Toward a Genome Scale Dynamic Model of Cell-Free Protein Synthesis in *Escherichia coli*

Nicholas Horvath, Michael Vilkhovoy, Joseph Wayman, Kara Calhoun[1], James Swartz[1] and Jeffrey D. Varner*

Robert Frederick Smith School of Chemical and Biomolecular Engineering

Cornell University, Ithaca NY 14853

[1]School of Chemical Engineering

Stanford University, Stanford, CA 94305

Running Title: Dynamic modeling of cell-free protein synthesis

To be submitted: *Scientific Reports*

*Corresponding author:

Jeffrey D. Varner,

Professor, Robert Frederick Smith School of Chemical and Biomolecular Engineering,

244 Olin Hall, Cornell University, Ithaca NY, 14853

Email: jdv27@cornell.edu

Phone: (607) 255 - 4258

Fax: (607) 255 - 9166

Abstract

Cell-free protein expression systems have become widely used in systems and synthetic biology. In this study, we developed an ensemble of dynamic *E. coli* cell-free protein synthesis (CFPS) models. Model parameters were estimated from measurements of glucose, organic acids, energy species, amino acids, and the protein product, chloramphenicol acetyltransferase (CAT). The ensemble described all of the training data, especially the central carbon metabolism. The model predicted a carbon yield for CAT production that was equal to 36% of yield for a physiologically realistic case, and an energy efficiency equal to 9% of the physiologically realistic case, calculated using sequence-specific flux balance analysis. This suggests that CAT production could be further optimized. The dynamic modeling approach predicted that substrate consumption and oxidative phosphorylation were most important to both CAT production and the system as a whole, while CAT production alone depended heavily on the CAT synthesis reaction. Conversely, CAT production was robust to allosteric control, as was most of the network, with the exception of the organic acids in central carbon metabolism. This study is the first to model dynamic protein production in *E. coli*, and should provide a foundation for genome-scale, dynamic modeling of cell-free *E. coli* protein synthesis.

Keywords: Biochemical engineering, systems biology, cell-free protein synthesis

Introduction

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 fundamentally important 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, 6]. 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.

Mathematical modeling has long contributed to our understanding of metabolism. Decades before the genomics revolution, mechanistically structured metabolic models arose from the desire to predict microbial phenotypes resulting from changes in intracellular or extracellular states [7]. The single cell *E. coli* models of Shuler and coworkers pioneered the construction of large-scale, dynamic metabolic models that incorporated multiple, regulated catabolic and anabolic pathways constrained by experimentally determined kinetic parameters [8]. Shuler and coworkers generated many single cell kinetic models, including single cell models of eukaryotes [9, 10], minimal cell architectures [11], as well as DNA sequence based whole-cell models of *E. coli* [12]. In the post genomics world, large-scale stoichiometric reconstructions of microbial metabolism popularized by techniques such as flux balance analysis (FBA) have become a standard approach [13]. Since the first genome-scale stoichiometric model of *E. coli*, developed by Edwards and

Palsson [14], well over 100 organisms, including industrially important prokaryotes are now available [15–17]. Stoichiometric models rely on a pseudo-steady-state assumption to reduce unidentifiable genome-scale kinetic models to an underdetermined linear algebraic system, which can be solved efficiently even for large systems. Traditionally, stoichiometric models have also neglected explicit descriptions of metabolic regulation and control mechanisms, instead opting to describe the choice of pathways by prescribing an objective function on metabolism. Interestingly, similar to early cybernetic models, the most common metabolic objective function has been the optimization of biomass formation [18], although other metabolic objectives have also been estimated [19]. Recent advances in constraint-based modeling have overcome the early shortcomings of the platform, including capturing metabolic regulation and control [20]. Thus, modern constraint-based approaches have proven extremely useful in the discovery of metabolic engineering strategies and represent the state of the art in metabolic modeling [21, 22]. However, genome-scale kinetic models of industrial important organisms such as *E. coli* have yet to be constructed.

In this study, we developed an ensemble of kinetic cell-free protein synthesis (CFPS) models using dynamic metabolite measurements in an *E. coli* cell free extract. Model parameters were estimated from measurements of glucose, organic acids, energy species, amino acids, and the protein product, chloramphenicol acetyltransferase (CAT). Characteristic values for model parameters and initial conditions, estimated from literature, were used to constrain the parameter estimation problem. The ensemble of parameter sets described the training data with a median cost that was greater than two orders of magnitude smaller than random sets constructed using the literature parameter constraints. We then used the ensemble of kinetic models to analyze the CFPS reaction. First, sensitivity analysis of the dynamic model suggested that CAT production was most sensitive to CAT synthesis parameters, as well as reactions in oxidative phosphorylation and pyruvate consumption. Sensitivity analysis also showed that the system as a whole was most sensitive to these same parts of the network and glucose consumption. CAT production and other metabolites, specifically organic acid intermediates such as pyruvate, were sensitive to the presence of allosteric control mechanisms. Next, to gauge the performance of the cell-free reaction, we compared the observed CAT carbon yield with the maximum theoretical CAT carbon yield calculated using sequence-specific flux balance analysis. The CAT yield estimated from the kinetic model was 36% of the theoretical yield when physiologically realistic constraints were used. Taken together, we have integrated traditional kinetics with a logical rule-based description of allosteric control to simulate a comprehensive CFPS dataset. This study provides a foundation for genome-scale, dynamic modeling of cell-free *E. coli* protein synthesis.

Results

The ensemble of kinetic CFPS models captured the time evolution of CAT biosynthesis (Fig. 1 - 3). The cell-free *E. coli* metabolic network was constructed by removing growth associated reactions from the MG1655 reconstruction [16], and by adding reactions describing chloramphenicol acetyltransferase (CAT) biosynthesis, a model protein for which there exists a comprehensive training dataset [23]. In addition, reactions that were knocked out from the cell extract preparation were removed from the network (ΔspeA, ΔtnaA, ΔsdaA, ΔsdaB, ΔgshA, ΔtonA, ΔendA). The CFPS model equations were formulated using the hybrid cell-free modeling framework of Wayman et al. [24]. An ensemble of model parameters (N $>$ 10,000) was estimated from measurements of glucose, CAT, organic acids (pyruvate, lactate, acetate, succinate, malate), energy species (A(x)P, G(x)P, C(x)P, U(x)P), and 18 of the 20 proteinogenic amino acids using a constrained Markov Chain Monte Carlo (MCMC) approach. The MCMC algorithm minimized the error between the training data and model simulations starting from an initial parameter set assembled from literature and inspection. Parameter sets were selected for the ensemble based upon their error, and the Pearson correlation coefficient between the candidate and existing sets in the ensemble. The parameter set with the lowest error value was defined as the best-fit set. Central carbon metabolism (Fig. 1, top), energy species (Fig. 2), and amino acids (Fig. 3) were captured by the ensemble and the best-fit set. The constrained MCMC approach estimated parameter sets with a median error greater than two-order of magnitude less than random parameter sets generated within the same parameter bounds (Fig. 4); thus, we have confidence in the predictive capability of the estimated parameters. The model captured the biphasic CAT production: during the first hour glucose powers production, and CAT is produced at ~10 μM/h; subsequently, pyruvate and lactate reserves are consumed to power metabolism, and CAT is produced less efficiently at ~5 μM/h. Allosteric control was important to biphasic CAT production;

without control, the CAT production rate increased and then ceased after 1.5 hr (Fig. 1, bottom). In addition, acetate no longer accumulated after 1.5 hours, in the absence of allosteric control. Interestingly, the simulated malate abundance tracked the experimental measurements during the glucose consumption phase, but increased sharply during the pyruvate consumption phase, without allosteric control. Taken together, we produced an ensemble of kinetic models that was consistent with time series measurements of the production of a model protein. However, while the ensemble described the experimental data, it was unclear which kinetic parameters most influenced CAT production, and whether the performance of the CFPS reaction was optimal.

To better understand which parameters and parameter combinations influenced the performance of the kinetic model, we performed sensitivity analysis (Fig. 5). We perturbed each V^{max} parameter, either individually or in pairwise combinations and measured the change in either CAT production or the overall system state. CAT production was most sensitive to the CAT synthesis reaction, oxidative phosphorylation, and the pyruvate-consuming alanine synthesis reaction (Fig. 5, top, section A). We saw a common theme of the most important reactions producing or consuming the cofactors ATP, NADH, NADPH, and coenzyme A, as well as the metabolites pyruvate and glutamate. Of the 25 reactions to which CAT production was most sensitive, 9 produced or consumed ATP, making it the most represented in these top reactions (with the exception of hydrogen and phosphate ion). The next most represented were pyruvate, glutamate, and ADP with 7 reactions each, followed by coenzyme A, α-ketoglutarate, NAD/NADH, and NADP/NADPH, with 6 reactions each. This makes sense, as glutamate was an important precursor for the synthesis of other amino acids required by CAT production. Meanwhile, the cofactors provided energy to power CAT synthesis, while pyruvate was important for energy generation following glucose depletion. In addition, pyruvate was required for the synthesis of several amino acids. The pairwise sensitivities (off-diagonal elements) were different from the corresponding first-order sensitivities (diagonal elements), and led to interesting outcomes. The combination of certain reactions had a much greater or lesser effect on CAT production than that of the individual reactions by themselves. For example, glutamine synthesis and arginine degradation were both among the most important reactions to CAT production (they ranked 5th and 10th, respectively). This was likely because they both affected the sensitive glutamine-glutamate balance; glutamine synthesis consumes glutamate, while arginine degradation produces it. However, when both were perturbed, their combined effect on the model was low, as the respective contributions to consumption and production of glutamate cancelled. An example of positive synergy can be seen in cometabolite interconversion. Pyridine nucleotide transhydrogenase catalyzes two reactions: one converts NAD and NADPH into NADH and NADP, while the other does the reverse and also generates a proton gradient. Increasing one or the other has little effect on the model, as any increase in reaction flux is hampered by the imbalance created. Increasing both, however, allows all cometabolites to balance, so that the reaction fluxes increase unhindered. Thus, the pairwise sensitivity is much higher than the sum of the first-order sensitivities.

