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. 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 observed CAT yield was 45% of the maximum theoretical yield, suggesting CAT production could be further optimized. The ensemble of dynamic models predicted the majority of carbon flux was routed through glycolysis and the TCA cycle, while flux balance analysis predicted significant flux through the Entner-Doudoroff pathway. Local and global sensitivity analysis suggested CAT production was most sensitive to parameters and initial conditions directly associated with CAT synthesis, as well as GTP/GMP synthesis, amino acid synthesis, and to a lesser extent amino acid initial conditions. On the other hand, CAT production was robust to allosteric control parameters and the initial conditions of glucose and oxygen. Taken together, we presented the first dynamic model of *E. coli* cell-free protein synthesis. This study provides 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 advantages is direct access to metabolites and the microbial biosynthetic machinery without the interference of a cell wall. This allows us to control as well as interrogate the chemical environment while the biosynthetic machinery is operating, potentially at a fine time resolution. Second, cell-free systems also allow us to study biological processes without the complications associated with cell growth. 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]. Interestingly, many of the challenges confronting genome-scale kinetic modeling can potentially be overcome in a cell-free system. For example, there is no complex transcriptional regulation to consider, transient metabolic measurements are easier to obtain, and we no longer have to consider cell growth. Thus, cell-free operation holds several significant advantages for model development, identification and validation. Theoretically, genome-scale cell-free kinetic models may be possible for industrially important organisms, such as *E. coli* or *B. subtilis*, if a simple, tractable framework for integrating allosteric regulation with enzyme kinetics can be formulated.

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 [6]. 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 [7]. Shuler and coworkers generated many single cell kinetic models, including single cell models of eukaryotes [8, 9], minimal cell architectures [10], as well as DNA sequence based whole-cell models of *E. coli* [11]. Conversely, highly abstracted kinetic frameworks, such as the cybernetic framework, represented a paradigm shift, viewing cells as growth-optimizing strategists [12]. Cybernetic models have been highly successful at predicting metabolic choice behavior, e.g., diauxie behavior [13], steady-state multiplicity [14], as well as the cellular response to metabolic engineering modifications [15]. Unfortunately, traditional, fully structured cybernetic models also suffer from an identifiability challenge, as both the kinetic parameters and an abstracted model of cellular objectives must be estimated simultaneously. However, recent cybernetic formulations from Ramkrishna and colleagues have successfully treated this identifiability challenge through elementary mode reduction [16, 17].

In the post genomics world, large-scale stoichiometric reconstructions of microbial metabolism popularized by static, constraint-based modeling techniques such as flux balance analysis (FBA) have become standard tools [18]. Since the first genome-scale stoichiometric model of *E. coli*, developed by Edwards and Palsson [19], well over 100 organisms, including industrially important prokaryotes such as *E. coli* [20] or *B. subtilis* [21], are now available [22]. 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 [23], although other metabolic objectives have also been estimated [24]. Recent advances in constraint-based modeling have overcome the early shortcomings of

the platform, including capturing metabolic regulation and control [25]. 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 [26, 27]. 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 *E. coli* cell-free protein synthesis (CFPS) models using the hybrid cell-free modeling approach of Wayman et al [REFHERE]. Model parameters were estimated from measurements of glucose, organic acids, energy species, amino acids and the protein product, chloramphenicol acetyltransferase (CAT). The ensemble described the training data, with the exception of some of the amino acid dynamics. 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 observed CAT yield was 45% of the maximum theoretical yield, suggesting CAT production could be further optimized. The metabolic flux distribution predicted by the dynamic model and flux balance analysis were significantly different. The ensemble of dynamic models predicted the majority of carbon flux was routed through glycolysis and the TCA cycle, while flux balance analysis predicted significant flux through the Entner-Doudoroff pathway. Local and global sensitivity analysis suggested CAT production was most sensitive to parameters and initial conditions directly associated with CAT synthesis, as well as GTP/GMP synthesis, amino acid synthesis, and to a lesser extent amino acid initial conditions. On the other hand, CAT production was robust to allosteric control parameters and the initial conditions of glucose and oxygen. Taken together, we presented the first dynamic model of *E. coli* cell-free protein synthesis. We 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

Estimation of an ensemble of cell-free protein synthesis models. We used the hybrid cell-free modeling framework of Wayman et al. to simulate the production of a model protein [REFHERE]. The cell-free *E. coli* metabolic model was constructed by removing the growth-associated processes from the model of Palsson and coworkers [19], and by adding reactions for the synthesis of chloramphenicol acetyltransferase (CAT), a model protein for which we have a comprehensive training dataset [28]. Thus, the model described core central carbon metabolism (glycolysis, pentose phosphate, Enter-Doudoroff, TCA cycle), as well as the synthesis of energy species, amino acids biosynthesis and degradation, and biosynthesis of the CAT protein. An ensemble of model parameters was estimated from dynamic 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. We generated an ensemble of N = 100 diverse parameter sets by minimizing the error between the training dataset and the metabolite concentrations predicted by the model, and choosing sets so as to minimize the Pearson correlation coefficient between sets in the ensemble. We also defined the single set with lowest cost function as our best-fit set.

The ensemble of models captured the time evolution of cell-free CAT biosynthesis (Fig. 1 - 4). The central carbon metabolism (Fig. 1, top), energy species (Fig. 3), and amino acids (Fig. 4) were all well captured by the ensemble and the best-fit set. Allosteric control was seen to be particularly important to the dynamics of the organic acid intermediates, as without it several of the measurements are not captured by the ensemble or the best-fit set (Fig. 1, bottom).

Sensitivity analysis We performed a local sensitivity analysis to determine the network reactions with the greatest effect on protein production and overall system state. CAT production was most sensitive to the CAT synthesis reaction, oxidative phosphorylation ac-

tivity, and alanine synthesis (Fig. 6, top, section A). The 16 next most important reactions to CAT production (section B) came from various pathways across the network: 4 each from glycolysis, the TCA cycle, and amino acid synthesis/degradation; 2 from pentose phosphate; and 1 each from the Entner-Doudoroff pathway and the energy species reactions. The pairwise sensitivities (off-diagonal elements) are often quite different from the corresponding first-order sensitivities (diagonal elements), and lead to some interesting outcomes. For example, glutamine synthesis and arginine degradation are both among the most important reactions to CAT production (they rank 5th and 10th, respectively). This is likely because they both affect the sensitive glutamine-glutamate balance; glutamine synthesis consumes glutamate, while arginine degradation produces it. However, when both are perturbed, their combined effect on the model is very low, as the respective consumption and production of glutamate cancel out.

The system state as a whole was most sensitive to glucose uptake via GTP and the forward reaction of lactate dehydrogenase (pyruvate being consumed to produce lactate) (Fig. 6, bottom, section F). The 30 next most important reactions to the system state (section G) came from various pathways across the network: 8 from amino acid synthesis/degradation; 6 from glycolysis; 4 from TCA; 2 each from pentose phosphate, Entner-Doudoroff, energy species reactions, NAD species reactions, and small molecule transport; 1 from oxidative phosphorylation, and 1 pyrophosphatase. The system state has even 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 is very high; it consumes both pruvate and glutamate, two key species in the network. In addition, a handful of alanine synthesis pairwise sensitivities are high. However, there are enough reactions that, when paired with alanine synthesis, have 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 is low, placing it at the

very bottom of section I. Another interesting result of this analysis is seen in the intersection of sections F and G with section J. The 53 reactions in section J were turned off in the best-fit set (rate constants were set to 0); therefore, the perturbation of these reactions (multiplying the rate constant by 1.01) had no actual effect on the model. Thus, all pairwise sensitivities with reactions in section J can essentially be considered first-order sensitivities for the other reactions. Interestingly, the reactions in section F and several in section G showed most of thier 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 uptake via GTP and CTP, fumarate reductase, and SO4 transport. This suggests 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.

