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 captured all of the training data, especially the central carbon metabolism. CAT was produced with a carbon yield of 7% and an energy efficiency of 5%, suggesting that CAT production could be further optimized. Reaction group knockouts showed that protein productivity and the metabolism as a whole depend most on oxidative phosphorylation and glycolysis and gluconeogenesis. Amino acid biosynthesis is also important for productivity, while the overflow metabolism and TCA cycle affect the overall system state. In addition, 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 metabolite 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 more 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. By knocking out metabolic reactions in groups, we showed that CAT productivity depends most on oxidative phosphorylation, glycolysis/gluconeogenesis, and cer-

tain amino acid synthesis reactions, while the system state is most sensitive to glycolysis/gluconeogenesis, oxidative phosphorylation, the overflow metabolism, and the TCA cycle reactions. 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 *i*AF1260 reconstruction of K-12 MG1655 *E. coli* [16], and by adding reactions describing chloramphenicol acetyltransferase (CAT) biosynthesis, a model protein for which there exists a comprehensive training dataset [23]. In addition, reactions that were knocked out from the cell extract preparation were removed from the network (ΔspeA, ΔtnaA, ΔsdaA, ΔsdaB, ΔgshA, ΔtonA, ΔendA). The CFPS model equations were formulated using the hybrid cell-free modeling framework of Wayman et al. [24]. An initial ensemble of model parameter sets (N $>$ 30,000) 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. A final ensemble of parameter sets (N $=$ 100) was constructed by selecting the sets with the lowest errors, the lowest of which was defined as the best-fit set. Parameter sets in the final ensemble had a mean Pearson correlation coefficient of 0.77; thus, an accurate yet diverse ensemble was created.

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 more than two orders of magnitude less than random parameter sets generated within the same parameter bounds (Fig. 4); thus, we have confidence in the predictive capability of the estimated parameters. The model captured the biphasic CAT production: during the first hour glucose powers production, and CAT is produced at ~10 μM/h; subsequently, pyruvate and lactate reserves are consumed to power metabolism, and CAT is produced less efficiently at ~5 μM/h. Allosteric control was important to central carbon metabolism, especially pyruvate, acetate, and succinate (Fig. 1, bottom). The difference between the allosteric control and no-control cases is mostly seen in the second phase of CAT production, after glucose is exhausted. Taken together, we produced an ensemble of kinetic models that was consistent with time series measurements of the production of a model protein. Although the ensemble described the experimental data, it was unclear which kinetic parameters most influenced CAT production, and whether the performance of the CFPS reaction was optimal.

To better understand the effect of network reactions on system performance we conducted a group knockout analysis (Fig. 5). The network was divided into 19 groups of reactions, spanning central carbon metabolism, energetics, and amino acid biosynthesis. The reactions in each of these groups were turned off, and the resulting change in productivity and system state were recorded. Then each pair of groups was knocked out to determine pairwise effects. These were summed with the first-order effect to obtain a total-order coefficient for each group for the change in productivity and system state. Glycolysis/gluconeogenesis and oxidative phosphorylation were seen to have the greatest effect on both productivity and system state. This is explained by their role in both central carbon metabolism and energy generation. In addition, CAT productivity is affected by two sectors of amino acid biosynthesis: alanine/aspartate/asparagine, and glutamate/glutamine. This is likely because aspartate, glutamate, and glutamine are key reactants in the biosynthesis of many other amino acids, all of which are required for CAT synthesis. Meanwhile, the TCA cycle and the overflow metabolism (which includes acetyl-coA/acetate reactions and the interconversion of pyruvate and lactate) have a significant effect on the system state. These reactions directly impact key species in the system state: succinate and malate in the TCA cycle, and acetate, pyruvate, and lactate in the overflow metabolism.