The overall system state was also sensitive to cofactors and substrates; however, instead of pyruvate and glutamate, the substrates driving metabolism were pyruvate and G3P. The system was most sensitive to an oxidative phosphorylation reaction (cytochrome oxidase), which converted ubiquinol to ubiquinone while generating a proton gradient. The next 4 most important reactions were all consumers of pyruvate: lactate dehydrogenase, pyruvate formate lyase, alanine synthesis, and PEP synthase. Of the 25 reactions to which the system was most sensitive, 7 produced or consumed pyruvate and 7 participated in NAD/NADH exchange. With the exception of hydrogen, these were the most represented species in the top 25 reactions. Also important were coenzyme A (5 reactions) and acetyl coenzyme A (4 reactions), as well as ubiquinone/ubiquinol, NADP/NADPH, G3P, and ATP (4 reactions each). The system state also had pairwise sensitivities that differed from the corresponding first-order sensitivities and stood out as significant. For example, alanine degradation was among the most important reactions, as it produced pyruvate; GMP synthesis was also moderately important, as it produced glutamate. However, when both reactions were increased, the combined effect on the model was almost zero. This can be understood by considering the reactions that involve both pyruvate and glutamate: they all either produced both of these substrates or consumed both. When alanine degradation was increased, the excess pyruvate stimulated these reactions to consume both pyruvate and glutamate; the amounts of both of these substrates could not be conserved. But when GMP synthesis was perturbed as well, the glutamate deficiency was corrected, cancelling much of the effect on the system. One of the pyruvate- and glutamate-consuming reactions was alanine synthesis; in this case, the levels of pyruvate, glutamate, and alanine were all virtually unchanged. An example of positive synergy can be seen in histidine synthesis, one of the least influential reactions that consumes three units of ATP. When perturbed in combination with the reverse reaction of succinyl coenzyme A synthetase, another reaction with little overall effect on the model, the combined effect on the system is much greater. This may be because the reverse reaction of succinyl coenzyme A synthetase produced ATP, which further stimulated histidine synthesis. Taken together, sensitivity analysis identified blocks of parameters that either individually, or in combination influenced model performance.

To understand whether the CFPS performance was optimal, we calculated the carbon yield and energy efficiency of CAT production (Fig. **??**). The best-fit parameter set for the kinetic model predicted a CAT carbon yield of 7.9%, while the experimental dataset had a CAT carbon yield of 8.2%. This was calculated as the increase in CAT concentration times the CAT carbon number, divided by the sum of the consumption terms for glucose and all amino acids excpet arginine and glutamate, as no data were available for these, weighted by their respective carbon numbers. To explain where the remainder of carbon was going, we performed a carbon balance for the best-fit set (Fig. 8). Of the other 92% of carbon, 35% accumulated as organic acids (lactate, acetate, succinate and malate) and 9% accumulated as amino acids (alanine and glutamine). The remaining 48% went to the net accumulation of all other metabolites, particularly carbon dioxide. The best-fit set produced CAT with an energy efficiency of 6.4%. This was calculated as the increase in CAT concentration times the CAT number of equivalent ATP molecules, divided by glucose consumption times times the number of equivalent ATP molecules for glucose, equal to 15 in the optimal case. An additional 62% of energy went to the accumulation of glycolysis metabolites, 6.4% to TCA cycle metabolites, 0.2% to pentose phosphate metabolites, and 25% to acetate and lactate. This shows that there is much room for improvement of the efficiency of CFPS. Gene knockouts in the electron transport chain further reduced the performance of the CFPS reaction (Fig. 6). A key finding of both the CAT and overall system state sensitivity analysis was the importance of oxidative phosphorylation. To investigate this further, we knocked out key oxidative phosphorylation reactions in the ensemble of kinetic models to examine the effect on CAT production and carbon yield. A single *cyd* knockout reduced the CAT carbon yield from 7.9% to 2.6% (Table 1). On the other hand, a *nuo* knockout showed a less dramatic decrease in yield, reducing the CAT carbon yield to 6.9%. Knocking out *app* increased CAT yield to 8.1%, but this increase was not statistically significantly different from that of the control. Lastly, knocking out all three reactions reduced the CAT yield to 0.7%, similar to knocking out the *cyd* alone. Thus, the model suggested the key step in oxidative phosphorylation was catalyzed by the gene product of *cyd*. However, while disruption of *cyd* significantly reduced the CAT carbon yield, it did not completely eliminate the production of CAT. This suggested there was a mixture of energy sources supporting CAT production, with the most significant being oxidative phosphorylation.

Sequence-specific flux balance analysis (ssFBA) predicted optimal CAT yields with no adjustable parameters (Fig. 6). Before exploring CFPS optimality, we first validated the ssFBA approach by comparing simulated and measured concentrations of CAT for the first hour of glucose consumption. We chose this time window (during the first phase of CAT production) because it was approximately linear both in glucose consumption and in the accumulation of most organic acids. As the ssFBA calculation had no adjustable parameters, bounds on transcription and translation rates and biochemical fluxes were either estimated from data or from mechanistic models parameterized from literature. Uncertainty in experimental factors such as RNA polymerase, ribosome concentrations, elongation rates, or the upper bounds for oxygen and glucose consumption rates was addressed by sampling plausible ranges for these parameters. The ensemble of ssFBA simulations predicted CAT formation as a function of time during the first hour of production when constrained by the experimental metabolite data (Fig. 6C). Thus, the molecular description of transcription and translation were consistent with experimental measurements. Next, to gauge the performance of the CFPS reaction, we next calculated the CAT carbon yield for three classes of constraints: (i) theoretical maximum glucose, amino acid and oxygen upper bounds, and realistic transcriptional/translational constraints; (ii) theoretical maximum glucose, amino acid and oxygen upper bounds, realistic transcriptional/translational constraints and knockouts of amino acid synthesis reactions of amino acids supplemented in the *E. coli* extract preparation. (iii) metabolite fluxes constrained by the CAT data, and realistic transcriptional/translational constraints and knockouts of amino acid synthesis reactions of amino acids supplemented in the *E. coli* extract preparation (Fig. 6D). The physiological theoretical maximum CAT carbon yield (case i) was 49.3% $\pm$ 3.5% (Fig. 6D, left); this represents optimal network performance if glucose, oxygen and amino acids were produced or consumed at their upper bounds, with bounded transcription and translation rates (96% without glucose contribution in the carbon yield calculation). For case ii, the optimal CAT carbon yield decreased to 48.9% $\pm$ 3.5% (Fig. 6D, middle). Lastly, when metabolite constraints based on experimental measurements were applied (case iii), the predicted carbon yield was 6.4% $\pm$ 2.9% (Fig. 6D, right). Unsurprisingly, this range of carbon yield encompasses both the best-fit set (7.9%) and the experimental dataset (8.2%). For cases i and ii the energy efficiencies were 72.1% and 71.2%, respectively, while for case iii it was only 5.1%. This dramatic decrease in efficiency when fluxes are constrained to data makes sense, as the network is forced toward a multitude of pathways that may not contribute to CAT production. However,

To investigate Taken together, a *gnd* knockout decreased acetate production and required less amino acid consumption, thus it is a promising strategy to increase the CAT carbon yield.

Discussion

In this study we present an ensemble of *E. coli* cell-free protein synthesis (CFPS) models that accurately predict a comprehensive CFPS dataset of glucose, CAT, central carbon metabolites, energy species, and amino acid measurements. We used the hybrid cell-free modeling approach of Wayman and coworkers, which integrates traditional kinetic modeling with a logic-based description of allosteric regulation. CFPS is seen to be biphasic relying on glucose during the first hour and pyruvate and lactate afterward. Allosteric control was essential to the maintenance of the network and production of CAT, as without it, central carbon metabolism is exhuasted within 1.5 hours leading to low CAT production. Having captured the experimental data, we investigated if CAT yield and CFPS performance could be further improved. We showed that the model produces CAT with a carbon yield equal to 36%, and an energy efficiency equal to 9%, of that of a physiological case in which transcription and translation are constrained. The accumulation of waste byproducts, especially acetate, is responsible for this sub-optimal yield. Sensitivity analysis showed that certain substrates and energy species are instrumental to CAT production and overall metabolism. The system heavily relied on oxidative phosphorylation for the system's energetic needs as well as for CAT synthesis. A single knockout in oxidative phosphorylation reduced the CAT carbon yield ~3-fold, as well as disrupting the system state showing its crucial role in CFPS. In comparing flux distributions between low and high yield cases, carbon flux could be potentially diverted toward CAT by reducing acetate overflow and minimizing flux through the Entner-Doudoroff pathway. Taken together, these findings represent the first dynamic model of *E. coli* cell-free protein synthesis, and an important step toward a functional genome scale description.

We present an ensemble of models that quantitatively describes the system behavior of cell-free metabolism and production of CAT. Experimental observations of the metabolites and cometabolites validate the structure of the model and the estimation of kinetic parameters. This is important in applying metabolic engineering principles to rationally design cell-free production processes and predict the redirection of carbon fluxes to product forming pathways. In analyzing the model parameters' effect on CAT production, CAT synthesis is the most important, followed by oxidative phosphorylation and the glutamate and pyruvate consuming reactions, as well as cofactor reactions which are necessary to drive CAT synthesis. For example, the conversion of ATP to GTP shows significance since it is necessary for CAT synthesis. While Jewett and coworkers have shown that ATP may be at saturation in CFPS [1], GTP is also required for CAT synthesis and may be a limiting reactant. Thus, supplementation with additional GTP may improve the efficiency of CAT production. A similar theme is seen in the sensitivity of overall model state, where the most important reactions are glucose and pyruvate consuming reactions and cofactor reactions which are vital to drive CFPS. This can be seen in the biphasic operation of CFPS, with the first phase operating on glucose and the second phase operating on pyruvate. During the first phase, there is an accumulation of byproducts from central carbon with the majority of flux going toward acetate and some toward pyruvate, lactate, and succinate; with the exception of acetate, these are all consumed in the second phase. This shows that CAT production can be sustained by pyruvate and glutamate in the absence of glucose, which provides alternative strategies to optimize CFPS performance. This is in accordance with literature, which showed pyruvate provided a relatively slow but continuous supply of ATP [25]. Taken together, this shows CFPS can be designed towards a specified application either requiring a slow stable energy source or faster production. This outstanding control on model performace was expected as these metabolites are responsible for driving CFPS and represent the first step in the model network. Nevertheless, there are further reactions with considerable impact on model performance. In examining oxidative phosphorylation activity, knockouts in the electron transport pathways disrupt metabolism across the network and show CAT carbon yield dropping from 8.6%

to 2.7%; Jewett and coworkers also saw a decrease in CAT yield, ranging from 1.5-fold to 4-fold, when knocking out oxidative phosphorylation reactions[1]. Oxidative phosphorylation is vital, since it provides most of the energetic needs of CFPS. However, it is unknown how active oxidative phosphorylation is compared to that of *in vivo* systems, and both of our modeling approaches suggest its importance to CAT production and CFPS. Thus, oxidative phosphorylation is a potential area for improvement for CFPS performance and protein yield. Comparing the physiologically realistic carbon yield of CAT from ssFBA predictions to those of the kinetic model and experimental measurements suggests that there is potential for increasing CAT yield as well as CFPS performance. A knockout of *gnd* and shows that carbon can be diverted away from acetate and toward CAT or other proteins of interest expressed in CFPS. Another limitation to be addressed in CFPS is the transcription and translation description, since protein production is ultimately bounded by these kinetic rates. Li et al. have increased productivity of firefly lucifease by 5-fold in CFPS systems by adding and adjusting factors that affect transcription and translation such as elongation factors, ribosome recyclcing factor, release factors, chaperones, BSA, and tRNAs [26]. Underwood and coworkers have also shown that an increase in ribosome levels does not significantly increase protein yields or rates; however, adding elongation factors increased yields by 23% at 30 minutes [27].