Maximum theoretical CAT yield showed CFPS can be optimized. We calculated the carbon yield of CAT production for our experimental data and our best-fit parameter set as a function of the initial and final concentrations and the carbon numbers of CAT, glucose, and amino acids. The experimental data displayed a CAT yield of 0.0865, while the best-fit parameter set displayed a CAT yield of 0.0871. While the model ensemble described the experimental data, 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 detailed promoter model under a T7 polymerase to compute the maximum theoretical carbon yield. However, we first validated the ssFBA approach by comparing an ensemble of simulated versus measured concentrations of CAT over a one hour period (Fig. 7A). The ensemble of 100 sets captured the CAT concentration profile which was generated by sampling RNA polymerase levels, ribosome levels and elongation rates in a physiological range. We then used sequence-specific FBA to calculate a theoretical maximum CAT yield under three different cases: unconstrained, constrained by transcription/translation rates, and constrained by transcription/translation

rates and measurements (Fig. 7B). The theoretical maximum carbon yield of CAT was 0.35 ± 0.006 for an unconstrained case and 0.225 ± 0.03 for the transcription and translation constrained case. Thus, we showed that our experimental dataset and best-fit parameter set were each producing CAT at 25% of the theoretical maximum and 38% of a theoretical physiological case. Whereas, the case constrained by experimental data had a carbon yield of 0.062 ± 0.02, similar to the experimental yield. In comparing the flux distributions between the unconstrained and constrained cases (Fig. 8), the constrained cases heavily utilize the first step in the pentose phosphate pathway to generate NADPH. The majority of the flux continues through the Entner–Doudoroff pathway whereas in the unconstrained case, the majority of flux travels through glycolysis. In all cases, the energy source comes from oxidative phosphorylation with some from the citric acid cycle. Whereas, in just the TX/TL case, there is a high flux through fumerate dehydrogenase from aspartic acid uptake. In the unconstrained and most constrained case, we see a mixture of acetate and lactate accumulation. This shows 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 TX/TL 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 towards lactate, however since CAT production is constrained by the translation rate, we expected there to be no increase in CAT production. The decrease in acetate production is a promising result to increase CAT yield. A second simulation with a knockout of *gnd* and phosphate acetyltransferase,

showed carbon being diverted towards 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 developed an ensemble of *E. coli* cell-free protein synthesis (CFPS) models using the hybrid cell-free modeling approach of Wayman et al [REFHERE]. Model parameters were estimated from measurements of glucose, organic acids, energy species, amino acids and the protein product, chloramphenicol acetyltransferase (CAT). The ensemble described the training data, with the exception of some of the amino acid dynamics. 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 observed CAT yield was 25% of the maximum theoretical yield, suggesting CAT production could be further optimized. The metabolic flux distribution predicted by the dynamic model and flux balance analysis were significantly different. The ensemble of dynamic models predicted the majority of carbon flux was routed through glycolysis and the TCA cycle, while flux balance analysis predicted significant flux through the Entner-Doudoroff pathway. Local and global sensitivity analysis suggested CAT production was most sensitive to parameters and initial conditions directly associated with CAT synthesis, as well as GTP/GMP synthesis, amino acid synthesis, and to a lesser extent amino acid initial conditions. On the other hand, CAT production was robust to allosteric control parameters and the initial conditions of glucose and oxygen.

In comparing the theoretical maximum carbon yield of CAT from ssFBA predictions to the kinetic model and experimental measurements, suggests that there is potential for increasing CAT yield in CFPS as well as CFPS performance. The theoretical maximum yield of CAT was 0.35 for an unconstrained case and 0.225 for a transcription/translation constrained case. Knockouts of *gnd* and phosphate acetyltransferase, show carbon can be diverted away from acetate and potentially towards CAT or other proteins of interest expressed in CFPS. The other limitations to address in CFPS would be to enhance the transcription and translation rates since the protein of interest to be expressed 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 [29]. Underwood et al. has also shown the increase in ribosome levels does not significantly increase protein yields and rates, however adding elongation factors increased yields by 23% at 30 minutes[30]. In addition to improving CFPS performance, Jewett et al. has showed that oxidative phosphorylation operates in cell-free systems and knocking out oxidative phosphorylation reactions is detrimental to protein yield [31]. However, it is inconclusive how much oxidative phosphorylation activity there is compared to *in vivo* systems and both of our models suggest oxidative phosphorylation is vital to CAT production. Thus, this is a potential place for improvement to optimize CFPS for better performance and protein yield.

The cell-free model ensemble described the training data with the exception of some of the amino acids. Specifically, adding more reactions that consume amino acids would improve the model's ability to predict those that show a decrase in the experimental data. Also, including specific transcription and translation steps for CAT would allow us to more accurately model the complexity and the resource cost of protein synthesis. Another area for future work is to more thoroughly sample parameter space. For the metabolites in the dataset, initial conditions were fixed at the initial data values. All other parameters were varied in a manner so as to best fit the dataset. However, the resulting ensemble may not represent every biological or practical possibility. In a different region of parameter space, the system could behave differently, including the flux distribution through the network, the accuracy and spread of ensemble fits, the relative sensitivities, and the yield as a percentage of the theoretical maximum. 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 degradation 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), $r_j \left(\mathbf{x}, \mathbf{k}\right) = \bar{r}_j v_j$. In this study, we used either saturation or mass action kinetics. 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 regulatory 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. Each regulatory

transfer function took the form:

$$f_{ij}\left(\mathcal{Z}_i, k_{ij}, \eta_{ij}\right) = k_{ij}^{\eta_{ij}} \mathcal{Z}_i^{\eta_{ij}} / \left(1 + k_{ij}^{\eta_{ij}} \mathcal{Z}_i^{\eta_{ij}}\right) \tag{3}$$

where $\mathcal{Z}_i$ denotes the abundance factor i, k_{ij} denotes a gain parameter, and η_{ij} denotes a cooperativity parameter. In this study, we used $\mathcal{I}_j \in \{mean\}$ [**?**]. If a process has no modifying factors, $v_j = 1$. We used multiple saturation kinetics to model the reaction term $\bar{r}_j$:

$$\bar{r}_j = k_j^{max} \epsilon_i \left(\prod_{s \in m_j^-} \frac{x_s}{K_{js} + x_s} \right) \tag{4}$$

where k_j^{max} denotes the maximum rate for reaction j, ϵ_i denotes the scaled enzyme activity which catalyzes reaction j, and K_{js} denotes the saturation constant for species s in reaction j. The product in Equation (4) was carried out over the set of *reactants* for reaction j (denoted as m_j^-).

We added regulation to the network as informed by literature, for a total of 17 interactions. PEP was modeled as an inhibitor for phosphofructokinase [32, 33], PEP carboxykinase [32], PEP synthetase [32, 34], isocitrate dehydrogenase [32, 35], and isocitrate lyase/malate synthase [32, 35, 36], and as an activator for fructose-biphosphatase [32, 37–39]. AKG was modeled as an inhibitor for citrate synthase [32, 40, 41] and isocitrate lyase/malate synthase [32, 36]. 3PG was modeled as an inhibitor for isocitrate lyase/malate synthase [32, 36]. FDP was modeled as an activator for pyruvate kinase [32, 42] and PEP carboxylase [32, 43]. Pyruvate was modeled as an inhibitor for pyruvate dehydrogenase [32, 44, 45] and as an activator for lactate dehydrogenase [46]. Acetyl CoA was modeled as an inhibitor for malate dehydrogenase [32].