To understand whether the CFPS performance was optimal, we calculated the carbon yield and energy efficiency of CAT production for the best-fit set (Fig. 6). Of the carbon taken up as glucose and consumed as amino acids, 7% went to CAT production. Another 7% went to the accumulation of alanine and glutamine, in line with experimental data (these were not included in the amino acid consumption term). 25% of the carbon was accumulated as organic acids (lactate, acetate, succinate, and malate), leaving 61% as accumulation of other byproducts, species for which data do not exist. CFPS also showed an energy efficiency of 5%; this was formulated as the equivalent ATP value of CAT produced divided by the equivalent ATP value of glucose consumed. The remainder of the balance was very similar to that of the carbon balance: 8% to alanine and glutamine, 26% to organic acids, and 61% to other byproducts. The equivalent ATP numbers for glucose, amino acids, and organic acids were calculated from the network stoichiometry; for example, 21 molecules of ATP shold be generated from one glucose molecule if the optimal path through glycolysis, the TCA cycle, and oxidative phosphorylation is taken. The ATP number for CAT was simply equal to the cost of transcription and translation. These results show that there is much room for improvement of the efficiency of CFPS. This demonstrates that CAT carbon yield could be improved by diverting flux away from some of the less efficient or altogether unnecessary metabolic pathways.

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. Our ensemble of models accurately predicts dynamic experimental measurements of central carbon metabolism, energy species, and amino acids over 100 times better than random sets in the same region of parameter space. CFPS is seen to be biphasic, relying on glucose during the first hour and pyruvate and lactate afterward. Allosteric control was essential to the maintenance of the network, specifically the central carbon metabolism. Without it, pyruvate, succinate, and malate are consumed more quickly following glucose exhaustion, presumably to power downstream reactions and ultimately CAT synthesis. Meanwhile, acetate accumulation during the increased after the first hour, suggesting a decrease in sytem efficiency. Interestingly, CAT production is virtually unaffected. This is because the amino acids and energy species used in CAT synthesis are not dependent on allosteric control. Having captured the experimental data, we investigated if CAT yield and CFPS performance could be further improved. We showed that the model predicts CAT production with a carbon yield of 7% and energy efficiency of 5%. The accumulation of waste byproducts, especially acetate and carbon dioxide, is responsible for this sub-optimal performance. However, there is also some accumulation of useful metabolites in the central carbon metabolism; if these could be utilized fully, CAT production would likely increase. Knocking out sections of network metabolism revealed that glycolysis/gluconeogenesis and oxidative phosphorylation were the most important to CAT production and the system as a whole. Productivity was also heavily dependent on the synthesis reactions of alanine, aspartate, asparagine, glutamate, and glutamine, while TCA cycle and overflow reactions affected the system state. Taken together, these findings represent the first dynamic model of *E. coli* cell-free protein synthesis, and an important step toward a functional genome scale description.

We present an ensemble of models that quantitatively describes the system behavior of cell-free metabolism and production of CAT. Experimental observations of the metabolites 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 predicting the redirection of carbon fluxes to product-forming pathways. In analyzing the effect of reaction groups on CAT production and the system state, the regions of metabolism associated with substrate utilization and subsequent energy generation are the most important. Oxidative phosphorylation is vital, since it provides most of the energetic needs of CFPS. Jewett and coworkers observed a decrease in CAT yield, between 1.5-fold and 4-fold, when knocking out oxidative phosphorylation reactions [1]. While it is unknown how active oxidative phosphorylation is compared to that of *in vivo* systems, our modeling approach suggests its importance to CFPS performance and protein yield. However, the biphasic operation of CFPS highlights the ability of the system to respond to an absence of glucose. During the first phase, there is an accumulation of central carbon metabolites with the majority of flux going toward acetate and some toward pyruvate, lactate, succinate and malate. While acetate continues to accumulate as a byproduct, the other organic acids are consumed as secondary substrates after glucose is no longer available. Glutamate also serves as a substrate throughout both phases, powering amino acid synthesis. These results show that CAT production can be sustained by other substrates in the absence of glucose, providing alternative strategies to optimize CFPS performance. While CAT synthesis can be powered by other substrates, the rate of production about half (~5 μM/h, as opposed to ~10 μM/h). This is in accordance with literature, where pyruvate provided a relatively slow but continuous supply of ATP [25].