A logical next step for this work would be sequence-specific dynamic modeling, as the kinetic modeling approach in this study used a single reaction to approximate CAT synthesis. Including specific transcription and translation steps for CAT would allow more accurate modeling of the complexity and the resource cost of protein synthesis. In addition, sensitivity analysis could be performed on these new parameters to determine the robustness of CAT synthesis to the processes of transcription and translation. Another area for future work is to more thoroughly sample parameter space. Parameters were varied so as to best fit the dataset; however, the resulting ensemble may not represent every biological possibility. In a different region of parameter space, the system may behave differently but still fit the experimental data. This could include the flux distribution through the network, the variation of predictions across the ensemble, and the relative sensitivity values. Testing the model under a variety of conditions could strengthen or challenge the findings of this study. Further experimentation could also be used to gain a deeper understanding of model performance under a variety of conditions. Specifically, CAT production performed in the absence of amino acids could inform the system's ability to manufacture them, while experimentation in the absence of glucose or oxygen could shed light on how important they are to protein synthesis, and under which conditions. Finally, the approach should be extended to other protein products. CAT is only a test protein used for model identification; the modeling framework, and to some extent the parameter values, should be protein agnostic. An important extension of this study would be to apply its insights to other protein applications, where possible.

Materials and Methods

Formulation and solution of the model equations. We used ordinary differential equations (ODEs) to model the time evolution of metabolite (x_i) and scaled enzyme abundance (ϵ_i) in hypothetical cell-free metabolic networks:

$$\frac{dx_i}{dt} = \sum_{j=1}^{\mathcal{R}} \sigma_{ij} r_j \left(\mathbf{x}, \epsilon, \mathbf{k}\right) \qquad i = 1, 2, \ldots, \mathcal{M} \tag{1}$$

$$\frac{d\epsilon_i}{dt} = -\lambda_i \epsilon_i \qquad i = 1, 2, \ldots, \mathcal{E} \tag{2}$$

where $\mathcal{R}$ denotes the number of reactions, $\mathcal{M}$ denotes the number of metabolites and $\mathcal{E}$ denotes the number of enzymes in the model. The quantity $r_j \left(\mathbf{x}, \epsilon, \mathbf{k}\right)$ denotes the rate of reaction j. Typically, reaction j is a non-linear function of metabolite and enzyme abundance, as well as unknown kinetic parameters $\mathbf{k}$ ($\mathcal{K} \times 1$). The quantity σ_{ij} denotes the stoichiometric coefficient for species i in reaction j. If $\sigma_{ij} > 0$, metabolite i is produced by reaction j. Conversely, if $\sigma_{ij} < 0$, metabolite i is consumed by reaction j, while $\sigma_{ij} = 0$ indicates metabolite i is not connected with reaction j. Lastly, λ_i denotes the scaled enzyme activity decay constant. The system material balances were subject to the initial conditions $\mathbf{x}\left(t_o\right) = \mathbf{x}_o$ and $\epsilon\left(t_o\right) = \mathbf{1}$ (initially we have 100% cell-free enzyme abundance).

The reaction rate was written as the product of a kinetic term $(\bar{r}_j)$ and a control term (v_j), $r_j \left(\mathbf{x}, \mathbf{k}\right) = \bar{r}_j v_j$. We used multiple saturation kinetics to model the reaction term $\bar{r}_j$:

$$\bar{r}_j = V_j^{max} \epsilon_i \prod_{s \in m_j^-} \frac{x_s}{K_{js} + x_s} \tag{3}$$

where V_j^{max} denotes the maximum rate for reaction j, ϵ_i denotes the scaled enzyme activity which catalyzes reaction j, K_{js} denotes the saturation constant for species s in reaction j and m_j^- denotes the set of *reactants* for reaction j. On the other hand, the control term $0 \leq v_j \leq 1$ depended upon the combination of factors which influenced rate process j. For each rate, we used a rule-based approach to select from competing control factors. If rate j was influenced by $1, \ldots, m$ factors, we modeled this relationship as $v_j = \mathcal{I}_j\left(f_{1j}\left(\cdot\right), \ldots, f_{mj}\left(\cdot\right)\right)$ where $0 \leq f_{ij}\left(\cdot\right) \leq 1$ denotes a transfer function quantifying the influence of factor i on rate j. The function $\mathcal{I}_j\left(\cdot\right)$ is an integration rule which maps the output of regulatory transfer functions into a control variable. We used hill-like transfer functions and $\mathcal{I}_j \in \{min, max\}$ in this study [24].

We included 17 allosteric regulation terms, taken from literature, in the CFPS model. PEP was modeled as an inhibitor for phosphofructokinase [28, 29], PEP carboxykinase [28], PEP synthetase [28, 30], isocitrate dehydrogenase [28, 31], and isocitrate lyase/malate synthase [28, 31, 32], and as an activator for fructose-biphosphatase [28, 33–35]. AKG was modeled as an inhibitor for citrate synthase [28, 36, 37] and isocitrate lyase/malate synthase [28, 32]. 3PG was modeled as an inhibitor for isocitrate lyase/malate synthase [28, 32]. FDP was modeled as an activator for pyruvate kinase [28, 38] and PEP carboxylase [28, 39]. Pyruvate was modeled as an inhibitor for pyruvate dehydrogenase [28, 40, 41] and as an activator for lactate dehydrogenase [42]. Acetyl CoA was modeled as an inhibitor for malate dehydrogenase [28].

Estimation of kinetic model parameters. We estimated an ensemble of diverse parameter sets using a constrained Markov Chain Monte Carlo (MCMC) random walk strategy. Starting from a single best fit parameter set estimated by inspection and literature, we calculated the cost function, equal to the sum-squared-error between experimental data and model predictions:

$$\texttt{cost} = \sum_{i=1}^{\mathcal{D}} \left[\frac{w_i}{\mathcal{Y}_i^2} \sum_{j=1}^{\mathcal{T}_i} \left(y_{ij} - x_i\big|_{t(j)}\right)^2\right] \tag{4}$$

where $\mathcal{D}$ denotes the number of datasets ($\mathcal{D}$ = 37), w_i denotes the weight of the i^{th} dataset, $\mathcal{T}_i$ denotes the number of timepoints in the i^{th} dataset, $t(j)$ denotes the j^{th} timepoint, y_{ij} denotes the measurement value of the i^{th} dataset at the j^{th} timepoint, and $x_i|_{t(j)}$ denotes the simulated value of the metabolite corresponding to the i^{th} dataset, interpolated to the j^{th} timepoint. Lastly, the cost calculation was scaled by the maximum experimental value in the i^{th} dataset, $\mathcal{Y}_i = \max_j(y_{ij})$. We then perturbed each model parameter between an upper and lower bound that varied by parameter type:

$$k_i^{new} = \min\left(\max\left(k_i \cdot \exp(a \cdot r_i), l_i\right), u_i\right) \qquad i = 1, 2, \ldots, \mathcal{P} \tag{5}$$

where $\mathcal{P}$ denotes the number of parameters ($\mathcal{P}$ = 815), which includes 163 maximum reaction rates (V^{max}), 163 enzyme activity decay constants, 455 saturation constants (K_{js}), and 34 control parameters, k_i^{new} denotes the new value of the i^{th} parameter, k_i denotes the current value of the i^{th} parameter, a denotes a distribution variance, r_i denotes a random sample from the normal distribution, l_i denotes the lower bound for that parameter type, and u_i denotes the upper bound for that parameter type. Maximum reaction rates were bounded between 0 and 500,000 mM/h [43]. Assuming a total enzyme concentration of 5.0 μM, this corresponds to catalytic rate bounds of 0 and 27,780 s^{-1}. These bounds resulted in a median catalytic rate of 0.16 s^{-1} across the ensemble. Enzyme activity decay constants were bounded between 0 and 1 h^{-1}, corresponding to half lives of 42 minutes and infinity; median = 25 h. Saturation constants were bounded between 0.001 and 10 mM; median = 0.16 mM. Control parameters (gains and orders) were left unbounded; gain median = 0.076, order median = 0.69. For each newly generated parameter set, we re-solved the balance equations and calculated the cost function. All sets with a lower cost (and some with higher cost) were accepted into the ensemble. After generating greater that 10,000 sets, we selected N = 100 sets with minimal set to set correlation to avoid over-sampling any region of parameter space.

Sensitivity analysis of the kinetic CFPS model. We determined the reactions most important to protein production by computing the local sensitivity of CAT concentration (denoted as $\mathtt{CAT}$) to each individual maximum reaction rate, and each pair of maximum reaction rates in the network. The sensitivity index was formulated as:

$$\mathcal{S}_{ij}^{\mathtt{CAT}} = \left\|\mathtt{CAT}(p_i, p_j, t) - \mathtt{CAT}(\alpha \cdot p_i, \alpha \cdot p_j, t)\right\|_2 \qquad i, j = 1, 2, \dots \mathcal{P} \tag{6}$$

where $\mathcal{S}_{ij}^{\mathtt{CAT}}$ denotes the sensitivity of CAT production to the i^{th} and j^{th} parameters, $\mathtt{CAT}(p_i, p_j, t)$ denotes CAT concentration as a function of time and the i^{th} and j^{th} parameters, α denotes the perturbation factor, and $\mathcal{P}$ denotes the number of maximum reaction rates ($\mathcal{P}$ = 163). In calculating the pairwise sensitivities, each parameter was perturbed by 1%; first-order sensitivities (i = j) were subject to two 1% perturbations. Parameters and parameter combinations were stratified into five degrees of importance, from least to most sensitive.