Generation of model ensemble We generated an ensemble of 100 diverse parameter sets via a Markov chain Monte Carlo random walk. Beginning with a single parameter

set as a starting point, we calculated its cost function, equal to the sum-squared-error between experimental data and model predictions:

$$cost = \sum_{i=1}^{\mathcal{D}} \left(w_i \sum_{j=1}^{\mathcal{T}_i} abs\left(x_{ij}^{data} - x_i^{sim}|_{t(j)}\right) \right) \tag{5}$$

where $\mathcal{D}$ denotes the number of datasets, equal to 37, w_i denotes the weight of the ith dataset, $\mathcal{T}_i$ denotes the number of timepoints in the ith dataset, $t(j)$ denotes the jth timepoint, x_{ij}^{data} denotes the value of the ith dataset at the jth timepoint, and $x_i^{sim}|_{t(j)}$ denotes the simulated value of the metabolite corresponding to the ith dataset, interpolated to the jth timepoint. We then perturbed model parameters:

$$k_i^{new} = k_i * exp(a\, r_i) \qquad i = 1, 2, \ldots, \mathcal{P} \tag{6}$$

where $\mathcal{P}$ denotes the number of parameters, equal to 815, which includes 163 rate constants, 163 enzyme degradation rate constants, 455 saturation constants, and 34 control parameters, k_i^{new} denotes the new value of the ith parameter, k_i denotes the current value of the ith parameter, a denotes a distribution variance, and r_i denotes a random sample from the normal distribution. 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 correlation to each other so as to avoid over-sampling any region of parameter space. The original 12,437-set ensemble had a mean Pearson correlation coefficient [REFERENCE NEEDED?] 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.

Local sensitivity analysis We determined the reactions most important to protein production by computing the local sensitivity of CAT concentration to each rate constant in the network. Each sensitivity index was formulated as:

$$S_{ij} = norm(CAT(p_i, p_j, t) - CAT(\alpha * p_i, \alpha * p_j, t)) \qquad i, j = 1, 2, \ldots \mathcal{P} \tag{7}$$

where S_{ij} denotes the sensitivity of CAT production to the ith and jth parameters, $CAT(p_i, p_j, t)$ denotes CAT concentration as a function of time and the ith and jth parameters, α denotes the perturbation factor, equal to 1.01, and $\mathcal{P}$ denotes the number of rate constants, equal to 163. In calculating the pairwise sensitivities, each parameter was perturbed by 1%; first-order sensitivities (i = j) were subject to two 1% perturbations, equivalent to a perturbation of 2.01%. The array of sensitivity indices was organized into a clustergram (Fig. 6, top) with both axes corresponding to the 163 rate constants rearranged from most sensitive (top, right) to least sensitive (bottom, left).

Likewise, we determined the reactions most important to the system as a whole by computing the local sensitivity of the species for which data exist to each rate constant in the network. Each sensitivity index was formulated as:

$$S_{ij} = norm(X(p_i, p_j, t) - X(\alpha * p_i, \alpha * p_j, t)) \qquad i, j = 1, 2, \ldots \mathcal{P} \tag{8}$$

where S_{ij} denotes the sensitivity of the system state to the ith and jth parameters, and $X(p_i, p_j, t)$ denotes the system state, an array consisting of the concentration of every species for which data exist as a function of time and the ith and jth parameters. This array of sensitivity indices was also organized into a clustergram (Fig. 6, bottom).

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):

$$Yield = \frac{\Delta CAT\ C_{CAT}}{\sum_{i=1}^{\mathcal{R}} \max(\Delta m_i, 0)\ C_{m_i}} \tag{9}$$

where ΔCAT 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 ith reactant consumed, never allowed to be negative, and C_{m_i} denotes the carbon number of the ith reactant. Because no data was available for arginine or glutamate, these reactants were left out of all three calculations. In the experimental case and the best-fit set case, yield was 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. The theoretical yield was calculated using flux balance anaylsis (FBA) with a sequence-specific based analysis on CAT. The sequence specific FBA [47] problem was formulated as:

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

where $\mathbf{S}$ denotes the stoichiometric matrix, $\mathbf{w}$ denotes the unknown flux vector, $\boldsymbol{\theta}$ denotes the objective selection vector and α_i and β_i denote the lower and upper bounds on flux w_i, respectively. The 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/hr according to experimental data; the specific amino acid uptake rates were bound to allow a maximum flux of 30 mM/hr, but did not reach this maximum flux. The transcription and translation template reactions were added to the metabolic network and are based off sequence specific analysis [47] involving transcription initiaion, transcription, mRNA degradation, translation initiation, translation, and tRNA charging. The flux bal-

ance analysis problem was solved using the GNU Linear Programming Kit (v4.52) [48]. The solution flux vector was used to calculate the theoretical carbon yield of CAT. Glucose, oxygen, and amino acids were modeled as being imported into the system, whereas CAT synthesis and metabolite byproduct formation was modeled as an export from the system. The rest of the network followed a pseudo steady-state asusmption where all other metabolites were not allowed to accumulate; thus, the network could be solved by linear programming.

The transcription rate was constrained as:

$$w_{tx} = RNAP \frac{v_{RNAP}}{l_{mRNA}} \left(\frac{Gene}{km + Gene} \right) P$$

where $RNAP$ is the concentration of RNA polymerase, v_{RNAP} is the elongation rate (nucleotides/hr) by the RNA polymerase, l_{mRNA} is the number of nucleotides in the mRNA, $Gene$ is the gene concentration, km is the plasmid saturation coefficient, and P is 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:

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

where K_1 represents the state of T7 RNA polymerase binding.

The translation rate was constrained as:

$$w_{tl} = K_P Ribo \frac{v_{Ribo}}{l_{protein}} [mRNA_{ss}]$$

where K_P is the polysome amplification constant, $Ribo$ is the ribosome concentration, v_{Ribo} is the elongation rate (amino acids/hr) of the ribosome, $l_{protein}$ is the number of amino acids in the protein of interest, and $mRNA_{ss}$ is the mRNA concentration at steady state

363 determined by the transcription rate divided by the degradation rate of mRNA.

364 An ensemble of 100 sets of flux distributions was calculated for three different cases,
365 unconstrained, constrained by transcription and translation (TX/TL) rates, and constrained
366 by TX/TL rates and experimental data. For the unconstrained case, all rates were left un-
367 bounded, except for the specific glucose uptake rate. An ensemble of flux distributions
368 was calculated by randonmly sampling the maximum specific glucose uptake rate from 30
369 to 40 mM/hr determined from experimental data. For the case constrained by TX/TL rates,
370 an ensemble was generated by randomly sampling RNAP polymerase levels, ribosome
371 levels, and elongation rates in a physiological range determined from literature. RNA poly-
372 merase levels were sampled between 60 and 80 nM, ribosome levels between 7 and 16
373 µM, the elongation rate by RNA polymerase between 20 and 30 nts/sec, and the elon-
374 gation rate by ribosomes between 1.5 and 3 AA/sec [30, 49]. For the case constrained
375 by TX/TL rates and experimental data, an ensemble was generated by randomly sam-
376 pling RNAP polymerase levels, ribosome levels, and elongation rates in a physiological
377 range determined from literature. The lower and upper bound constraints were randomly
378 sampled in the physiological range of the experimental noise where data was available,
379 except for CAT flux, which was determined from the transcription and translation rates.

