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 23% of the maximum theoretical yield, calculated using sequence-specific flux balance analysis. This suggests that CAT production could be further optimized. The dynamic modeling approach predicted that substrate consumption of glucose and pyruvate 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 23% of the maximum theoretical yield, but 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 iAF1260 reconstruction [16], and by adding reactions describing chloramphenicol acetyltransferase (CAT) biosynthesis, a model protein for which we have a comprehensive training dataset [23]. 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. Allosteric control was important to the dynamics of the organic acid intermediates and CAT biosynthesis (Fig. 1, bottom). The acetate, lactate, pyruvate, succinate, malate and CAT trajectories were qualitatively different in the absence of allosteric control following glucose exhaustion. In particular, the rate of CAT biosynthesis and lactate accumulation and subsequent consumption decreased following glucose exhaustion in the absence of control.

We used sequence-specific flux balance analysis (ssFBA) to calculate the theoretical maximum CAT carbon yield for different constraint values (Fig. 6). The experimental CAT carbon yield was 0.0821, while the best-fit parameter set had a carbon yield of 0.086 $\pm$ 0.004. Thus, although the kinetic model described the experimental data including the yield, it was unclear whether the performance of the CFPS system was optimal. To address this question, we used ssFBA in combination with the cell-free metabolic network and a T7 promoter model to estimate the maximum theoretical CAT carbon yield. Toward this, we first validated the ssFBA approach by comparing the simulated versus measured concentrations of CAT over the first hour of the CFPS reaction (Fig. 6A). We sampled different RNA polymerase/ribosome levels and elongation rates in the physiological range to establish the uncertainty in the ssFBA simulation. The ssFBA estimate of the CAT abundance was consistent with the measured values. Next, we calculate the CAT carbon yield for three classes of constraints: (i) theoretical max glucose, amino acid and oxygen upper bounds, and no transcriptional/translational constraints; (ii) theoretical maximum glucose, amino acid and oxygen upper bounds, and realistic transcriptional/translational constraints and (iii) metabolite values constrained by the data, and realistic transcriptional/translational constraints. The unconstrained theoretical maximum CAT carbon yield was 0.363 $\pm$ 0.02 (Fig. 6B, left); this case had no upper bound on the transcription and translation reactions, and was only constrained by glucose, oxygen and amino acid consumption rates. On the other hand, for realistic constraints on transcription and translation, the CAT carbon yield was 0.226 $\pm$ 0.03 (Fig. 6B, middle). Lastly, when using realistic metabolite and transcription and translation constraints the predicted carbon yield was 0.062 $\pm$ 0.02, similar to the experimental yield Fig. 6B, end). Thus, the experimental dataset and best-fit parameter set each produced CAT at 23% of the theoretical maximum and 36% of a theoretical physiological case.

To better understand which parameters and parameter combinations influenced model performance we performed sensitivity analysis (Fig. 5). CAT production was most sensitive to the CAT synthesis reaction, oxidative phosphorylation activity, and alanine synthesis (Fig. 5, top, section A). The 16 next most important reactions to CAT production (section B) came from various pathways across the network: four each from glycolysis, the TCA cycle, and amino acid synthesis/degradation; two from the pentose phosphate pathway; and one each from the Entner-Doudoroff pathway and the energy species reactions. The pairwise sensitivities (off-diagonal elements) were different from the corresponding first-order sensitivities (diagonal elements), and led to interesting outcomes. For example, glutamine synthesis and arginine degradation were both among the most important reactions to CAT production (they rank 5th and 10th, respectively). This was likely because they both affect 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. The system state as a whole was most sensitive to glucose uptake via GTP and the forward reaction of lactate dehydrogenase (Fig. 5, bottom, section F). The 30 next most important reactions to the system state (section G) came from various pathways across the network: eight from amino acid synthesis/degradation; six from glycolysis; four from the TCA cycle; and two each from the pentose phosphate pathway, Entner-Doudoroff, energy/reducing pathways, and small molecule transport; one from oxidative phosphorylation; and one from pyrophosphatase consumption. The system state had more pairwise sensitivities that differ from the corresponding first-order sensitivities and stand out as significant. For example, the first-order effect of alanine synthesis was large; it consumes both pyruvate and glutamate, two key species in the network. In addition, a handful of alanine synthesis pairwise sensitivities were also large. However, there were enough reactions that, when paired with alanine synthesis, had little effect on the model; malic enzyme is one of these, as it produces the pyruvate that alanine synthesis consumes. Thus, the total-order alanine synthesis sensitivity was low, placing it at the very bottom of section I. Another interesting result was the intersection of sections F and G with section J. The 53 reactions in section J were turned off in the best-fit set ($V^{max} = 0$); therefore, the perturbation of these reactions had no effect on the model. Thus, all pairwise sensitivities with reactions in section J were pseudo first-order sensitivities for the other reactions. Interestingly, many reactions in section F and several in section G showed their highest sensitivities when paired with the ”non-effects” of section J. Of these, three involved pyruvate, strengthening its role as a key metabolite; the others were glucose consumption via GTP/CTP-specific hexokinases, fumarate reductase, and SO_4 utilization. This suggested that these reactions' effects on the model were canceled out or lessened by most other reactions, but were of course not affected by the reactions in section J. This was also likely the reason that reactions in section J rank above those in section K, despite having no effect themselves on the model. Taken together, sensitivity analysis identified blocks of parameters that either individually, or in combination influenced model performance. However, the sensitivity analysis did not establish what the maximum performance of the system was. To answer that question we performed sequence-specific flux balance analysis of CAT production.

