometry of the kinetic model was used for the ssFBA calculations, with the execpetion of the transcription and translation rates. The transcription (TX) and translation (TL) stoichiometry was modeled using the template reactions taken from Allen

and Palsson (*11*):

$$
\begin{aligned}
\mathrm{G_{\mathcal{P}}} + \mathrm{R_1} &\longrightarrow \mathrm{G^*_{\mathcal{P}}} \\
\mathrm{G^*_{\mathcal{P}}} + \sum_{\mathrm{k}\in\{\mathrm{A,C,G,U}\}} \eta_k \cdot \{\mathrm{k}\}\,\mathrm{TP} &\xrightarrow{\mathrm{TX}} \mathrm{mRNA} + \mathrm{G_{\mathcal{P}}} + \mathrm{R_1} + \sum_{\mathrm{k}\in\{\mathrm{A,C,G,U}\}} 2\eta_{\mathrm{k}} \cdot \mathrm{Pi} \\
\mathrm{mRNA} &\longrightarrow \sum_{\mathrm{k}\in\{\mathrm{A,C,G,U}\}} \eta_{\mathrm{k}} \cdot \{\mathrm{k}\}\,\mathrm{MP} \\
\mathrm{mRNA} + \mathrm{R_2} &\longrightarrow \mathrm{R^*_2} \\
\mathrm{ff_j} \cdot \mathrm{AA_j} + \mathrm{ff_j} \cdot \mathrm{tRNA} + \mathrm{ff_j} \cdot \mathrm{ATP} &\longrightarrow \mathrm{ff_j} \cdot \mathrm{AA_j} - \mathrm{tRNA_j} + \\
&\qquad \mathrm{ff_j} \cdot \mathrm{AMP} + 2\mathrm{ff_j} \cdot \mathrm{Pi} \qquad j = 1, 2, \ldots, 20 \\
\mathrm{R^*_2} + \sum_{\mathrm{j}\in\{\mathrm{AA}\}} \mathrm{ff_j} \cdot \Big(\mathrm{AA_j} - \mathrm{tRNA_j} + 2 \cdot \mathrm{GTP}\Big) &\xrightarrow{\mathrm{TL}} \mathcal{P} + \mathrm{R_2} + \mathrm{mRNA} + \\
&\quad + \sum_{\mathrm{j}\in\{\mathrm{AA}\}} \mathrm{ff_j} \cdot \Big(\mathrm{tRNA} + 2 \cdot \mathrm{GDP} + 2 \cdot \mathrm{Pi}\Big)
\end{aligned}
$$

where $G_{\mathcal{P}}$ denotes the gene encoding protein product $\mathcal{P}$, $\mathrm{R_1}$ denotes the concentration of RNA polymerase, $G^*_{\mathcal{P}}$ denotes the gene bounded by the RNA polymerase, η_i and α_j denote the stoichiometric coefficients for nucleotide and amino acid, respectively, Pi denotes inorganic phosphate, $\mathrm{R_2}$ denotes the ribosome concentration, $\mathrm{R^*_2}$ denotes bounded ribosome, and AA_j denotes the j^{th} amino acid.

The transcription rate (w_{TX}) was fixed in the ssFBA calculation at:

$$w_{TX} = V^{max}_{TX}\left(\frac{G}{K_{TX} + G}\right) \tag{2}$$

where G denotes the gene concentration and K_{TX} denotes a transcription saturation coefficient. The maximum rate of transcription V^{max}_{TX} was formulated as:

$$V^{max}_{TX} \equiv \left[R_1\left(\frac{v_{TX}}{l_G}\right)\mathcal{P}\right] \tag{3}$$

The term R_1 denotes the RNA polymerase abundance, v_{TX} denotes the RNA polymerase

elongation rate (nt/h), l_G denotes the gene length in nucleotides, and the last term $\mathcal{P}$ describes a model of promoter activity. In this study, we considered two promoters: T7 and σ_{70}. The promoter function for the T7 promoter, $\mathcal{P}_{T7}$, was given by:

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

where K_{T7} denotes a T7 RNA polymerase binding constant (*24*). The σ_{70} binding promoter, used for all other proteins, was formulated as:

$$P_{\sigma_{70}} = \frac{K_1 + K_2 f_{p70}}{1 + K_1 + K_2 f_{p70}} \tag{5}$$

where K_1 denotes the state of RNA polymerase binding, K_2 is the state of σ_{70} binding along with RNA polymerase, and f_{p70} denotes the fraction of the σ_{70} transcription factor bound to the promoter (modeled as a Hill function).