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. Fredrickson AG (1976) Formulation of structured growth models. Biotechnol Bioeng 18: 1481-6.
7. 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.
8. Steinmeyer D, Shuler M (1989) Structured model for Saccharomyces cerevisiae. Chem Eng Sci 44: 2017 - 2030.
9. Wu P, Ray NG, Shuler ML (1992) A single-cell model for cho cells. Ann N Y Acad Sci 665: 152-87.
10. 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.
11. Atlas JC, Nikolaev EV, Browning ST, Shuler ML (2008) Incorporating genome-wide

410 dna sequence information into a dynamic whole-cell model of escherichia coli: appli-
411 cation to dna replication. IET Syst Biol 2: 369-82.
412 12. Dhurjati P, Ramkrishna D, Flickinger MC, Tsao GT (1985) A cybernetic view of micro-
413 bial growth: modeling of cells as optimal strategists. Biotechnol Bioeng 27: 1-9.
414 13. Kompala DS, Ramkrishna D, Jansen NB, Tsao GT (1986) Investigation of bacterial
415 growth on mixed substrates: experimental evaluation of cybernetic models. Biotech-
416 nol Bioeng 28: 1044-55.
417 14. Kim JI, Song HS, Sunkara SR, Lali A, Ramkrishna D (2012) Exacting predictions by
418 cybernetic model confirmed experimentally: steady state multiplicity in the chemostat.
419 Biotechnol Prog 28: 1160-6.
420 15. Varner J, Ramkrishna D (1999) Metabolic engineering from a cybernetic perspective:
421 aspartate family of amino acids. Metab Eng 1: 88-116.
422 16. Song HS, Morgan JA, Ramkrishna D (2009) Systematic development of hybrid cy-
423 bernetic models: application to recombinant yeast co-consuming glucose and xylose.
424 Biotechnol Bioeng 103: 984-1002.
425 17. Song HS, Ramkrishna D (2011) Cybernetic models based on lumped elementary
426 modes accurately predict strain-specific metabolic function. Biotechnol Bioeng 108:
427 127-40.
428 18. Lewis NE, Nagarajan H, Palsson BØ (2012) Constraining the metabolic genotype-
429 phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 10:
430 291-305.
431 19. Edwards JS, Palsson BØ (2000) The escherichia coli mg1655 in silico metabolic
432 genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci U S
433 A 97: 5528-33.
434 20. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A genome-
435 scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for

436 1260 orfs and thermodynamic information. Mol Syst Biol 3: 121.
437 21. Oh YK, Palsson BØ, Park SM, Schilling CH, Mahadevan R (2007) Genome-scale re-
438 construction of metabolic network in bacillus subtilis based on high-throughput phe-
439 notyping and gene essentiality data. J Biol Chem 282: 28791-9.
440 22. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ (2009) Reconstruction of bio-
441 chemical networks in microorganisms. Nat Rev Microbiol 7: 129-43.
442 23. Ibarra RU, Edwards JS, Palsson BØ (2002) Escherichia coli k-12 undergoes adaptive
443 evolution to achieve in silico predicted optimal growth. Nature 420: 186-9.
444 24. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions
445 for predicting intracellular fluxes in escherichia coli. Mol Syst Biol 3: 119.
446 25. Hyduke DR, Lewis NE, Palsson BØ (2013) Analysis of omics data with genome-scale
447 models of metabolism. Mol Biosyst 9: 167-74.
448 26. McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale
449 metabolic network reconstructions of escherichia coli. Mol Syst Biol 9: 661.
450 27. Zomorrodi AR, Suthers PF, Ranganathan S, Maranas CD (2012) Mathematical opti-
451 mization applications in metabolic networks. Metab Eng 14: 672-86.
452 28. Calhoun KA, Swartz JR (2005) An economical method for cell-free protein synthesis
453 using glucose and nucleoside monophosphates. Biotechnology Progress 21: 1146–
454 1153.
455 29. Li J, Gu L, Aach J, Church GM (2014) Improved cell-free rna and protein synthesis
456 system. PLoS ONE 9: 1-11.
457 30. Underwood KA, Swartz JR, Puglisi JD (2005) Quantitative polysome analysis iden-
458 tifies limitations in bacterial cell-free protein synthesis. Biotechnology and Bioengi-
459 neering 91: 425–435.
460 31. Jewett MC, Calhoun KA, Voloshin A, Wuu JJ, Swartz JR (2008) An integrated cell-free
461 metabolic platform for protein production and synthetic biology. Molecular Systems

462 Biology 4.
463 32. Kotte O, Zaugg JB, Heinemann M (2010) Bacterial adaptation through distributed
464 sensing of metabolic fluxes. Mol Syst Biol 6: 355.
465 33. Cabrera R, Baez M, Pereira HM, Caniuguir A, Garratt RC, et al. (2011) The crys-
466 tal complex of phosphofructokinase-2 of Escherichia coli with fructose-6-phosphate:
467 kinetic and structural analysis of the allosteric ATP inhibition. J Biol Chem 286: 5774–
468 5783.
469 34. Chulavatnatol M, Atkinson DE (1973) Phosphoenolpyruvate synthetase from Es-
470 cherichia coli. Effects of adenylate energy charge and modifier concentrations. J
471 Biol Chem 248: 2712–2715.
472 35. Ogawa T, Murakami K, Mori H, Ishii N, Tomita M, et al. (2007) Role of phospho-
473 enolpyruvate in the NADP-isocitrate dehydrogenase and isocitrate lyase reaction in
474 Escherichia coli. J Bacteriol 189: 1176–1178.
475 36. MacKintosh C, Nimmo HG (1988) Purification and regulatory properties of isocitrate
476 lyase from Escherichia coli ML308. Biochem J 250: 25–31.
477 37. Donahue JL, Bownas JL, Niehaus WG, Larson TJ (2000) Purification and character-
478 ization of glpX-encoded fructose 1, 6-bisphosphatase, a new enzyme of the glycerol
479 3-phosphate regulon of Escherichia coli. J Bacteriol 182: 5624–5627.
480 38. Hines JK, Fromm HJ, Honzatko RB (2006) Novel allosteric activation site in Es-
481 cherichia coli fructose-1,6-bisphosphatase. J Biol Chem 281: 18386–18393.
482 39. Hines JK, Fromm HJ, Honzatko RB (2007) Structures of activated fructose-1,6-
483 bisphosphatase from Escherichia coli. Coordinate regulation of bacterial metabolism
484 and the conservation of the R-state. J Biol Chem 282: 11696–11704.
485 40. Pereira DS, Donald LJ, Hosfield DJ, Duckworth HW (1994) Active site mutants of
486 Escherichia coli citrate synthase. Effects of mutations on catalytic and allosteric prop-
487 erties. J Biol Chem 269: 412–417.

41. 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: 51–54.
42. 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: 116–120.
43. Wohl RC, Markus G (1972) Phosphoenolpyruvate carboxylase of Escherichia coli. Purification and some properties. J Biol Chem 247: 5785–5792.
44. 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: 28106–28116.
45. 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: 5213–5221.
46. 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: 449–454.
47. Allen TE, Palsson BØ (2003) Sequence-based analysis of metabolic demands for protein synthesis in prokaryotes. Journal of Theoretical Biology 220: 1 - 18.
48. (2016). GNU Linear Programming Kit, Version 4.52. URL http://www.gnu.org/software/glpk/glpk.html.
49. Garamella J, Marshall R, Rustad M, Noireaux V .