In comparing the flux distributions between the unconstrained and constrained cases (Fig. 7), the constrained cases heavily utilized the first step in the pentose phosphate pathway to generate NADPH. In these cases the majority of the flux continued through the Entner–Doudoroff pathway, whereas in the unconstrained case the majority of flux traveled through glycolysis. In all cases, the energy source came primarily from oxidative phosphorylation, as well as partly from the TCA cycle. In the transcription/translation constrained case, there was a high flux through fumerate dehydrogenase from aspartic acid uptake, whereas in the unconstrained and most constrained cases, acetate and lactate accumulation occurred. This shows that the system is producing NADH through lactate dehydrogenase as well as through pyridine nucleotide transhydrogenase (*pntAB*) to supply enough NADH for oxidative phosphorylation. As a result, high oxidative phosphorylation activity relative to our cell-free system leads to an acetate overflow. This suggests that there is potential for increasing CAT production by reducing this diversion of carbon. To simulate potential knockouts, we constrained the specific glucose and amino acid uptake rates to the same values as simulated with no knockouts. In an ssFBA simulation with constrained transcription/translation rates, knocking out the *gnd* reaction decreases flux of acetate production but increases flux through *pntAB*, which is responsible for regenerating NADPH. The simulation showed carbon was diverted toward lactate; however, since CAT production is constrained by the translation rate, we expected no increase in CAT production. The decrease in acetate production is promising as a mechanism to increase CAT yield. A second simulation with a knockout of *gnd* and phosphate acetyltransferase showed carbon being diverted toward lactate and succinate; however, it required a higher flux through oxidative phosphorylation and the TCA cycle to meet the energetic needs of the system.

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. We showed that the model produces CAT at 23% of the theoretical maximum in terms of carbon yield, and at 36% of a physiological case in which transcription and translation are constrained. The theoretical maximum and transcription/translation constrained case were obtained using FBA, which predicted a different flux distribution for each case. The unconstrained case predicted most of the carbon flux going through glycolysis, while the constrained cases predicted significant flux through the Entner-Doudoroff pathway and some through TCA cycle. However, all cases relied on oxidative phosphorylation for the system's energetic needs as well as for tRNA charging for CAT synthesis. Sensitivity analysis of the dynamic model suggested that both CAT production and the entire metabolic network were most sensitive to amino acid synthesis and degradation reactions, and reactions in glycolysis and the TCA cycle. CAT production was also very sensitive to the CAT synthesis reaction, unsurprisingly. The allosteric control component of the hybrid modeling approach was shown as important to central carbon metabolism, but not very important to CAT production. Taken together, this is 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. The most sensitive parameter for the model as a whole is the uptake of glucose, followed by pyruvate. 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 that excercise considerable sensitivity to model performance. Oxidative phosphorylation activity is vital, since it provides most of the energetic needs of CFPS. 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%, consistent with literature observations [25]. Jewett and coworkers also saw a decrease in CAT yield, ranging from 1.5-fold to 4-fold, when knocking out oxidative phosphorylation. Other reactions with considerable sensitivity are those responsible for the consumption of glucose and the consumption and production of pyruvate, which is consumed during the second phase of CFPS. Also important is the first step in pentose phosphate pathway, as it converts G6P to 6PG and produces NADPH. The next most sensitive reaction is the conversion of NADPH to NADH, which helps to supply NADH to oxidative phosphorylation reactions to meet the energetic needs of the system. Many of the insignificant reactions lead to the accumulation of unwanted byproducts such as ethanol and formate.

In analyzing the model parameters' effect on CAT production, CAT synthesis is the most important, followed by oxidative phosphorylation and the glucose and pyruvate consuming reactions, as well as the downstream reactions which are necessary to drive CFPS. In addition, the conversion of ATP to GTP shows significance since it facilitates CAT synthesis. Thus, supplementation with additional GTP may improve the efficiency of CAT production. A similar theme is seen in the sensitivity of overall model performace; the most important reactions are glucose and pyruvate consuming reactions and downstream reactions. The overall sensitivity results show that the substrates driving CFPS are vital to system performance and CAT production. 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, but with the exception of acetate these are all consumed in the second phase. CFPS operating under glucose consumption has higher CAT production than that of pyruvate/lactate consumption, as can be seen in the experimental data (~10 μM/h vs ~5 μM/h).

Comparing the theoretical maximum 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. The theoretical maximum CAT yield was 36% for an unconstrained case and 23% for a transcription/translation constrained case. Knockouts of *gnd* and phosphate acetyltransferase show that carbon can be diverted away from acetate and potentially 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]. In addition to improving CFPS performance, Jewett and coworkers have shown that oxidative phosphorylation operates in cell-free systems, and that knocking out these reactions is detrimental to protein yield [25]. However, it is unknown how active oxidative phosphorylation is compared to that of *in vivo* systems, and both of the modeling approaches we present suggest its importance to CAT production. Thus, oxidative phosphorylation is a potential area for improvement of CFPS performance and protein yield.

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 generated 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:

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

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 50 μM, this corresponds to catalytic rate bounds of 0 and 2778 s^{-1}. These bounds resulted in a median catalytic rate of 0.016 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 min. 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 than the previous set, and some with higher cost, were added to the ensemble. After generating 12,437 sets, we selected 100 sets with minimal set to set correlation to avoid over-sampling any region of parameter space. The original 12,437-set ensemble had a mean Pearson correlation coefficient of 0.94 between pairs of sets, while the 100-set ensemble had a mean Pearson correlation coefficient of 0.83 between pairs of sets.

Sensitivity analysis of the 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:

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

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.

Likewise, we determined the reactions 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:

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

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.