Likewise, we determined which reactions were most important to global system performance by computing the sensitivity of all species for which data exists (denoted as $\mathtt{X}$) to each maximum reaction rate in the network. In this case, each sensitivity index was formulated as:

$$\mathcal{S}_{ij}^{\mathtt{X}} = \left\|\mathtt{X}(p_i, p_j, t) - \mathtt{X}(\alpha \cdot p_i, \alpha \cdot p_j, t)\right\|_2 \qquad i, j = 1, 2, \dots \mathcal{P} \tag{7}$$

where $\mathcal{S}_{ij}^{\mathtt{X}}$ denotes the sensitivity of the system state to the i^{th} and j^{th} parameters, and $\mathtt{X}(p_i, p_j, t)$ denotes the system state, an array consisting of the concentration of every species for which data exists as a function of time and the i^{th} and j^{th} parameters. The parameter sensitivities were stratified into five degrees of importance, from least to most sensitive, as above.

Sequence-specific calculation of carbon yield. We estimated the theoretical maximum CAT carbon yield using sequence-specific flux balance analysis (ssFBA) [44]. The sequence-specific flux balance analysis problem was formulated as a linear program:

$$\begin{aligned}
&\max_{\boldsymbol{w}} \left(w_{TL} = \boldsymbol{\theta}^{T}\boldsymbol{w}\right)\\
&\text{Subject to}: \ \mathbf{Sw} = \mathbf{0}\\
&\alpha_i \leq w_i \leq \beta_i \qquad i = 1, 2, \ldots, \mathcal{R}
\end{aligned} \tag{8}$$

where 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 [44]:

$$\begin{aligned}
\mathrm{G}_{\mathcal{P}} + \mathrm{R}_1 &\longrightarrow \mathrm{G}^{*}_{\mathcal{P}}\\
\mathrm{G}^{*}_{\mathcal{P}} + \sum_{\mathrm{k}\in\{\mathrm{A,C,G,U}\}} \eta_k \cdot \{\mathrm{k}\}\,\mathrm{TP} &\xrightarrow{\mathrm{TX}} \mathrm{mRNA} + \mathrm{G}_{\mathcal{P}} + \mathrm{R}_1 + \sum_{\mathrm{k}\in\{\mathrm{A,C,G,U}\}} 2\eta_{\mathrm{k}} \cdot \mathrm{Pi}\\
\mathrm{mRNA} &\longrightarrow \sum_{\mathrm{k}\in\{\mathrm{A,C,G,U}\}} \eta_{\mathrm{k}} \cdot \{\mathrm{k}\}\,\mathrm{MP}\\
\mathrm{mRNA} + \mathrm{R}_2 &\longrightarrow \mathrm{R}^{*}_2\\
\alpha_{\mathrm{j}} \cdot \mathrm{AA}_{\mathrm{j}} + \alpha_{\mathrm{j}} \cdot \mathrm{tRNA} + \alpha_{\mathrm{j}} \cdot \mathrm{ATP} &\longrightarrow \alpha_{\mathrm{j}} \cdot \mathrm{AA}_{\mathrm{j}} - \mathrm{tRNA}_{\mathrm{j}} +\\
&\qquad \alpha_{\mathrm{j}} \cdot \mathrm{AMP} + 2\alpha_{\mathrm{j}} \cdot \mathrm{Pi} \qquad j = 1, 2, \ldots, 20\\
\mathrm{R}^{*}_2 + \sum_{\mathrm{j}\in\{\mathrm{AA}\}} \alpha_{\mathrm{j}} \cdot \Big(\mathrm{AA}_{\mathrm{j}} - \mathrm{tRNA}_{\mathrm{j}} + 2 \cdot \mathrm{GTP}\Big) &\xrightarrow{\mathrm{TL}} \mathcal{P} + \mathrm{R}_2 + \mathrm{mRNA} +\\
&\qquad + \sum_{\mathrm{j}\in\{\mathrm{AA}\}} \alpha_{\mathrm{j}} \cdot \Big(\mathrm{tRNA} + 2 \cdot \mathrm{GDP} + 2 \cdot \mathrm{Pi}\Big)
\end{aligned}$$

where $G_{\mathcal{P}}$ denotes the gene encoding protein product $\mathcal{P}$, 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, R_2 denotes the ribosome concentration, R^*_2 denotes bounded ribosome, and AA_j denotes j^{th} amino acid.

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

$$w_{TX} = V^{max}_{TX} \left(\frac{G}{K_{TX} + G} \right) \tag{9}$$

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:

$$V^{max}_{TX} \equiv \left[R_1 \left(\frac{v_{TX}}{l_G} \right) \left(\frac{K_{T7}}{1 + K_{T7}} \right) \right] \tag{10}$$

The term R_1 denotes the RNA polymerase abundance, v_{TX} denotes the RNA polymerase elongation rate (nt/hr), l_G denotes the gene length in nucleotides, and the last term describes T7 promoter activity, where K_{T7} denotes a T7 RNA polymerase binding constant [45]. On the other hand, the translation rate (w_{TL}) was bounded by:

$$0 \leq w_{TL} \leq V^{max}_{TL} \left(\frac{\mathrm{mRNA}_{\mathrm{SS}}}{K_{TL} + \mathrm{mRNA}_{\mathrm{SS}}} \right) \tag{11}$$

where $\mathrm{mRNA}_{\mathrm{SS}}$ denotes the steady state mRNA abundance, and K_{TL} denotes the translation saturation constant. The maximum translation rate V^{max}_{TL} was formulated as:

$$V^{max}_{TL} \equiv \left[K_P R_2 \left(\frac{v_{TL}}{l_P} \right) \right] \tag{12}$$

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 $\mathrm{mRNA_{SS}}$ denotes the steady-state mRNA concentration:

$$\mathrm{mRNA_{SS}} \simeq \frac{\mathrm{w_{TX}}}{\lambda} \tag{13}$$

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 CAT 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. The CAT gene and protein sequences were taken from literature. The sequence-specific flux balance linear program was solved using the GNU Linear Programming Kit (GLPK) v4.52 [46].

Quantification of uncertainty. An ensemble of 100 sets of flux distributions was calculated for three different cases: constrained by transcription/translation rates, constrained by transcription/translation rates without amino acid synthesis reactions, and constrained by transcription/translation rates and experimental measurements without amino acid synthesis reactions. For the first case, all rates were left unbounded, except the specific glucose uptake rate, transcription and translation rate. An ensemble of flux distributions was then calculated by randomly sampling the maximum specific glucose uptake rate from within a range of 30 to 40 mM/h, determined from experimental data and randomly sampling RNAP polymerase levels, ribosome levels, and elongation rates in a physiological range determined from literature.. For the second case, an ensemble was generated by randomly sampling the same parameters as the first case, however certain amino acid synthesis reactions were removed from the network. This included all the amino acids that were present in the preparation of the *E. coli* extract (alanine, arginine, aspartate, cysteine, glutamate, glutamine and serine were excluded from the media), thus reactions producing the excluded amino acids were left in the network. 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 [27, 47]. For the third case, the ensemble was generated as in the second case, in addition to the lower and upper bounds on the fluxes for the data-informed metabolites were sampled within the range given by the experimental noise. This included the data for glucose, organic acids, energy species, and amino acids; CAT was not constrained by experimental data, but by the transcription/translation rates as stated above.

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

$$Y_C^{CAT} = \frac{\Delta\mathtt{CAT} \cdot C_{CAT}}{\displaystyle\sum_{i=1}^{\mathcal{R}} \max(\Delta m_i, 0) \cdot C_{m_i}} \tag{14}$$

where $\Delta\mathtt{CAT}$ denotes the abundance of CAT produced, C_{CAT} denotes carbon number of CAT, $\mathcal{R}$ denotes the number of reactants, Δm_i denotes the amount of the i^{th} reactant consumed (never allowed to be negative), and C_{m_i} denotes the carbon number of the i^{th} reactant. Arginine and glutamate were excluded from the yield calculations for the experimental yield calculation, as no experimental measurements were available for these amino acids. Yield of the best-fit parameter set and the experimental data were calculated by setting $\Delta\mathtt{CAT}$ equal to the final minus the initial CAT concentration, and setting Δm_i equal to the initial minus the final reactant concentration. The individual CAT production and substrate consumption terms for the best-fit set, kinetic models with knockouts, and experimental data are shown in Table 1. Total net consumption of amino acids and amino acid consumption via CAT synthesis were calculated for the best-fit set (Table **??**). Total net consumption was calculated as amino acid concentration at 0 hours minus concentration at 3 hours; it was negative if synthesis outweighed consumption. Consumption toward CAT was calculated as CAT concentration at 3 hours minus concentration at 0 hours, times the stoichiometric coefficient for that amino acid in the CAT synthesis reaction. The difference between these was defined as other consumption, equal to consumption from reactions other than CAT synthesis minus amino acid production.

Calculation of energy efficiency. Energy efficiency was also calculated for the best-fit set:

$$\mathrm{Efficiency} = \frac{\Delta\texttt{CAT} \cdot (2 \cdot \mathrm{ATP}_{\mathrm{CAT}} + \mathrm{GTP}_{\mathrm{CAT}})}{\Delta\texttt{ATP}} \tag{15}$$

where $\mathrm{Efficiency}$ denotes the energy efficiency of CAT production, $\mathrm{ATP}_{\mathrm{CAT}}$ denotes the stoichiometric coefficient of ATP in CAT synthesis (multiplied by 2 because AMP, rather than ADP, is a product of CAT synthesis), $\mathrm{GTP}_{\mathrm{CAT}}$ denotes the stoichiometric coefficient of GTP in CAT synthesis, and $\Delta\mathrm{ATP}$ denotes the amount of ATP production in 3 hours, equal to the sum of ATP-producing fluxes integrated over the timecourse. $\mathrm{ATP}_{\mathrm{CAT}}$ = 219, $\mathrm{GTP}_{\mathrm{CAT}}$ = 438.