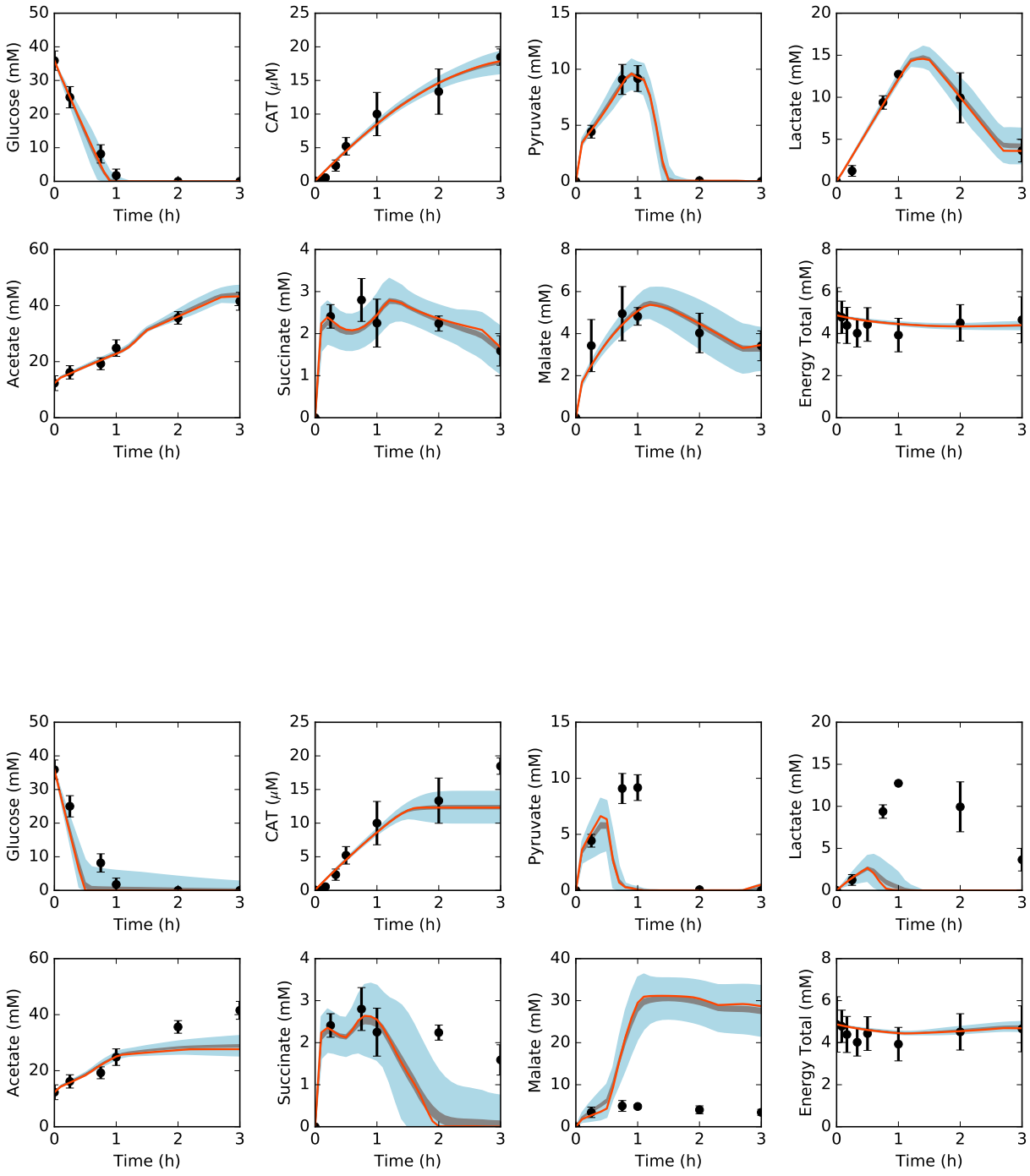

Fig. 1: Central carbon metabolism in the presence (top) and absence (bottom) of allosteric control, including glucose (substrate), CAT (product), and intermediates, as well as total concentration of energy species. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence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

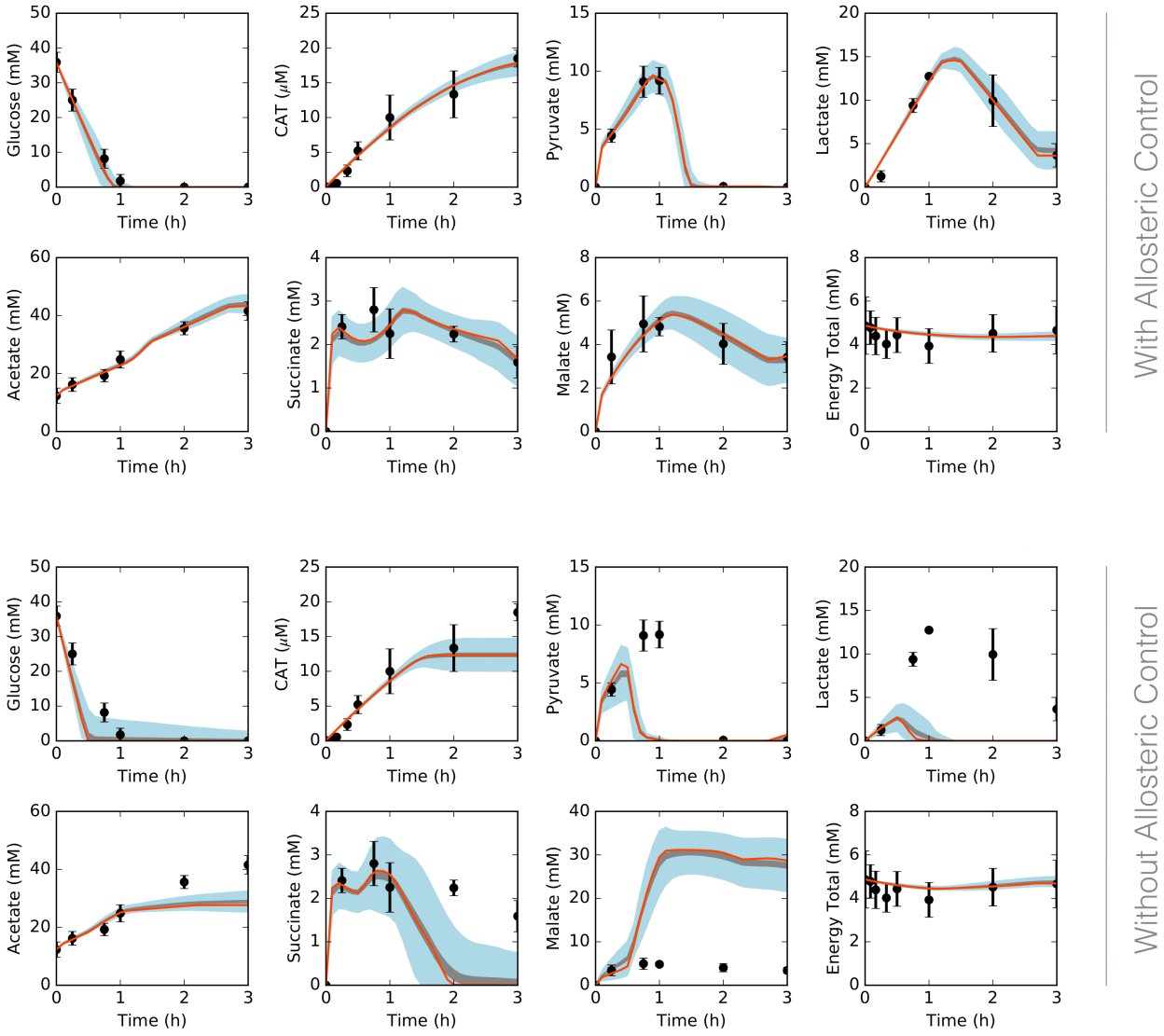

Fig. 2: Central carbon metabolism in the presence (top) and absence (bottom) of allosteric control, including glucose (substrate), CAT (product), and intermediates, as well as total concentration of energy species. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence interval (blue shaded region) and 95% confidence interval of the mean (gray shaded region) over the ensemble of 100 sets.