Sequence-specific FBA and calculation of CAT yield The yield on CAT production was calculated for each case as a ratio of carbon produced as CAT to carbon consumed as reactants (glucose and amino acids):

$$Yield = \frac{\Delta m_{CAT}\ C_{CAT}}{\sum_{i=1}^{\mathcal{R}} \max(\Delta m_i, 0)\ C_{m_i}} \tag{8}$$

where Δm_{CAT} denotes the amount 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. Because no data was available for arginine or glutamate, these reactants were left out of all yield calculations. Yield of the best-fit parameter set and the experimental data were calculated by setting ΔCAT equal to the final minus the initial CAT concentration and setting Δm_i equal to the initial minus the final reactant concentration. Theoretical maximum CAT carbon yields for three cases discussed below were calculated using flux balance anaylsis (FBA) with a sequence-specific description of CAT synthesis, where Δm_i denotes the flux of the i^{th} species. This sequence-specific FBA [44] problem was formulated as:

$$\begin{gathered} \max_{\boldsymbol{w}} \left(w_{obj} = \boldsymbol{\theta}^T \boldsymbol{w}\right) \\ \text{Subject to : } \mathbf{Sw} = \mathbf{0} \\ \alpha_i \leq w_i \leq \beta_i \qquad i = 1, 2, \ldots, \mathcal{R} \end{gathered} \tag{9}$$

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 objective w_{obj} was to maximize the specific rate of CAT formation. The specific glucose uptake rate was constrained to allow a maximum flux of 40 mM/h according to experimental data; the specific amino acid uptake rates were bound to allow a maximum flux of 30 mM/h, but did not reach this maximum flux. The transcription and translation template reactions come from sequence-specific analysis [44], and include transcription initiation, transcription, mRNA degradation, translation initiation, translation, and tRNA charging. The flux balance analysis problem was solved using the GNU Linear Programming Kit (v4.52) [45]. The solution flux vector was used to calculate the carbon yield of CAT for the three FBA cases. Glucose, oxygen, and amino acids were modeled as being imported into the system, while CAT synthesis and metabolite byproduct formation were modeled as export from the system. The rest of the network followed a pseudo steady-state assumption where metabolites were not allowed to accumulate; thus, the network could be solved by linear programming instead of solving differential equations.

The transcription rate was formulated as:

$$w_{TX} = RNAP \left(\frac{v_{RNAP}}{l_{mRNA}}\right) \left(\frac{Gene}{k_m + Gene}\right) P \tag{10}$$

where w_{TX} denotes the transcription rate, $RNAP$ denotes the concentration of RNA polymerase, v_{RNAP} denotes the elongation rate by the RNA polymerase in nucleotides per hour, l_{mRNA} denotes the mRNA length in nucleotides, $Gene$ denotes the gene concentration, k_m denotes the plasmid saturation coefficient, and P denotes the promoter activity. The mRNA and protein sequence of CAT was determined from literature. The promoter activity was formulated following Moon et al. for synthetic circuits as:

$$P = \frac{K_1}{1 + K_1} \tag{11}$$

where K_1 denotes the state of T7 RNA polymerase binding. The translation rate was formulated as:

$$w_{TL} = K_P \; Ribo \left(\frac{v_{Ribo}}{l_{Protein}}\right) mRNA_{SS} \tag{12}$$

where K_P denotes the polysome amplification constant, $Ribo$ denotes the ribosome concentration, v_{Ribo} denotes the elongation rate of the ribosome in amino acids per hour, $l_{Protein}$ denotes the number of amino acids in the protein of interest, and $mRNA_{SS}$ denotes the mRNA concentration at steady state, equal to the transcription rate divided by the degradation rate of mRNA.

An ensemble of 100 sets of flux distributions was calculated for three different cases: unconstrained, constrained by transcription/translation rates, and constrained by transcription/translation rates and experimental measurements. For the unconstrained case, all rates were left unbounded, except the specific glucose uptake 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. For the case constrained by transcription/translation rates, an ensemble was generated by randomly sampling RNAP 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 [27, 46]. For the case constrained by transcription/translation rates and experimental measurements, 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.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

J.V and A.Y directed the study. R.T, H.J and J.C conducted the cell culture measurements. J.V and W.D developed the reduced order HL-60 models and the parameter ensemble. W.D analyzed the model ensemble, and generated figures for the manuscript. The manuscript was prepared and edited for publication by W.D, A.Y 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 and by an award from the US Army and Systems Biology of Trauma Induced Coagulopathy (W911NF-10-1-0376) to J.V for the support of M.V.

References

1. Jewett MC, Calhoun KA, Voloshin A, Wuu JJ, Swartz JR (2008) An integrated cell-free metabolic platform for protein production and synthetic biology. Mol Syst Biol 4: 220.
2. Matthaei JH, Nirenberg MW (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 MW, Matthaei JH (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 JP, Swartz JR (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 CE, Jewett MC (2012) Cell-free synthetic biology: thinking outside the cell. Metab Eng 14: 261-9.
6. Pardee K, Slomovic S, Nguyen PQ, Lee JW, Donghia N, et al. (2016) Portable, on-demand biomolecular manufacturing. Cell 167: 248-59.e12.
7. Fredrickson AG (1976) Formulation of structured growth models. Biotechnol Bioeng 18: 1481-6.
8. Domach MM, Leung SK, Cahn RE, Cocks GG, Shuler ML (1984) Computer model for glucose-limited growth of a single cell of escherichia coli b/r-a. Biotechnol Bioeng 26: 203-16.
9. Steinmeyer D, Shuler M (1989) Structured model for Saccharomyces cerevisiae. Chem Eng Sci 44: 2017-30.
10. Wu P, Ray NG, Shuler ML (1992) A single-cell model for cho cells. Ann N Y Acad Sci 665: 152-87.
11. Castellanos M, Wilson DB, Shuler ML (2004) A modular minimal cell model: purine

and pyrimidine transport and metabolism. Proc Natl Acad Sci U S A 101: 6681-6.

12. Atlas JC, Nikolaev EV, Browning ST, Shuler ML (2008) Incorporating genome-wide dna sequence information into a dynamic whole-cell model of escherichia coli: application to dna replication. IET Syst Biol 2: 369-82.

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

14. Edwards JS, Palsson BØ (2000) The escherichia coli mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci U S A 97: 5528-33.

15. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7: 129-43.

16. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (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.