For sequence-specific flux balance analysis, with a more in-depth biological description of CAT synthesis, energy efficiency was calculated slightly differently:

$$\mathrm{Efficiency} = \frac{\Delta\texttt{CAT} \cdot (2\,(\mathrm{ATP}_{\mathrm{TX}} + \mathrm{CTP}_{\mathrm{TX}} + \mathrm{GTP}_{\mathrm{TX}} + \mathrm{UTP}_{\mathrm{TX}}) + 2 \cdot \mathrm{ATP}_{\mathrm{TL}} + \mathrm{GTP}_{\mathrm{TL}})}{\Delta\texttt{ATP}} \tag{16}$$

where $\mathrm{ATP}_{\mathrm{TX}}$, $\mathrm{CTP}_{\mathrm{TX}}$, $\mathrm{GTP}_{\mathrm{TX}}$, $\mathrm{UTP}_{\mathrm{TX}}$ denote the stoichiometric coefficients of each energy species for CAT transcription, and $\mathrm{ATP}_{\mathrm{TL}}$, $\mathrm{GTP}_{\mathrm{TL}}$ denote the stoichiometric coefficients of ATP and GTP for CAT translation. $\mathrm{ATP}_{\mathrm{TX}}$ = 176, $\mathrm{CTP}_{\mathrm{TX}}$ = 144, $\mathrm{GTP}_{\mathrm{TX}}$ = 151, $\mathrm{UTP}_{\mathrm{TX}}$ = 189, $\mathrm{ATP}_{\mathrm{TL}}$ = 219, $\mathrm{GTP}_{\mathrm{TL}}$ = 438.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

J.V directed the modeling study. K.C and J.S conducted the cell free protein synthesis experiments. J.V, J.W, and N.H developed the cell free protein synthesis mathematical model, and parameter ensemble. J.V and M.V performed the sequence-specific flux balance analysis calculations. The manuscript was prepared and edited for publication by J.S, N.H, M.V, J.W and J.V.

Acknowledgements

We gratefully acknowledge the suggestions from the anonymous reviewers to improve this manuscript.

Funding

This study was supported by a National Science Foundation Graduate Research Fellowship (DGE-1333468) to N.H. Research reported in this publication was also supported by the Systems Biology Coagulopathy of Trauma Program with support from the US Army Medical Research and Materiel Command under award number W911NF-10-1-0376.

References

1. Jewett MC, Calhoun KA, Voloshin A, Wuu JJ, Swartz JR. An integrated cell-free metabolic platform for protein production and synthetic biology. Mol Syst Biol. 2008;4:220. doi:10.1038/msb.2008.57.
2. Matthaei JH, Nirenberg MW. Characteristics and stabilization of DNAase-sensitive protein synthesis in E. coli extracts. Proc Natl Acad Sci U S A. 1961;47:1580–8.
3. Nirenberg MW, Matthaei JH. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci U S A. 1961;47:1588–602.
4. Lu Y, Welsh JP, Swartz JR. Production and stabilization of the trimeric influenza hemagglutinin stem domain for potentially broadly protective influenza vaccines. Proc Natl Acad Sci U S A. 2014;111(1):125–30. doi:10.1073/pnas.1308701110.
5. Hodgman CE, Jewett MC. Cell-free synthetic biology: thinking outside the cell. Metab Eng. 2012;14(3):261–9. doi:10.1016/j.ymben.2011.09.002.
6. Pardee K, Slomovic S, Nguyen PQ, Lee JW, Donghia N, Burrill D, et al. Portable, On-Demand Biomolecular Manufacturing. Cell. 2016;167(1):248–59.e12. doi:10.1016/j.cell.2016.09.013.
7. Fredrickson AG. Formulation of structured growth models. Biotechnol Bioeng. 1976;18(10):1481–6. doi:10.1002/bit.260181016.
8. Domach MM, Leung SK, Cahn RE, Cocks GG, Shuler ML. Computer model for glucose-limited growth of a single cell of Escherichia coli B/r-A. Biotechnol Bioeng. 1984;26(3):203–16. doi:10.1002/bit.260260303.
9. Steinmeyer DE, Shuler ML. Structured model for Saccharomyces cerevisiae. Chem Eng Sci. 1989;44:2017–30.
10. Wu P, Ray NG, Shuler ML. A single-cell model for CHO cells. Ann N Y Acad Sci. 1992;665:152–87.

11. Castellanos M, Wilson DB, Shuler ML. A modular minimal cell model: purine and pyrimidine transport and metabolism. Proc Natl Acad Sci U S A. 2004;101(17):6681–6. doi:10.1073/pnas.0400962101.

12. Atlas JC, Nikolaev EV, Browning ST, Shuler ML. Incorporating genome-wide DNA sequence information into a dynamic whole-cell model of Escherichia coli: application to DNA replication. IET Syst Biol. 2008;2(5):369–82. doi:10.1049/iet-syb:20070079.

13. Lewis NE, Nagarajan H, Palsson BØ. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol. 2012;10(4):291–305. doi:10.1038/nrmicro2737.

14. Edwards JS, Palsson BØ. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci U S A. 2000;97(10):5528–33.

15. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ. Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol. 2009;7(2):129–43. doi:10.1038/nrmicro1949.

16. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, et al. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol. 2007;3:121. doi:10.1038/msb4100155.

17. Oh YK, Palsson BØ, Park SM, Schilling CH, Mahadevan R. Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. J Biol Chem. 2007;282(39):28791–9. doi:10.1074/jbc.M703759200.

18. Ibarra RU, Edwards JS, Palsson BØ. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature. 2002;420(6912):186–9. doi:10.1038/nature01149.

19. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol. 2007;3:119. doi:10.1038/msb4100162.
20. Hyduke DR, Lewis NE, Palsson BØ. Analysis of omics data with genome-scale models of metabolism. Mol Biosyst. 2013;9(2):167–74. doi:10.1039/c2mb25453k.
21. McCloskey D, Palsson BØ, Feist AM. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol Syst Biol. 2013;9:661. doi:10.1038/msb.2013.18.
22. Zomorrodi AR, Suthers PF, Ranganathan S, Maranas CD. Mathematical optimization applications in metabolic networks. Metab Eng. 2012;14(6):672–86. doi:10.1016/j.ymben.2012.09.005.
23. Calhoun KA, Swartz JR. An Economical Method for Cell-Free Protein Synthesis using Glucose and Nucleoside Monophosphates. Biotechnology Progress. 2005;21(4):1146–53. doi:10.1021/bp050052y.
24. Wayman JA, Sagar A, Varner JD. Dynamic Modeling of Cell-Free Biochemical Networks Using Effective Kinetic Models. Processes. 2015;3(1):138. doi:10.3390/pr3010138.
25. Swartz J. A PURE approach to constructive biology. Nature Biotechnology. 2001;19:732–3.
26. Li J, Gu L, Aach J, Church GM. Improved Cell-Free RNA and Protein Synthesis System. PLoS ONE. 2014;9(9):1–11. doi:10.1371/journal.pone.0106232.
27. Underwood KA, Swartz JR, Puglisi JD. Quantitative polysome analysis identifies limitations in bacterial cell-free protein synthesis. Biotechnology and Bioengineering. 2005;91(4):425–35. doi:10.1002/bit.20529.
28. Kotte O, Zaugg JB, Heinemann M. Bacterial adaptation through distributed sensing of metabolic fluxes. Mol Syst Biol. 2010;6:355.

29. Cabrera R, Baez M, Pereira HM, Caniuguir A, Garratt RC, Babul J. The crystal complex of phosphofructokinase-2 of Escherichia coli with fructose-6-phosphate: kinetic and structural analysis of the allosteric ATP inhibition. J Biol Chem. 2011;286(7):5774–83.

30. Chulavatnatol M, Atkinson DE. Phosphoenolpyruvate synthetase from Escherichia coli. Effects of adenylate energy charge and modifier concentrations. J Biol Chem. 1973;248(8):2712–5.

31. Ogawa T, Murakami K, Mori H, Ishii N, Tomita M, Yoshin M. Role of phosphoenolpyruvate in the NADP-isocitrate dehydrogenase and isocitrate lyase reaction in Escherichia coli. J Bacteriol. 2007;189(3):1176–8.

32. MacKintosh C, Nimmo HG. Purification and regulatory properties of isocitrate lyase from Escherichia coli ML308. Biochem J. 1988;250(1):25–31.

33. Donahue JL, Bownas JL, Niehaus WG, Larson TJ. Purification and characterization of glpX-encoded fructose 1, 6-bisphosphatase, a new enzyme of the glycerol 3-phosphate regulon of Escherichia coli. J Bacteriol. 2000;182(19):5624–7.

34. Hines JK, Fromm HJ, Honzatko RB. Novel allosteric activation site in Escherichia coli fructose-1,6-bisphosphatase. J Biol Chem. 2006;281(27):18386–93.

35. Hines JK, Fromm HJ, Honzatko RB. Structures of activated fructose-1,6-bisphosphatase from Escherichia coli. Coordinate regulation of bacterial metabolism and the conservation of the R-state. J Biol Chem. 2007;282(16):11696–704.

36. Pereira DS, Donald LJ, Hosfield DJ, Duckworth HW. Active site mutants of Escherichia coli citrate synthase. Effects of mutations on catalytic and allosteric properties. J Biol Chem. 1994;269(1):412–7.

37. Robinson MS, Easom RA, Danson MJ, Weitzman PD. Citrate synthase of Escherichia coli. Characterisation of the enzyme from a plasmid-cloned gene and amplification of the intracellular levels. FEBS Lett. 1983;154(1):51–4.

38. Zhu T, Bailey MF, Angley LM, Cooper TF, Dobson RC. The quaternary structure of pyruvate kinase type 1 from Escherichia coli at low nanomolar concentrations. Biochimie. 2010;92(1):116–20.

39. Wohl RC, Markus G. Phosphoenolpyruvate carboxylase of Escherichia coli. Purification and some properties. J Biol Chem. 1972;247(18):5785–92.

40. Kale S, Arjunan P, Furey W, Jordan F. A dynamic loop at the active center of the Escherichia coli pyruvate dehydrogenase complex E1 component modulates substrate utilization and chemical communication with the E2 component. J Biol Chem. 2007;282(38):28106–16.

41. Arjunan P, Nemeria N, Brunskill A, Chandrasekhar K, Sax M, Yan Y, et al. Structure of the pyruvate dehydrogenase multienzyme complex E1 component from Escherichia coli at 1.85 A resolution. Biochemistry. 2002;41(16):5213–21.