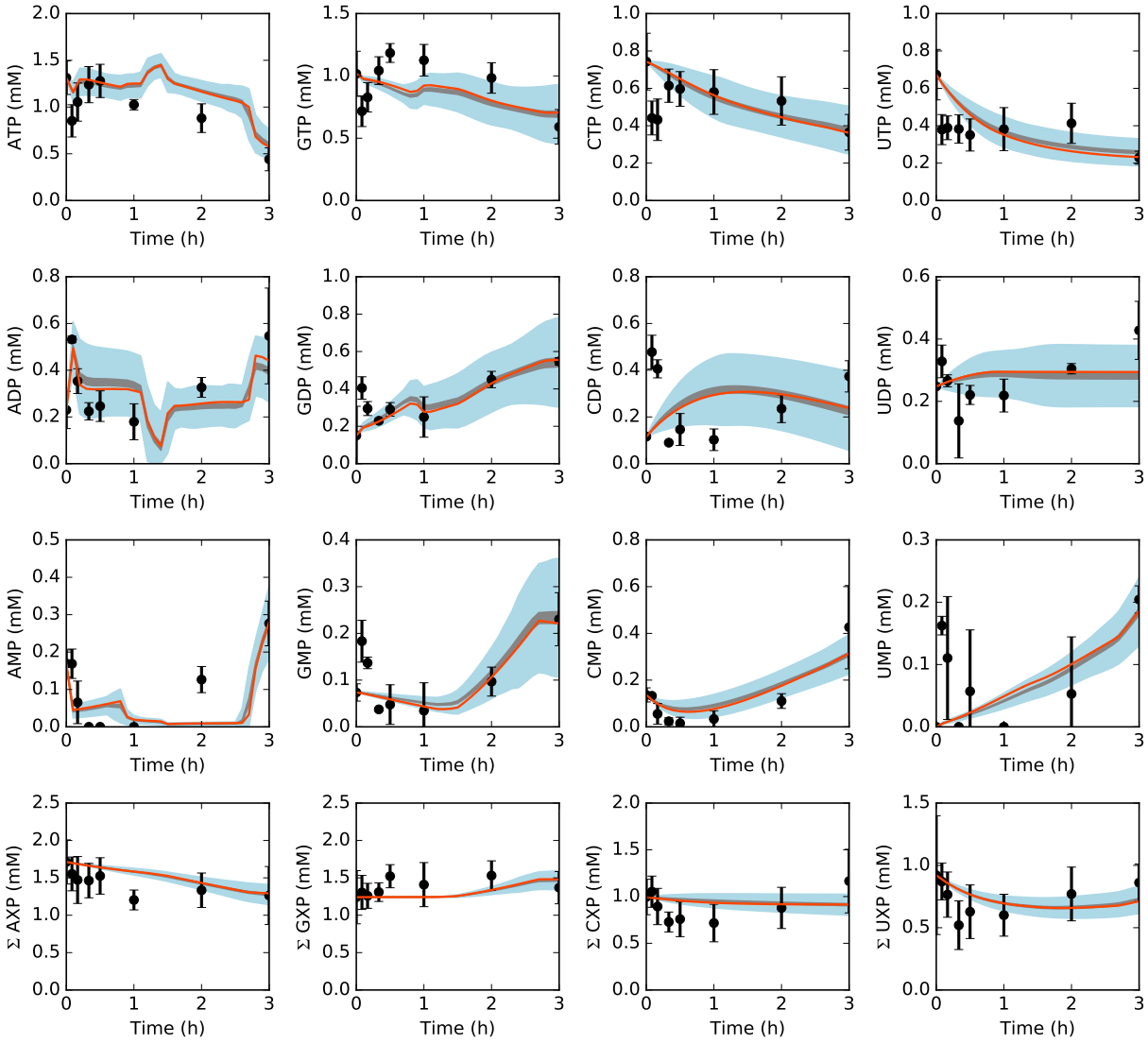

Fig. 3: Energy species and energy totals by base in the presence of allosteric control. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence interval (blue shaded region) and 95% confidence interval of the mean (gray shaded region) over the ensemble of 100 sets.

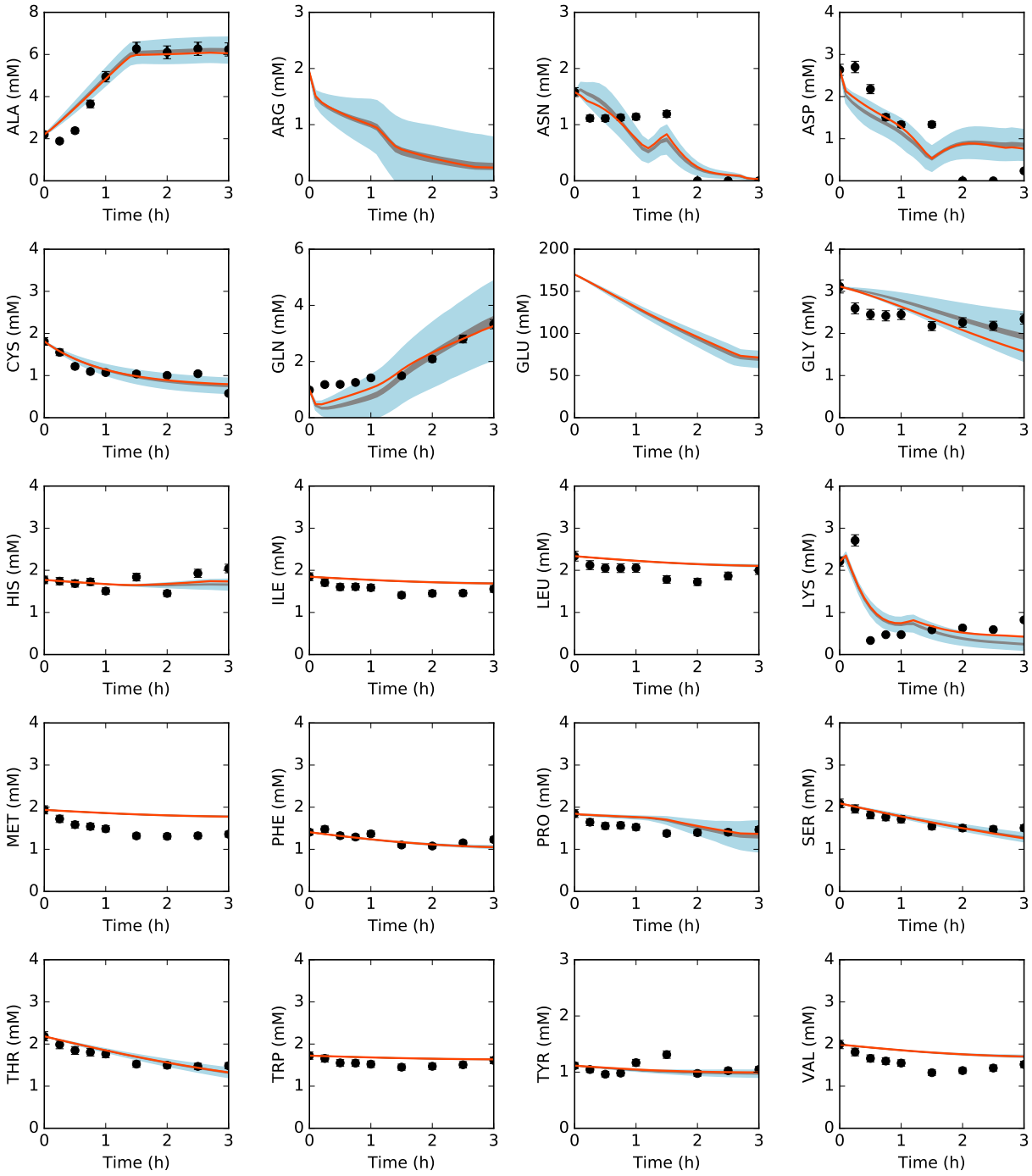

Fig. 4: Amino acids in the presence of allosteric control. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence interval (blue shaded region) and 95% confidence interval of the mean (gray shaded region) over the ensemble of 100 sets.