17. Oh YK, Palsson BØ, Park SM, Schilling CH, 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.

18. Ibarra RU, Edwards JS, Palsson BØ (2002) Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420: 186-9.

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

20. Hyduke DR, Lewis NE, Palsson BØ (2013) Analysis of omics data with genome-scale models of metabolism. Mol Biosyst 9: 167-74.

21. McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of escherichia coli. Mol Syst Biol 9: 661.

22. Zomorrodi AR, Suthers PF, Ranganathan S, Maranas CD (2012) Mathematical optimization applications in metabolic networks. Metab Eng 14: 672-86.

23. Calhoun KA, Swartz JR (2005) An economical method for cell-free protein synthesis using glucose and nucleoside monophosphates. Biotechnology Progress 21: 1146-53.

24. Wayman JA, Sagar A, Varner JD (2015) Dynamic modeling of cell-free biochemical networks using effective kinetic models. Processes 3: 138.

25. Jewett MC, Calhoun KA, Voloshin A, Wuu JJ, Swartz JR (2008) An integrated cell-free metabolic platform for protein production and synthetic biology. Molecular Systems Biology 4.

26. Li J, Gu L, Aach J, Church GM (2014) Improved cell-free rna and protein synthesis system. PLoS ONE 9: 1-11.

27. Underwood KA, Swartz JR, Puglisi JD (2005) Quantitative polysome analysis identifies limitations in bacterial cell-free protein synthesis. Biotechnology and Bioengineering 91: 425-35.

28. Kotte O, Zaugg JB, Heinemann M (2010) Bacterial adaptation through distributed sensing of metabolic fluxes. Mol Syst Biol 6: 355.

29. Cabrera R, Baez M, Pereira HM, Caniuguir A, Garratt RC, et al. (2011) 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 286: 5774-83.

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

31. Ogawa T, Murakami K, Mori H, Ishii N, Tomita M, et al. (2007) Role of phosphoenolpyruvate in the NADP-isocitrate dehydrogenase and isocitrate lyase reaction in

Escherichia coli. J Bacteriol 189: 1176-8.

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

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

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

35. Hines JK, Fromm HJ, Honzatko RB (2007) 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 282: 11696-704.

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

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

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

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

40. Kale S, Arjunan P, Furey W, Jordan F (2007) 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 282: 28106-16.

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

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

43. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, et al. (2013) Ecocyc: fusing model organism databases with systems biology. Nucleic Acids Res 41: 605-12.

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

45. (2016). GNU Linear Programming Kit, Version 4.52. URL http://www.gnu.org/software/glpk/glpk.html.

46. Garamella J, Marshall R, Rustad M, 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.

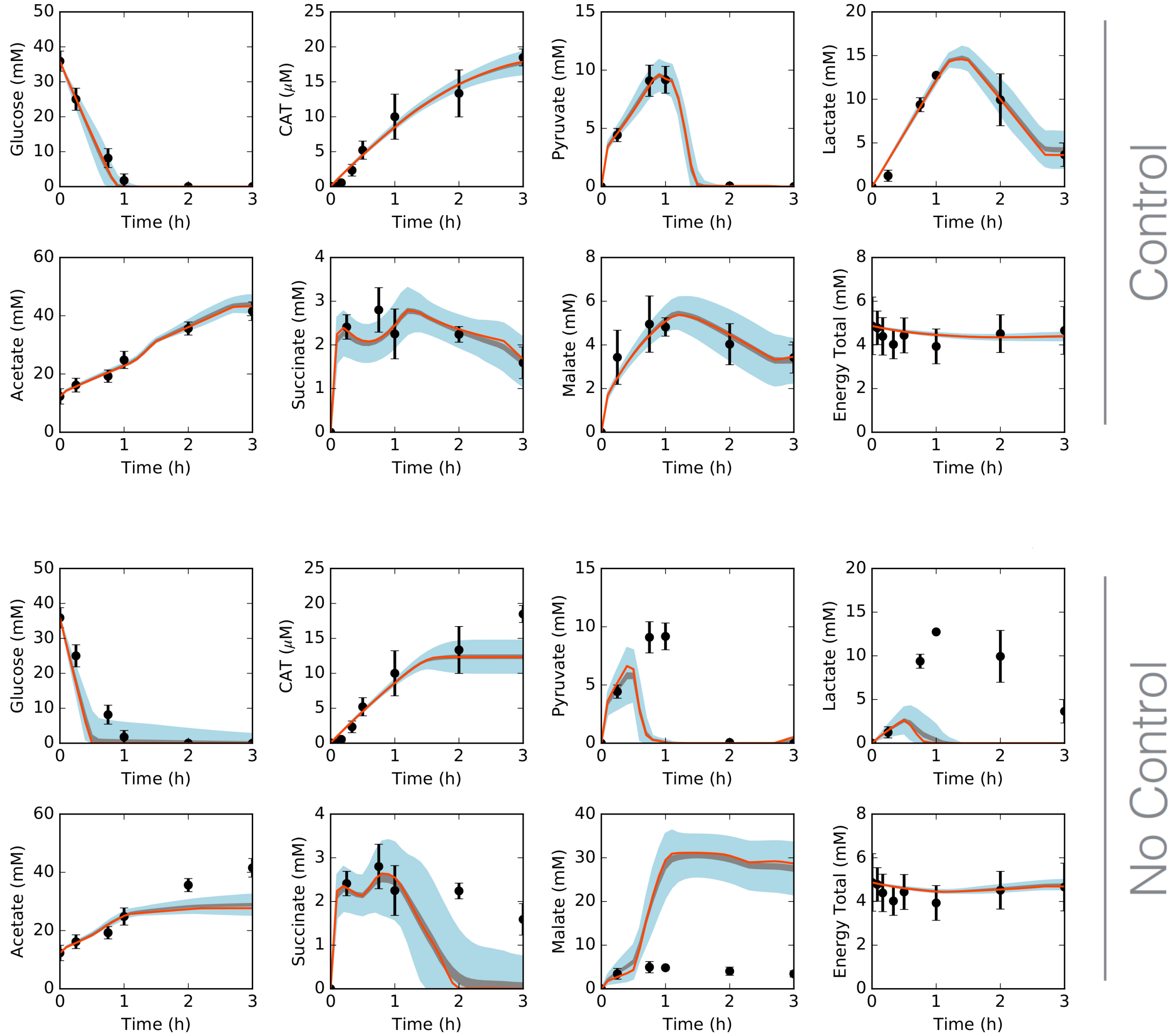

Fig. 1: Central carbon metabolism in the presence (top) and absence (bottom) of allosteric control, including glucose (substrate), CAT (product), and intermediates, as well as total concentration of energy species. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence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