42. Okino S, Suda M, Fujikura K, Inui M, Yukawa H. Production of D-lactic acid by Corynebacterium glutamicum under oxygen deprivation. Appl Microbiol Biotechnol. 2008;78(3):449–54.

43. Milo R, Jorgensen P, Moran U, Weber G, Springer M. BioNumbers–the database of key numbers in molecular and cell biology. Nucleic Acids Res. 2009;38:750–3.

44. Allen TE, Palsson BØ. Sequence-based analysis of metabolic demands for protein synthesis in prokaryotes. J Theor Biol. 2003;220(1):1–18.

45. Moon TS TASBVC Lou C. Genetic programs constructed from layered logic gates in single cells. Nature. 2012;491.

46. type; 2016. Available from: http://www.gnu.org/software/glpk/glpk.html.

47. Garamella J, Marshall R, Rustad M, Noireaux V. The All E. coli TX-TL Toolbox 2.0: A Platform for Cell-Free Synthetic Biology. ACS Synth Biol. 2016;5(4):344–55. doi:10.1021/acssynbio.5b00296.

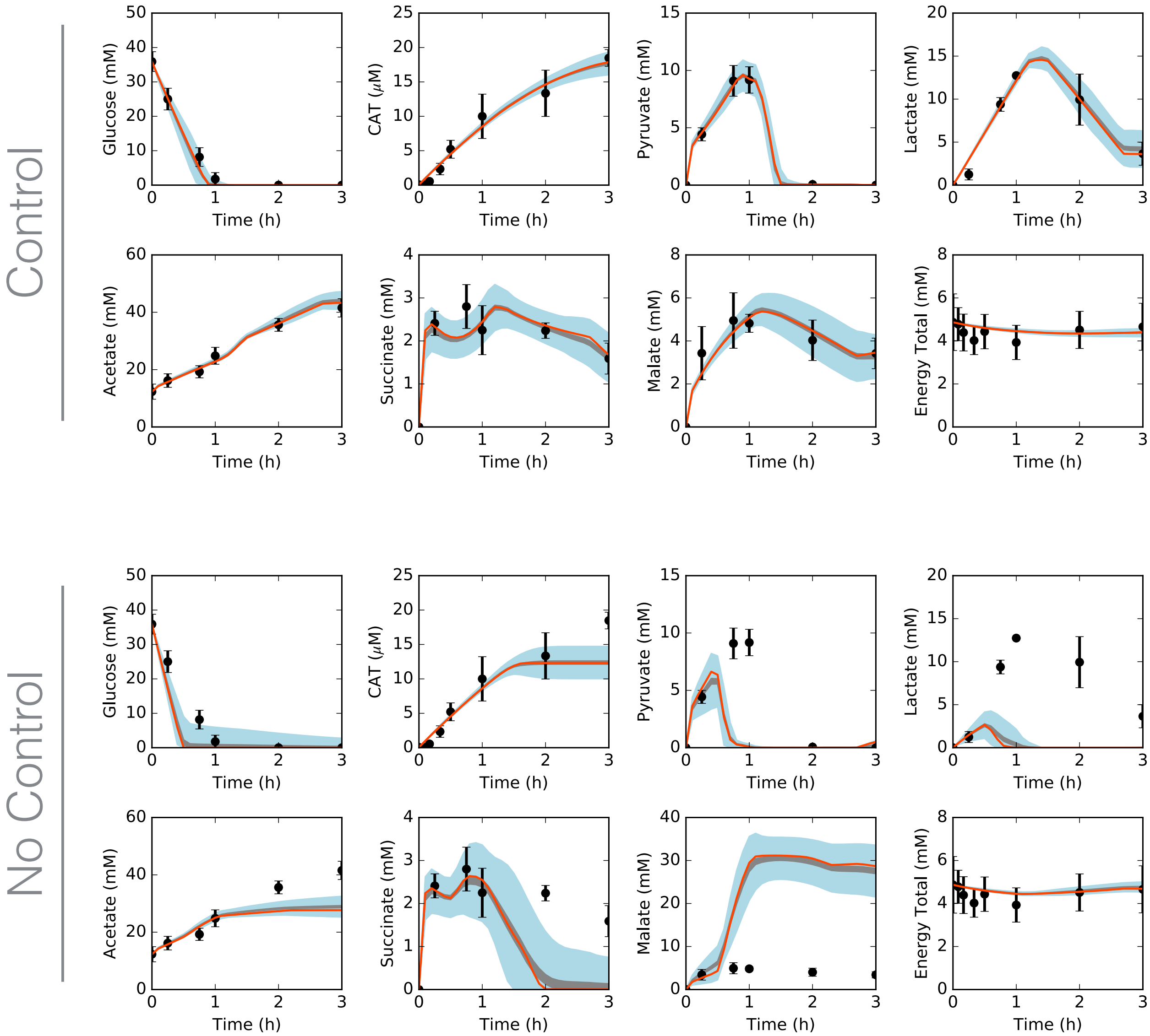

Fig. 1: Central carbon metabolism in the presence (top) and absence (bottom) of allosteric control, including glucose (substrate), CAT (product), and intermediates, as well as total concentration of energy species. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence interval (blue shaded region) and 95% confidence interval of the mean (gray shaded region) over the ensemble of 100 sets.

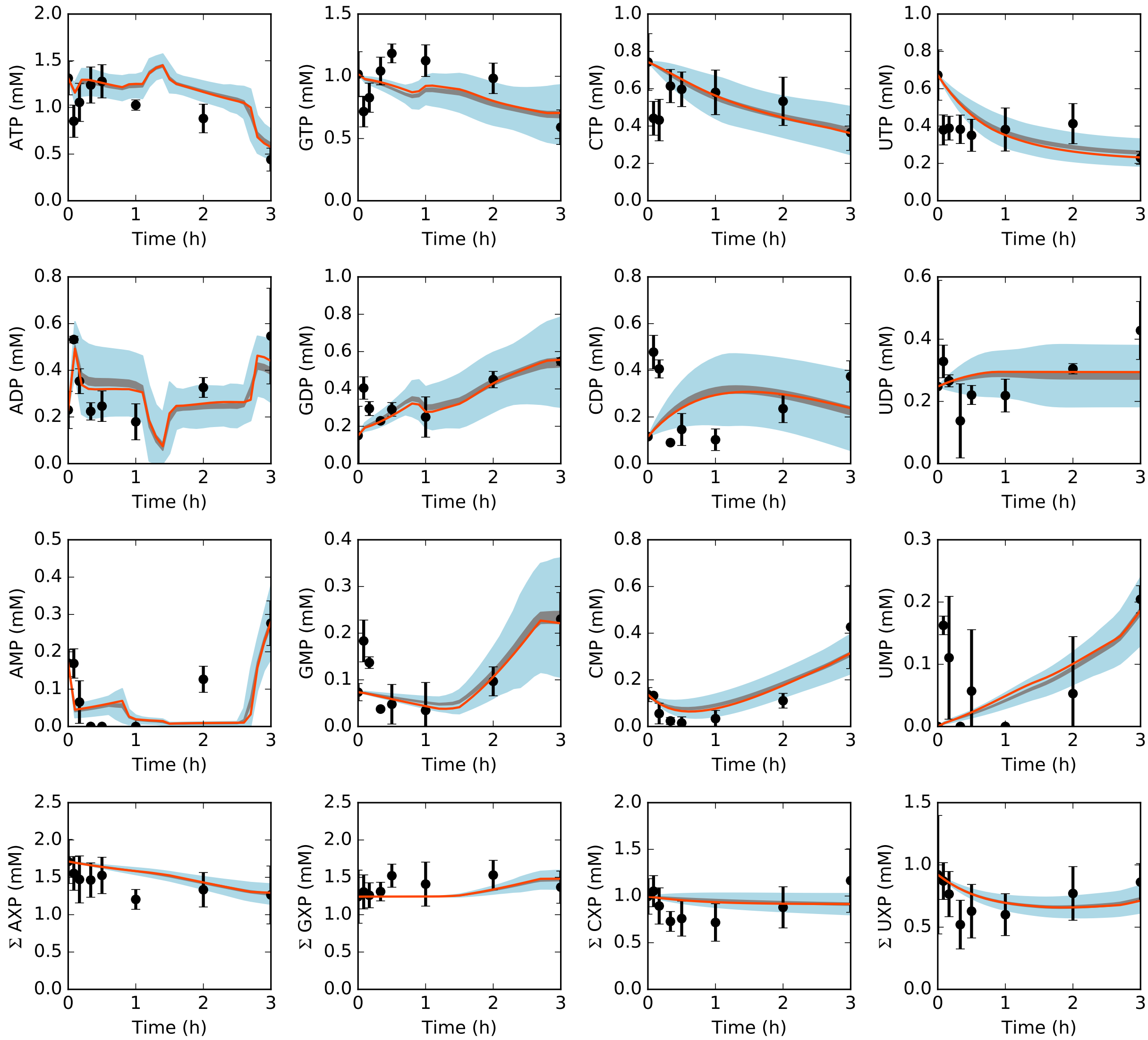

Fig. 2: Energy species and energy totals by base in the presence of allosteric control. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence interval (blue shaded region) and 95% confidence interval of the mean (gray shaded region) over the ensemble of 100 sets.