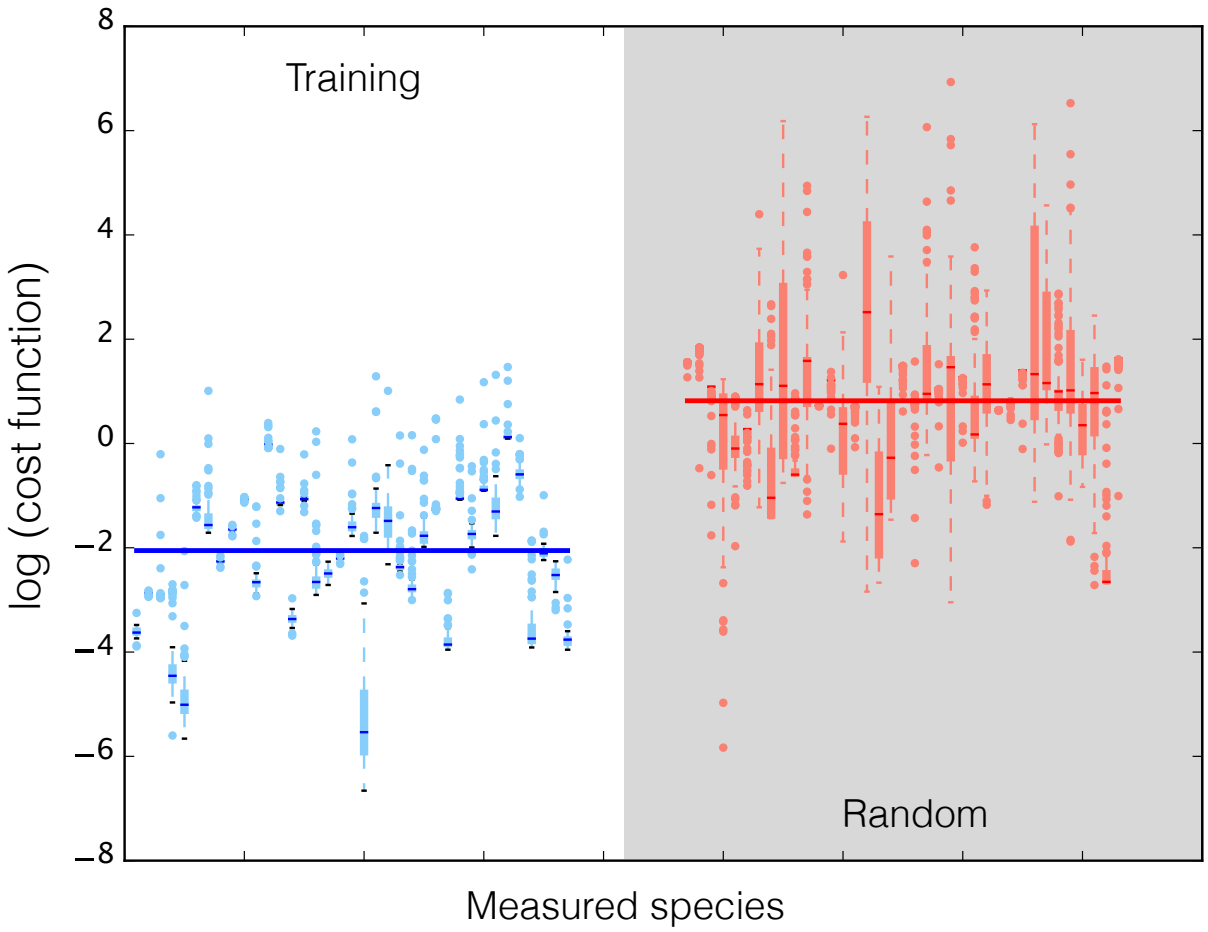

Fig. 5: Log of cost function across 37 datasets for data-trained ensemble (blue) and randomly generated ensemble (red, gray background). Median (bars), interquartile range (boxes), range excluding outliers (dashed lines), and outliers (circles) for each dataset. Median across all datasets (large bar overlaid).

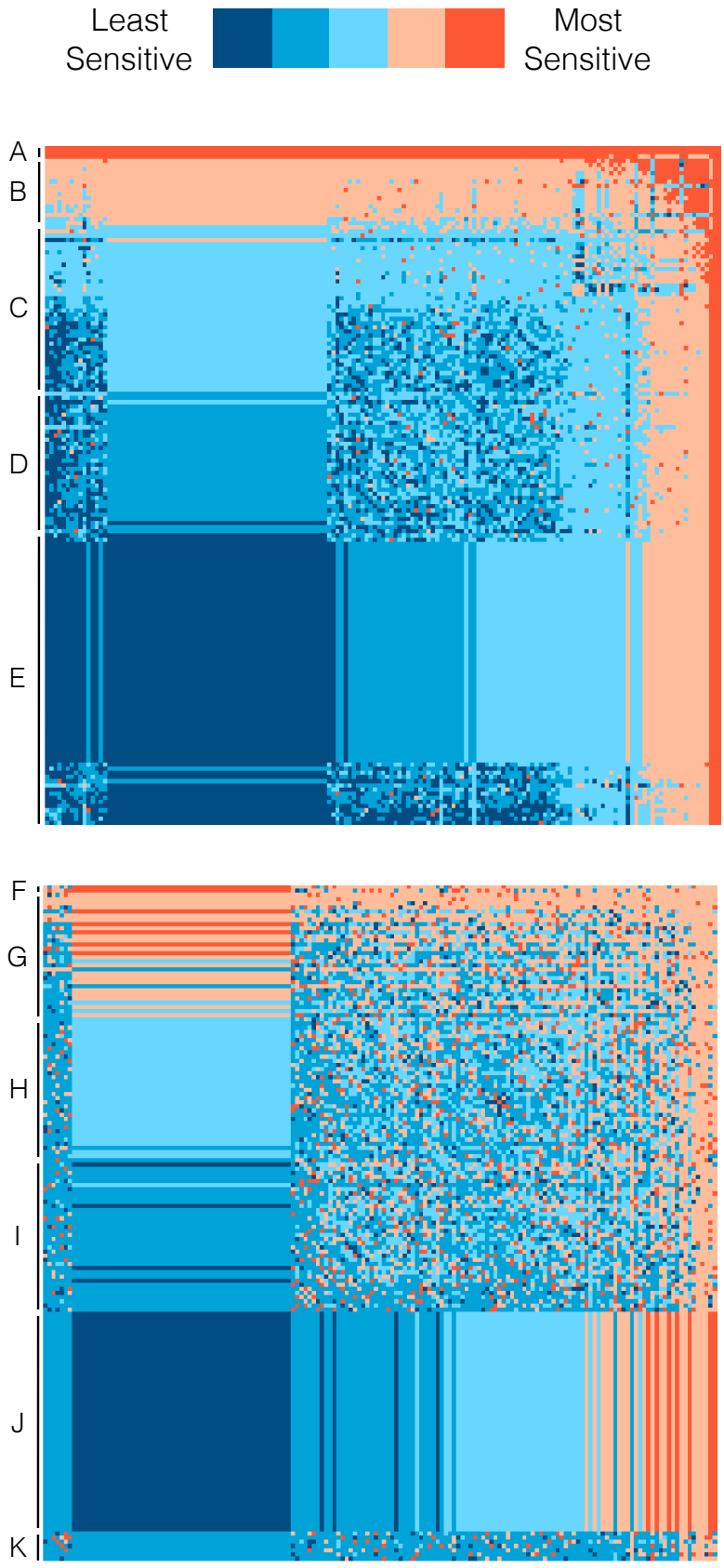

Fig. 6: Normalized first-order and pairwise sensitivities of CAT production (top) and system state (bottom) to rate constants.