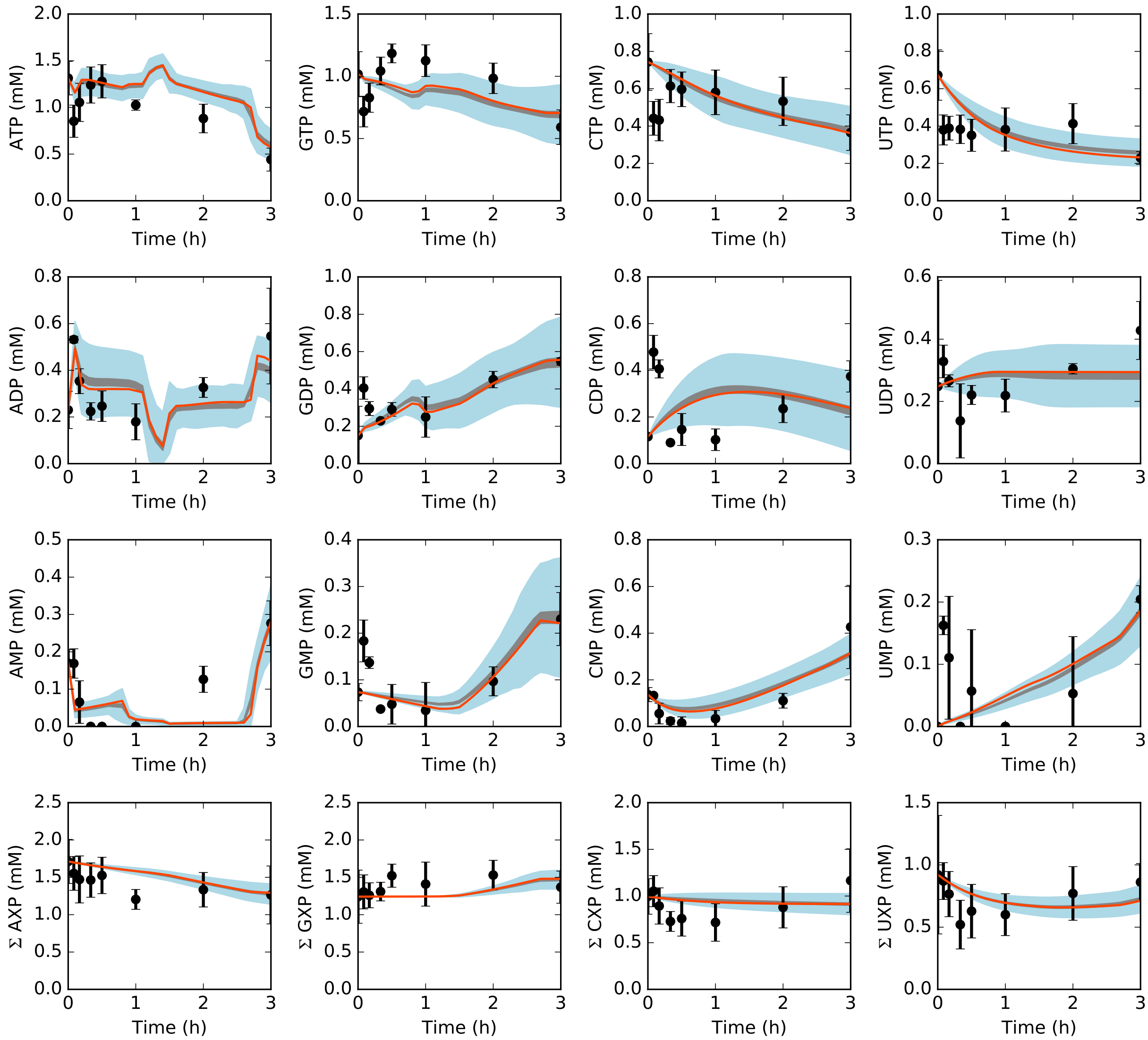

Fig. 2: Energy species and energy totals by base in the presence of allosteric control. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence interval (blue shaded region) and 95% confidence interval of the mean (gray shaded region) over the ensemble of 100 sets.