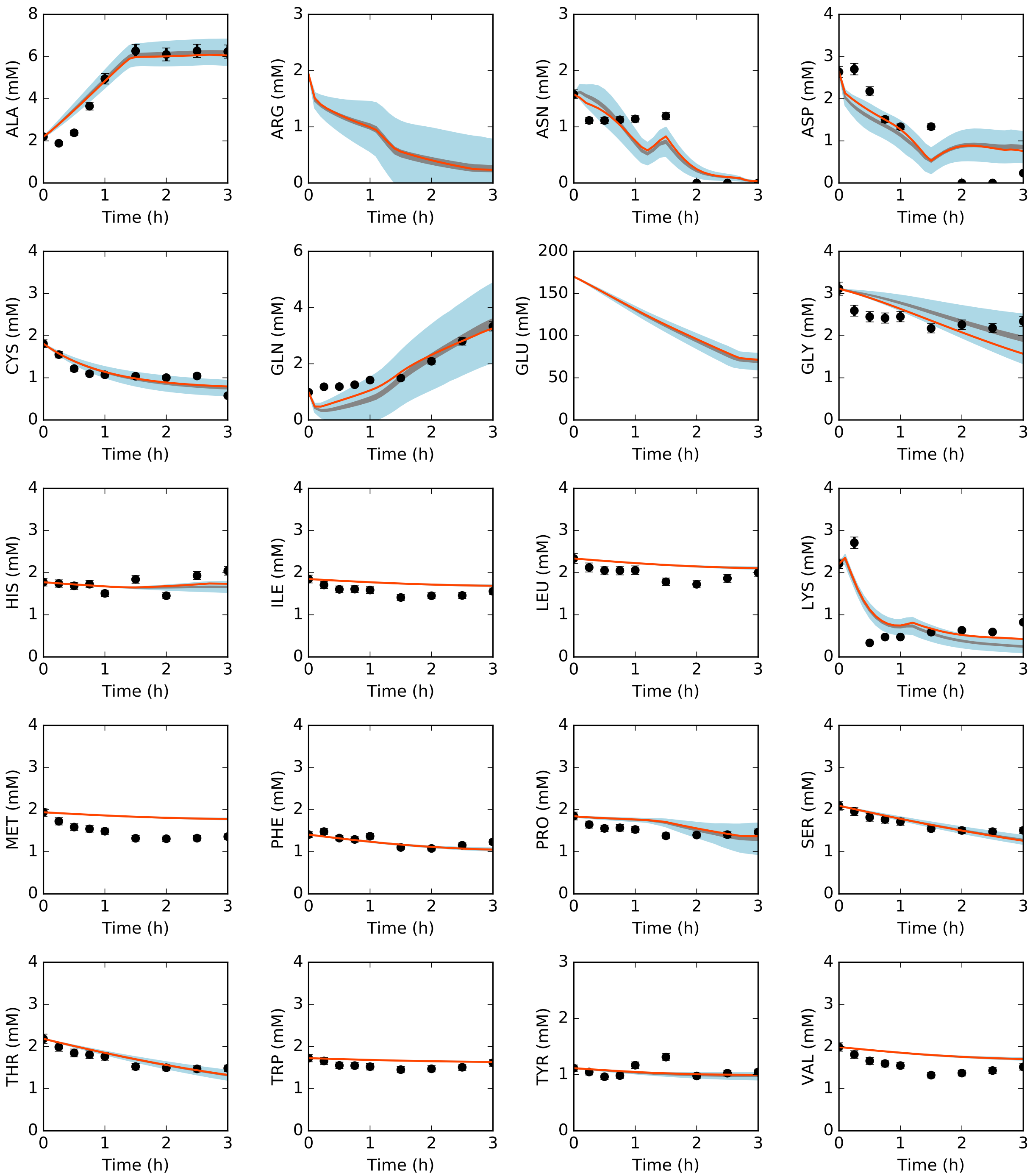

Fig. 3: Amino acids in the presence of allosteric control. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence interval (blue shaded region) and 95% confidence interval of the mean (gray shaded region) over the ensemble of 100 sets.

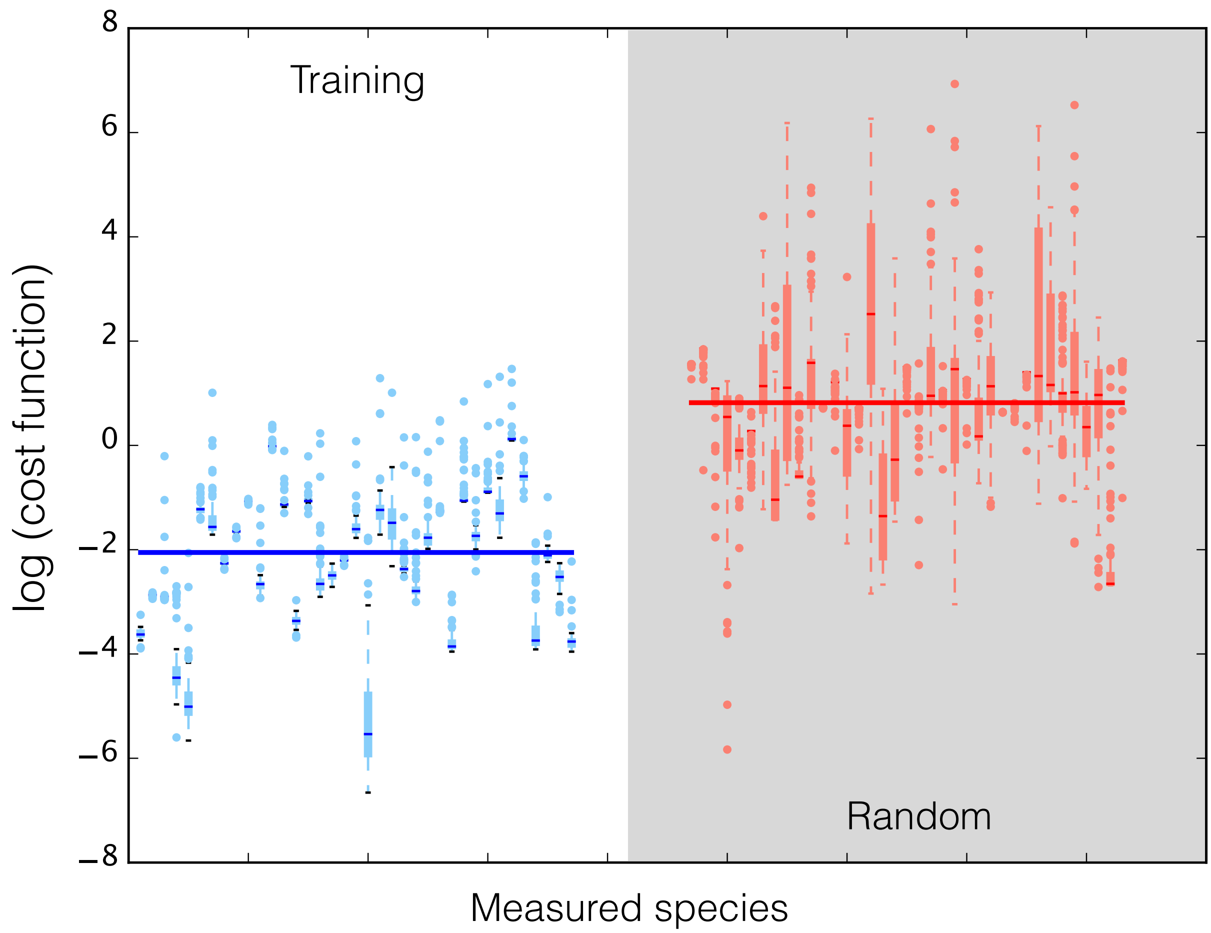

Fig. 4: Log of cost function across 37 datasets for data-trained ensemble (blue) and randomly generated ensemble (red, gray background). Median (bars), interquartile range (boxes), range excluding outliers (dashed lines), and outliers (circles) for each dataset. Median across all datasets (large bar overlaid).

Fig. 5: Normalized first-order and pairwise sensitivities of CAT production (top) and system state (bottom) to maximum reaction rates.

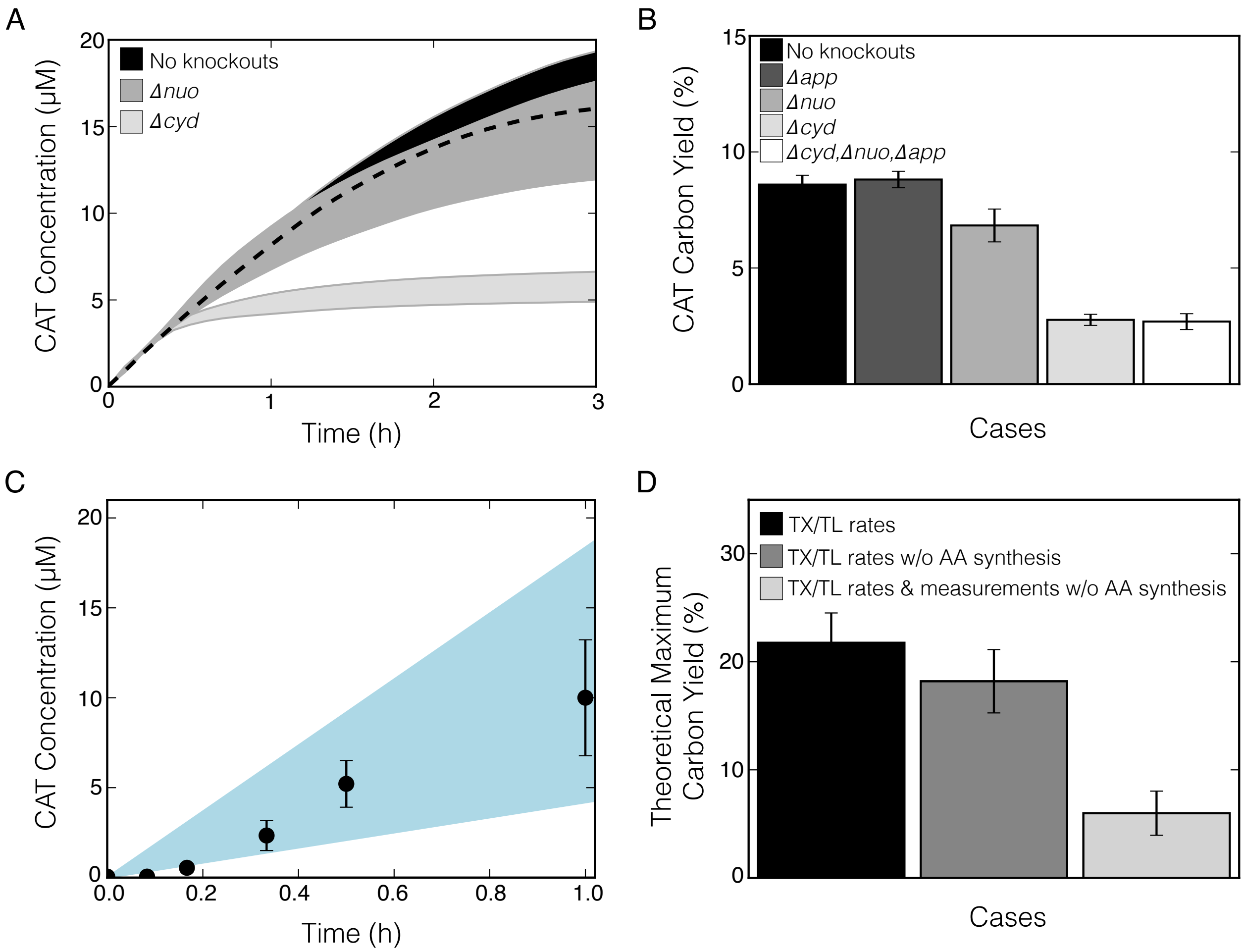

Fig. 6: The effects of oxidative phosphorylation and amino acid synthesis pathways on CAT production and carbon yield. A. 95% confidence interval of the ensemble of kinetic models for CAT concentration versus time, for the best-fit set with no knockouts (black shaded region and dashed line), *nuo* knockout (medium gray), and *cyd* knockout (light gray). B. CAT carbon yield of the ensemble of kinetic models for no knockouts (black), *app* knockout (dark gray), *nuo* knockout (medium gray), *cyd* knockout (light gray), and all three knockouts (white). Error bars represent standard deviation of the ensemble. C. 95% confidence interval of the ensemble of ssFBA simulations (light blue region) of CAT concentration over time, against experimental data (black). D. Theoretical maximum carbon yield of CAT production, calcualted by ssFBA for three different cases: constrained by transcription/translation (TX/TL) rates (black), same as previous but without amino acid synthesis reactions (medium gray), and same as previous but constrained by experimental measurements where available (light gray). Error bars represent standard deviation of the ensemble.