The translation rate (w_{TL}) was bounded by:

$$0 \leq w_{TL} \leq V_{TL}^{max} \left(\frac{\text{mRNA}_{\text{SS}}}{K_{TL} + \text{mRNA}_{\text{SS}}} \right) \tag{6}$$

where mRNA_{SS} denotes the steady state mRNA abundance and K_{TL} denotes the translation saturation constant. The maximum translation rate V_{TL}^{max} was formulated as:

$$V_{TL}^{max} \equiv \left[K_P R_2 \left(\frac{v_{TL}}{l_P} \right) \right] \tag{7}$$

The term K_P denotes the polysome amplification constant, v_{TL} denotes the ribosome elongation rate (amino acids per hour), l_P denotes the number of amino acids in the protein of interest, and mRNA_{SS} denotes the steady-state mRNA concentration:

$$mRNA_{SS} \simeq \frac{w_{TX}}{\lambda} \tag{8}$$

where λ denotes the rate constant controlling the mRNA degradation rate.

The objective of the sequence specific flux balance calculation was to maximize the rate of protein translation, w_{TL}. 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 (*25*). For all cases, amino acid degradation reactions were blocked since 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) (*20*), thus the synthesis reactions responsible for those 13 amino acid were set to 0 mM/hr.

Calculation of energy efficiency.

Energy efficiency was calculated as the ratio of protein production to glucose consumption, both in terms of equivalent ATP molecules:

$$Efficiency = \frac{\Delta\mathtt{POI} \cdot (2\,(ATP_{TX} + \mathrm{CTP_{TX}} + \mathrm{GTP_{TX}} + \mathrm{UTP_{TX}}) + 2 \cdot \mathrm{ATP_{TL}} + \mathrm{GTP_{TL}})}{\Delta\mathtt{GLC} \cdot \mathtt{ATP}_{GLC}} \quad (9)$$

where $\Delta\mathtt{POI}$ denotes the flux of the protein of interest produced, $\mathrm{ATP_{TX}}$, $\mathrm{CTP_{TX}}$, $\mathrm{GTP_{TX}}$, $\mathrm{UTP_{TX}}$ 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, $\Delta\mathtt{GLC}$ denotes the glucose flux, 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):

$$Y_C^{POI} = \frac{\Delta\texttt{POI} \cdot C_{POI}}{\sum_{i=1}^{\mathcal{R}} \max(\Delta m_i, 0) \cdot C_{m_i}} \tag{10}$$

where $\Delta\texttt{POI}$ 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, Δm_i denotes the amount of the i^{th} reactant consumed (not allowed to be negative), and C_{m_i} denotes the carbon number of the i^{th} reactant.

Quantification of uncertainty.

An ensemble of 100 sets of flux distributions was calculated for each of the three different cases: control (with amino acid synthesis and uptake), amino acid uptake without synthesis, and amino acid synthesis without uptake. The ensemble was calculated by randomly sampling the maximum specific glucose uptake rate from 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 7 and 16 µM, the RNA polymerase elongation rate between 20 and 30 nt/sec, and the ribosome elongation rate between 1.5 and 3 aa/sec (*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 synthetic circuits (*26*). 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.* (*27*) 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 (*28*).

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 and transcription/translation stoichiometric coefficients of energy species.

- Carbon Yield Sensitivity Analysis: Global sensitivity analysis on deGFP carbon yield.

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. Henry CS, H. V., Broadbelt LJ Thermodynamics-Based Metabolic Flux Analysis. *Biophysical Journal 92*, 1792–1805.

17. Hamilton, J., Dwivedi, V., and Reed, J. Quantitative Assessment of Thermodynamic Constraints on the Solution Space of Genome-Scale Metabolic Models. *Biophysical Journal 105*, 512–22.

18. Zamboni, N., Fendt, S., and Sauer, U. 13C-based metabolic flux analysis. *Nature Protocols 4*, 878–92.

19. 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.

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

21. 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.

22. 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.

23. 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.

24. Moon TS, T. A. S. B. V. C., Lou C (2012) Genetic programs constructed from layered logic gates in single cells. *Nature 491*.

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

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

27. 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.

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