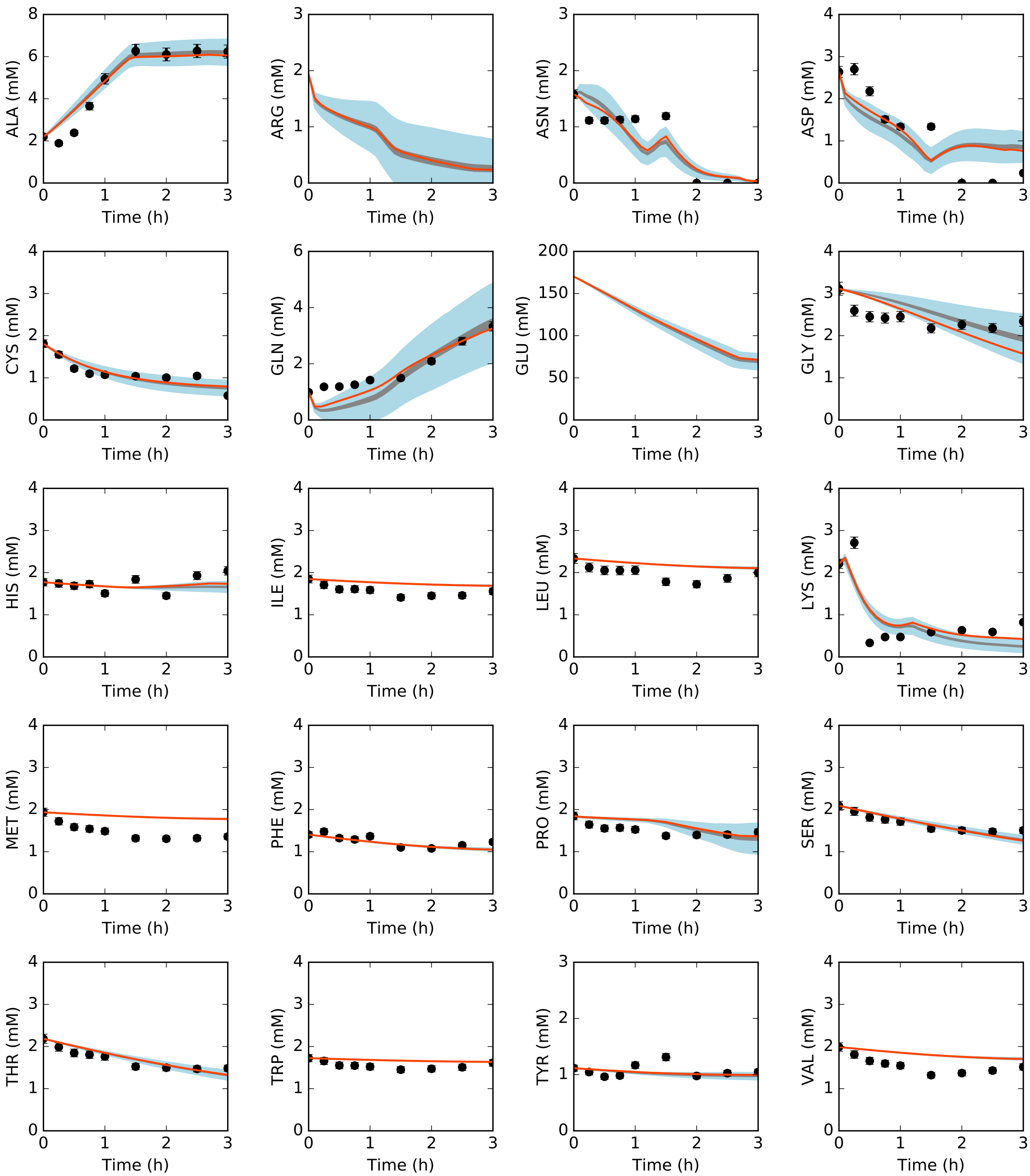

Fig. 3: Amino acids in the presence of allosteric control. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence interval (blue shaded region) and 95% confidence interval of the mean (gray shaded region) over the ensemble of 100 sets.