Taken together, this shows CFPS can be designed towards a specified application, either requiring a slow stable energy source or faster production.

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.

Our ensemble of models accurately predicts dynamic experimental measurements of central carbon metabolism, energy species, and amino acids over 100 times better than random sets in the same region of parameter space.

CFPS is seen to be biphasic, relying on glucose during the first hour and pyruvate and lactate afterward. Allosteric control was essential to the maintenance of the network, specifically the central carbon metabolism. Without it, pyruvate, succinate, and malate are consumed more quickly following glucose exhaustion, presumably to power downstream reactions and ultimately CAT synthesis. Meanwhile, acetate accumulation during the increased after the first hour, suggesting a decrease in sytem efficiency. Interestingly, CAT production is virtually unaffected. This is because the amino acids and energy species used in CAT synthesis are not dependent on allosteric control. Having captured the experimental data, we investigated if CAT yield and CFPS performance could be further improved. We showed that the model predicts CAT production with a carbon yield of 7% and energy efficiency of 5%. The accumulation of waste byproducts, especially acetate and carbon dioxide, is responsible for this sub-optimal performance. However, there is also some accumulation of useful metabolites in the central carbon metabolism; if these could be utilized fully, CAT production would likely increase. Knocking out sections of network metabolism revealed that glycolysis/gluconeogenesis and oxidative phosphorylation were the most important to CAT production and the system as a whole. Productivity was also heavily dependent on the synthesis reactions of alanine, aspartate, asparagine, glutamate, and glutamine, while TCA cycle and overflow reactions affected the system state.

Taken together, these findings represent the first dynamic model of *E. coli* cell-free protein synthesis, and an important step toward a functional genome scale description.

This work represents the first dynamic model of *E. coli* cell-free protein synthesis. We apply a hybrid modeling framework to capture an experimental dataset for production of a test protein, and identify system limitations and areas of improvement for production efficiency. This work could be extended through further experimentation 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 the importance of those substrates. in addition, 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:

$$\begin{aligned} \frac{dx_i}{dt} &= \sum_{j=1}^{\mathcal{R}} \sigma_{ij} r_j \left(\mathbf{x}, \epsilon, \mathbf{k}\right) \qquad i = 1, 2, \dots, \mathcal{M} && (1) \\ \frac{d\epsilon_i}{dt} &= -\lambda_i \epsilon_i \qquad i = 1, 2, \dots, \mathcal{E} && (2) \end{aligned}$$

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 [26, 27], PEP carboxykinase [26], PEP synthetase [26, 28], isocitrate dehydrogenase [26, 29], and isocitrate lyase/malate synthase [26, 29, 30], and as an activator for fructose-biphosphatase [26, 31–33]. AKG was modeled as an inhibitor for citrate synthase [26, 34, 35] and isocitrate lyase/malate synthase [26, 30]. 3PG was modeled as an inhibitor for isocitrate lyase/malate synthase [26, 30]. FDP was modeled as an activator for pyruvate kinase [26, 36] and PEP carboxylase [26, 37]. Pyruvate was modeled as an inhibitor for pyruvate dehydrogenase [26, 38, 39] and as an activator for lactate dehydrogenase [40]. Acetyl CoA was modeled as an inhibitor for malate dehydrogenase [26].

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

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

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.

Model parameters were constrained by literature. Rate maxima were bounded within one order of magnitude of the literature value where available (Table 1). All other rate maxima were bounded within two orders of magnitude of the geometric mean of the available values. The median maximum reaction rate was 7.8 mM/h; assuming a total cell-free enzyme concentration of 167 nM, this corresponds to a median catalytic rate of 0.08 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 = 156 h. Saturation constants were bounded between 0.0001 and 10 mM; median = 1.0 mM. Control parameters (gains and orders) were bounded between 0.1 and 10 (dimensionless); median = 0.74. For each newly generated parameter set, we re-solved the balance equations and calculated the cost function. All sets with a lower cost (and some with higher cost) were accepted into the ensemble. After generating over 30,000 sets, N = 100 sets with minimal error were selected for the final ensemble. The final ensemble had a mean Pearson correlation coefficient of 0.77.