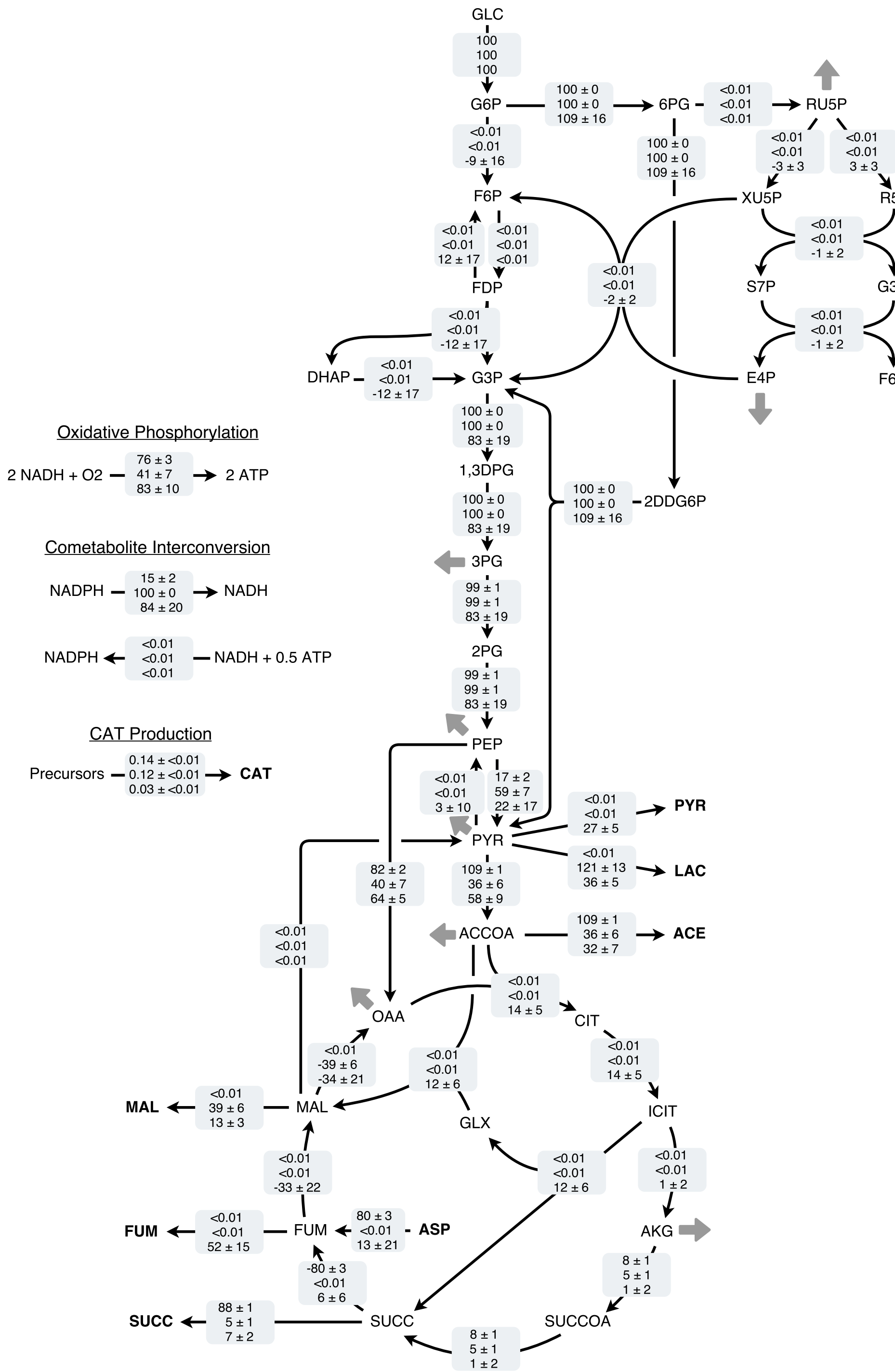

Fig. 7: Flux profile for glycolysis, pentose phosphate pathway, Entner-Doudoroff pathway, TCA cycle, NADPH/NADH transfer, and oxidative phosphorylation. Sequence-specific FBA flux value (mean ± standard deviation) across ensemble for 1 hr, normalized to glucose uptake flux. Flux distribution for three different cases: constrained by transcription and translation rates (top row), same as previous but without amino acid synthesis reactions (second row), and same as previous but constrained by experimental measurements where available (bottom row).

Table 1: CAT carbon yield breakdown for best-fit set, knockouts, and experimental data. Carbon produced as CAT, carbon consumed as glucose and each amino acid, sum of consumed species, and yield.

Carbon Produced (C-mM)	Best-fit	Δapp	Δnuo	Δcyd	Δapp Δnuo Δcyd	Data
CAT	20.9	21.4	18.1	6.5	5.1	21.6
Carbon Consumed (C-mM)						
GLC	215.4	215.4	215.4	215.4	159.8	215.4
ALA	0.0	0.0	1.7	0.0	0.0	0.0
ASN	6.2	6.2	6.2	6.3	6.3	6.3
ASP	7.5	7.5	3.9	0.0	0.0	9.6
CYS	3.0	3.1	3.0	2.9	2.9	3.7
GLN	0.0	0.0	0.0	1.8	2.7	0.0
GLY	3.1	3.1	2.6	1.1	0.9	1.5
HIS	0.2	0.2	1.1	0.4	0.3	0.0
ILE	1.0	1.0	0.8	0.3	0.2	1.7
LEU	1.4	1.4	1.2	0.4	0.3	2.0
LYS	10.7	10.7	13.1	13.2	13.2	8.3
MET	0.8	0.8	0.7	0.2	0.2	2.9
PHE	3.2	3.3	2.8	1.0	0.8	1.6
PRO	2.4	2.4	0.7	0.2	0.2	1.9
SER	2.5	2.5	2.4	2.1	2.1	1.8
THR	3.4	3.4	3.3	2.9	2.8	2.8
TRP	1.0	1.0	0.8	0.3	0.2	1.2
TYR	1.1	1.1	1.1	0.4	0.4	0.6
VAL	1.4	1.5	1.2	0.4	0.4	2.4
Sum	264.3	264.6	262.0	249.3	193.7	263.7
Yield	7.9%	8.1%	6.9%	2.6%	2.7%	8.2%

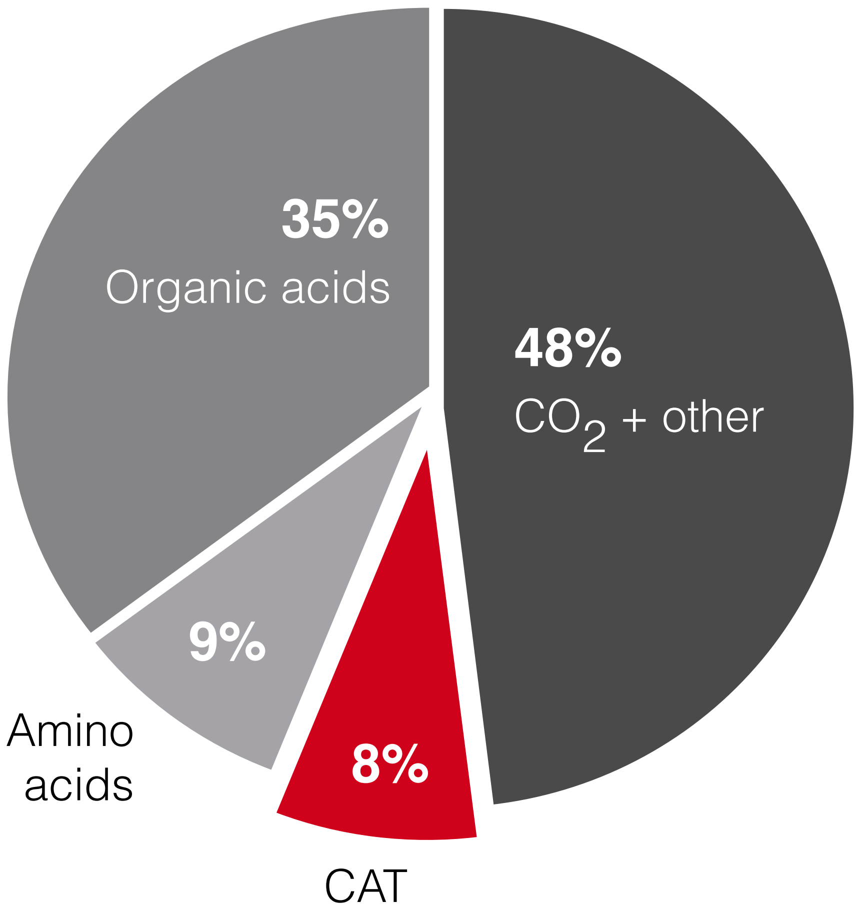

Fig. 8: Carbon balance for the best-fit set. Carbon moles produced as CAT, amino acids (alanine and glutamine), organic acids (lactate, acetate, succinate, and malate), and other byproducts including carbon dioxide, as percentages of total carbon consumption (glucose and all other amino acids).