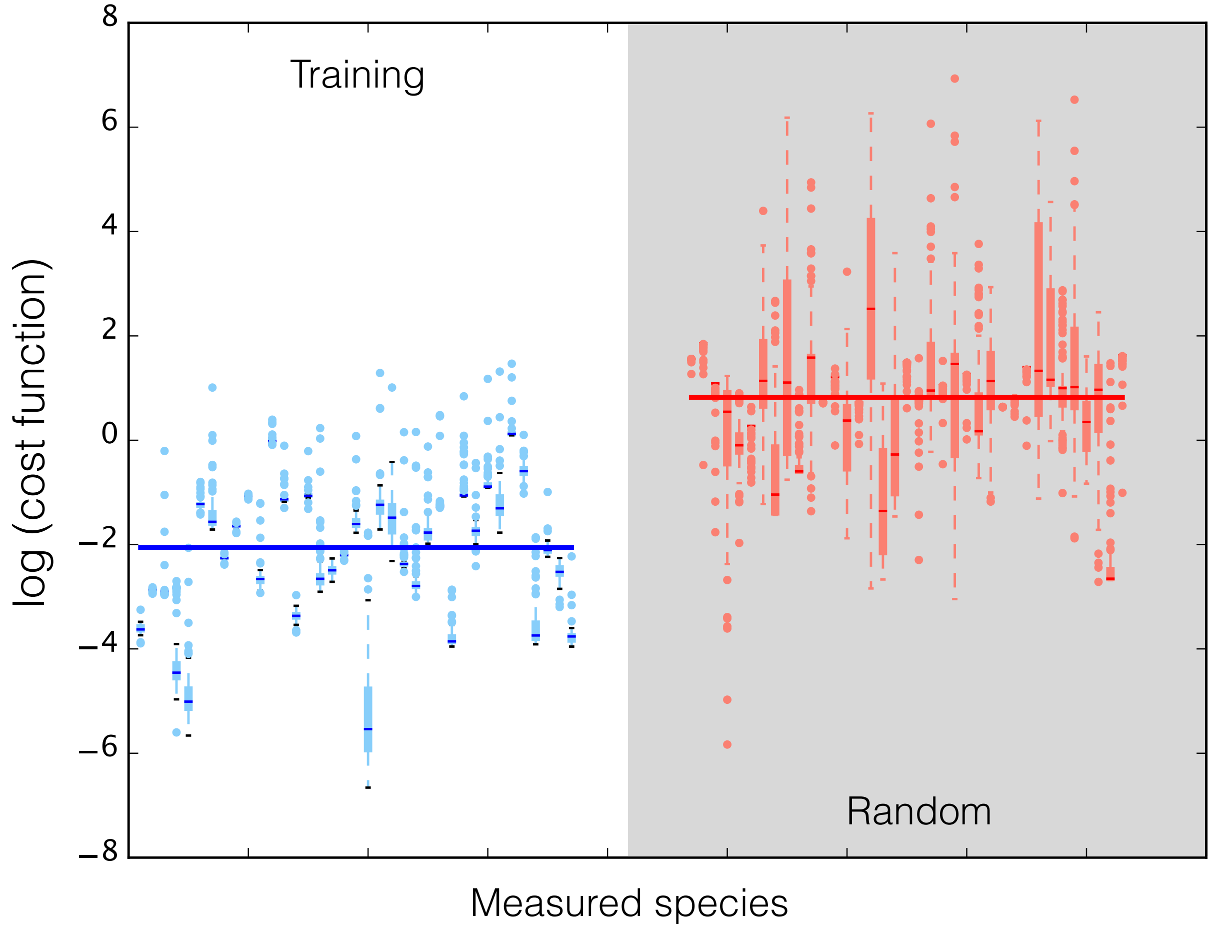

Fig. 4: Log of cost function across 37 datasets for data-trained ensemble (blue) and randomly generated ensemble (red, gray background). Median (bars), interquartile range (boxes), range excluding outliers (dashed lines), and outliers (circles) for each dataset. Median across all datasets (large bar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

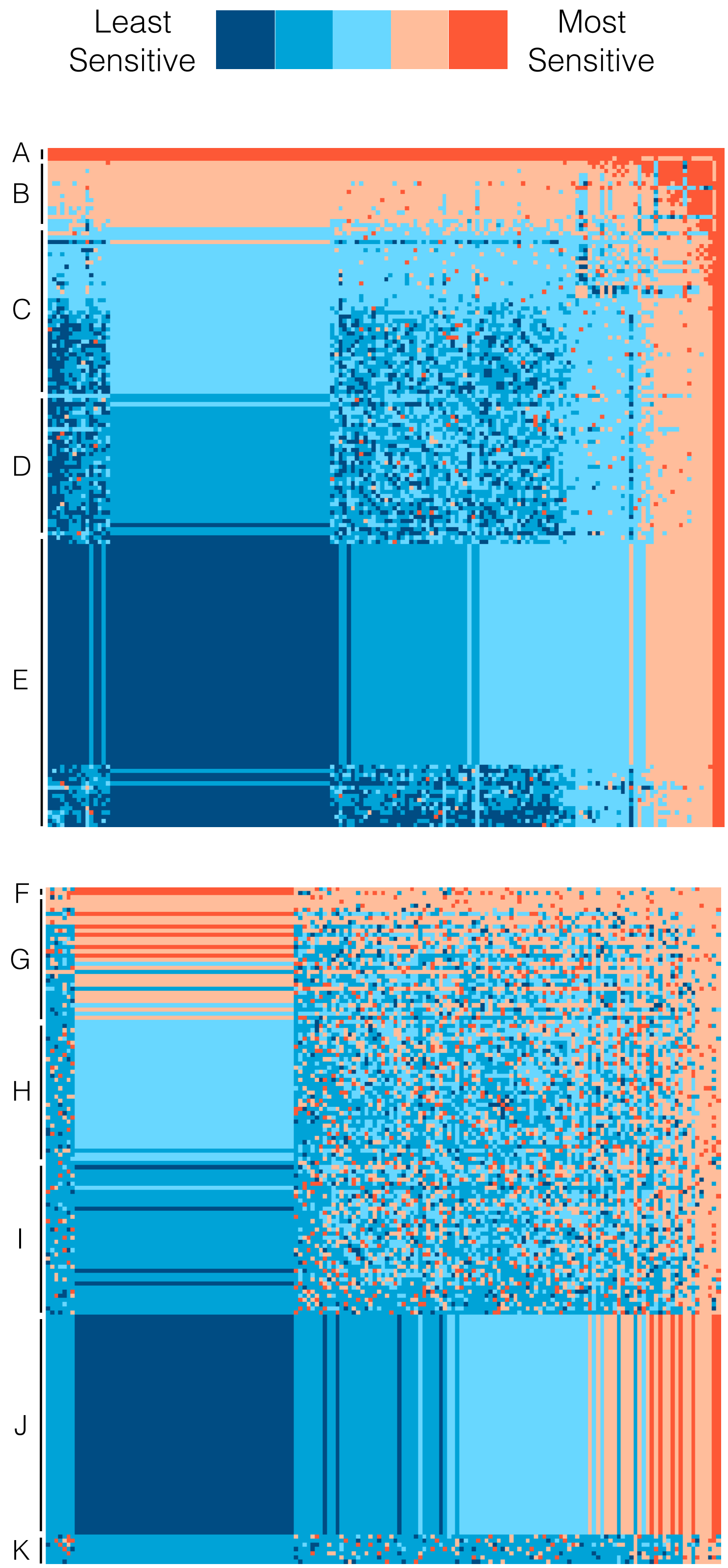

Fig. 5: Normalized first-order and pairwise sensitivities of CAT production (top) and system state (bottom) to maximum reaction rates.