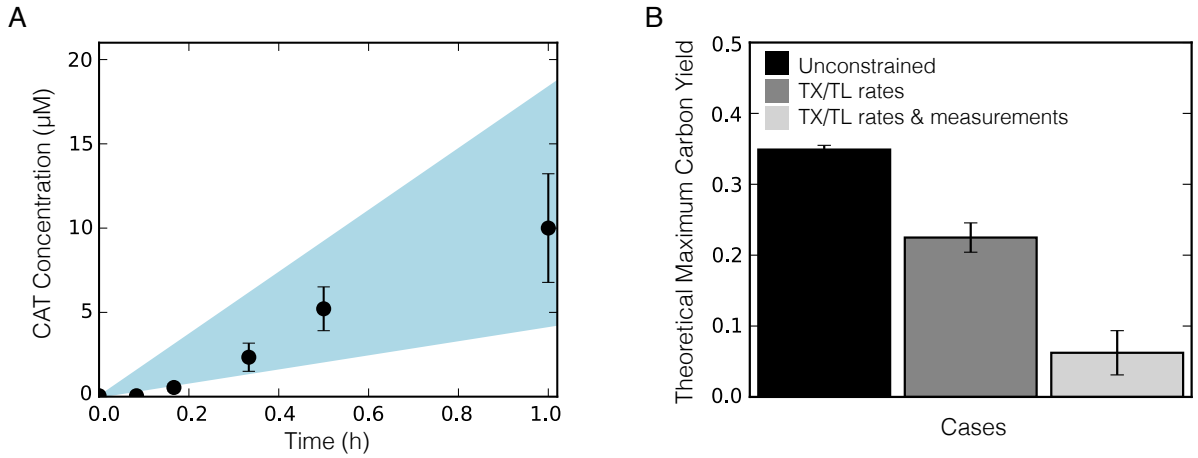

Fig. 7: Sequence specific flux balance analysis of CAT production and yield. A. 95% confidence interval of the ensemble (light blue region) for CAT concentration versus time. B. Theoretical maximum carbon yield of CAT calcualted by ssFBA for three different cases: unconstrained except for glucose uptake (black), constrained by transcription and translation rates (grey), and constrained by transcription, translation rates and experimental measurements where available (light grey). Error bars represent standard deviation of the ensemble.

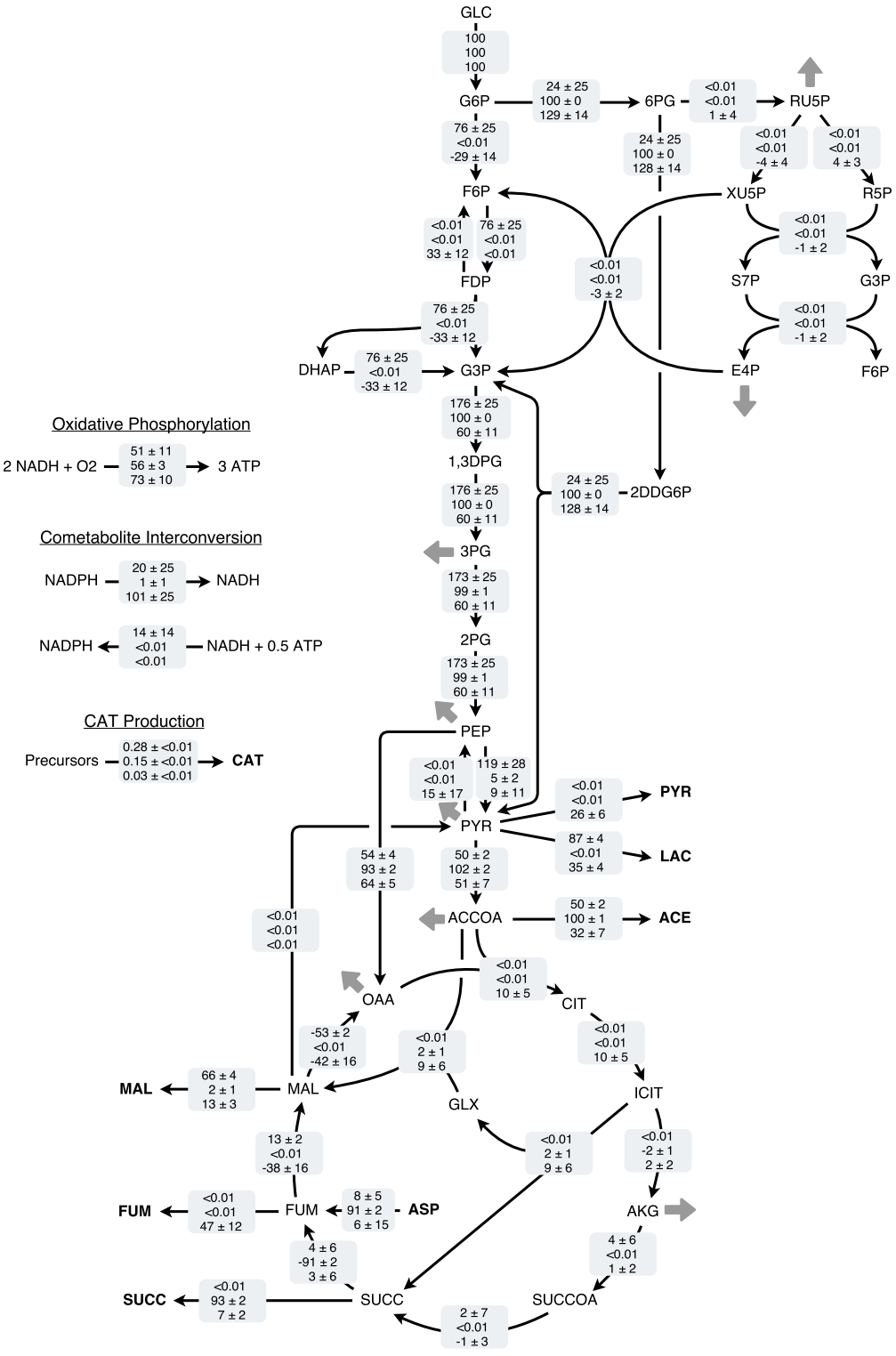

Fig. 8: Flux profile for glycolysis, pentose phosphate pathway, Entner-Doudoroff pathway, TCA cycle, NADPH/NADH transfer, and oxidative phosphorylation. Sequence specific FBA flux value (mean ± standard deviation) across ensemble for 1 hr, normalized to glucose uptake flux. Flux distribution for three 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).