Comparison against random ensemble. A random ensemble of 100 parameter sets was generated from within the same parameter bounds as the trained ensemble. Sets were sampled using a Monte Carlo approach: each parameter was taken from a uniform distribution constructed between its upper and lower bounds. The model equations were then solved and the cost function was calculated in terms of the 37 separate experimental datasets. The random ensemble had a log median error of 0.80 across the datasets, as compared with a log median error of -1.43 for the trained ensemble (Fig. 4). Thus, the trained ensemble fits the dataset over one hundred times better than a random ensemble generated within the same bounds.

Group knockouts. The network was divided into 19 groups: glycolysis/gluconeogenesis, pentose phosphate, Entner-Doudoroff, TCA cycle, oxidative phosphorylation, cofactor reactions, anaplerotic/glyoxylate reactions, overflow metabolism, folate synthesis, purine/pyrimidine reactions, alanine/aspartate/asparagine synthesis, glutamate/glutamine synthesis, arginine/proline synthesis, glycine/serine synthesis, cysteine/methionine synthesis, threonine/lysine synthesis, histidine synthesis, tyrosine/tryptophan/phenylalanine synthesis, and valine/leucine/isoleucin synthesis. Each group of reactions was turned off individually, and then in pairs, and the model equations were re-solved. The CAT productivity was calculated and compared to that of the best-fit set (Fig. 5A). The absolute difference in productivity was recorded for each first-order knockout (diagonal elements) and each pairwise knockout, and a total-order coefficient was calculated by summing the first-order effect with all pairwise effects. Total-order coefficients were then normalized to fit within the same colorbar range as the first-order and pairwise effects. The system state was also calculated for each simulation, defined as the model predictions for all species for which data exist. The norm of the difference between the knockout system state and the best-fit system state is shown in

(Fig. 5B).

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

$$Y_C^{CAT} = \frac{\Delta\texttt{CAT} \cdot C_{CAT}}{\sum_{i=1}^{\mathcal{R}} \max(\Delta m_i, 0) \cdot C_{m_i}} \tag{6}$$

where $\Delta\texttt{CAT}$ denotes the abundance of CAT produced, C_{CAT} denotes carbon number of CAT, $\mathcal{R}$ denotes the number of reactants, Δm_i denotes the amount of the i^{th} reactant consumed (never allowed to be negative), and C_{m_i} denotes the carbon number of the i^{th} reactant. This analysis was extended to the accumulation of amino acids, organic acids, and other byproducts, to create a complete carbon balance through the network (Fig. 6A).

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

$$\text{Efficiency} = \frac{\Delta\texttt{CAT} \cdot \left(2\left(\text{ATP}_{\text{TX}} + \text{CTP}_{\text{TX}} + \text{GTP}_{\text{TX}} + \text{UTP}_{\text{TX}}\right) + 2 \cdot \text{ATP}_{\text{TL}} + \text{GTP}_{\text{TL}}\right)}{\Delta\texttt{GLC} \cdot \text{ATP}_{\text{GLC}}} \tag{7}$$

where ATP_{TX}, CTP_{TX}, GTP_{TX}, UTP_{TX} denote the stoichiometric coefficients of each energy species for CAT transcription, ATP_{TL}, GTP_{TL} denote the stoichiometric coefficients of ATP and GTP for CAT translation, $\Delta\texttt{GLC}$ denotes the glucose consumption, equal to the initial minus the final glucose concentration, and ATP_{GLC} denotes the equivalent ATP number for glucose. ATP_{TX} = 176, CTP_{TX} = 144, GTP_{TX} = 151, UTP_{TX} = 189, ATP_{TL} = 219, GTP_{TL} = 438, ATP_{GLC} = 15. This analysis was also extended to the accumulation of amino acids, organic acids, and other byproducts to create a complete energy balance through the network (Fig. 6B). Equivalent ATP numbers for glucose, amino acids, and organic acids were calculated from the network stoichiometry.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