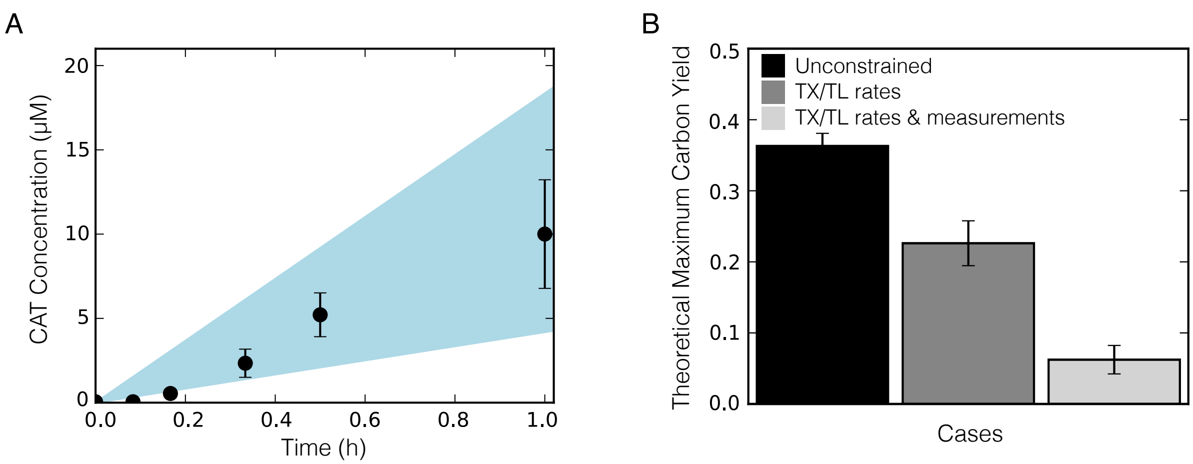

Fig. 6: Sequence-specific flux balance analysis of CAT production and yield. A. 95% confidence interval of the ensemble (light blue region) for CAT concentration versus time. B. Theoretical maximum carbon yield of CAT calcualted by ssFBA for three different cases: unconstrained except for glucose uptake (black), constrained by transcription/translation (TX/TL) rates (grey), and constrained by transcription/translation (TX/TL) rates and experimental measurements where available (light grey). Error bars represent standard deviation of the ensemble.

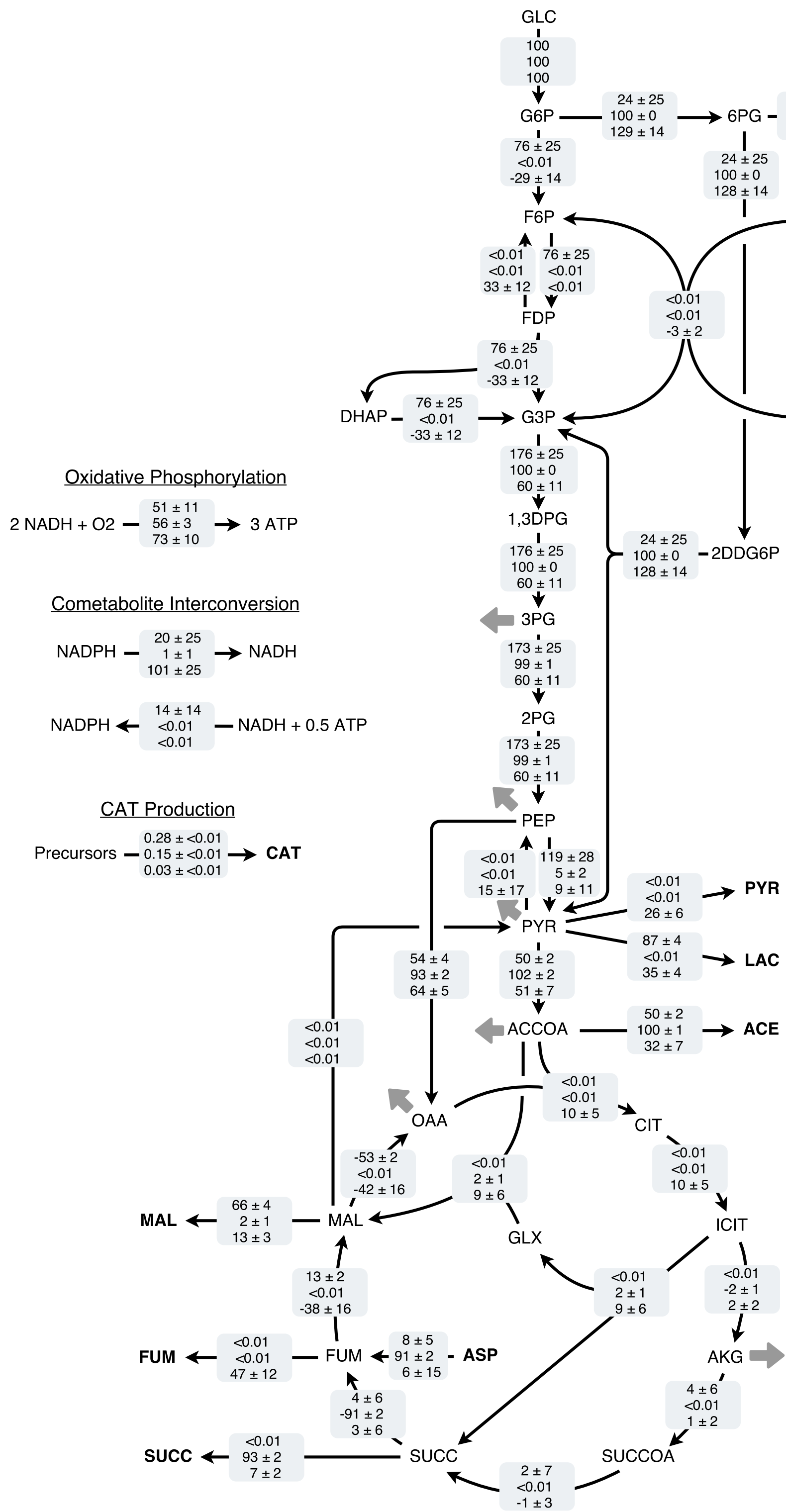

Fig. 7: Flux profile for glycolysis, pentose phosphate pathway, Entner-Doudoroff pathway, TCA cycle, NADPH/NADH transfer, and oxidative phosphorylation. Sequence-specific FBA flux value (mean ± standard deviation) across ensemble for 1 hr, normalized to glucose uptake flux. Flux distribution for three different cases: unconstrained except for glucose uptake (top row), constrained by transcription and translation rates (second row), and constrained by transcription, translation rates and experimental measurements where available (bottom row).