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

Acknowledgements

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

Funding

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

References

1. Jewett MC, Calhoun KA, Voloshin A, Wuu JJ, Swartz JR. An integrated cell-free metabolic platform for protein production and synthetic biology. Mol Syst Biol. 2008;4:220. doi:10.1038/msb.2008.57.

2. Matthaei JH, Nirenberg MW. Characteristics and stabilization of DNAase-sensitive protein synthesis in E. coli extracts. Proc Natl Acad Sci U S A. 1961;47:1580–8.

3. Nirenberg MW, Matthaei JH. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci U S A. 1961;47:1588–602.

4. Lu Y, Welsh JP, Swartz JR. Production and stabilization of the trimeric influenza hemagglutinin stem domain for potentially broadly protective influenza vaccines. Proc Natl Acad Sci U S A. 2014;111(1):125–30. doi:10.1073/pnas.1308701110.

5. Hodgman CE, Jewett MC. Cell-free synthetic biology: thinking outside the cell. Metab Eng. 2012;14(3):261–9. doi:10.1016/j.ymben.2011.09.002.

6. Pardee K, Slomovic S, Nguyen PQ, Lee JW, Donghia N, Burrill D, et al. Portable, On-Demand Biomolecular Manufacturing. Cell. 2016;167(1):248–59.e12. doi:10.1016/j.cell.2016.09.013.

7. Fredrickson AG. Formulation of structured growth models. Biotechnol Bioeng. 1976;18(10):1481–6. doi:10.1002/bit.260181016.

8. Domach MM, Leung SK, Cahn RE, Cocks GG, Shuler ML. Computer model for glucose-limited growth of a single cell of Escherichia coli B/r-A. Biotechnol Bioeng. 1984;26(3):203–16. doi:10.1002/bit.260260303.

9. Steinmeyer DE, Shuler ML. Structured model for Saccharomyces cerevisiae. Chem Eng Sci. 1989;44:2017–30.

10. Wu P, Ray NG, Shuler ML. A single-cell model for CHO cells. Ann N Y Acad Sci. 1992;665:152–87.

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

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

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

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

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

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

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

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

19. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol. 2007;3:119. doi:10.1038/msb4100162.

20. Hyduke DR, Lewis NE, Palsson BØ. Analysis of omics data with genome-scale models of metabolism. Mol Biosyst. 2013;9(2):167–74. doi:10.1039/c2mb25453k.

21. McCloskey D, Palsson BØ, Feist AM. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol Syst Biol. 2013;9:661. doi:10.1038/msb.2013.18.

22. Zomorrodi AR, Suthers PF, Ranganathan S, Maranas CD. Mathematical optimization applications in metabolic networks. Metab Eng. 2012;14(6):672–86. doi:10.1016/j.ymben.2012.09.005.

23. Calhoun KA, Swartz JR. An Economical Method for Cell-Free Protein Synthesis using Glucose and Nucleoside Monophosphates. Biotechnology Progress. 2005;21(4):1146–53. doi:10.1021/bp050052y.

24. Wayman JA, Sagar A, Varner JD. Dynamic Modeling of Cell-Free Biochemical Networks Using Effective Kinetic Models. Processes. 2015;3(1):138. doi:10.3390/pr3010138.

25. Swartz J. A PURE approach to constructive biology. Nature Biotechnology. 2001;19:732–3.

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

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

28. Chulavatnatol M, Atkinson DE. Phosphoenolpyruvate synthetase from Escherichia

coli. Effects of adenylate energy charge and modifier concentrations. J Biol Chem. 1973;248(8):2712–5.

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

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

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

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

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

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

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

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

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

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

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

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

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

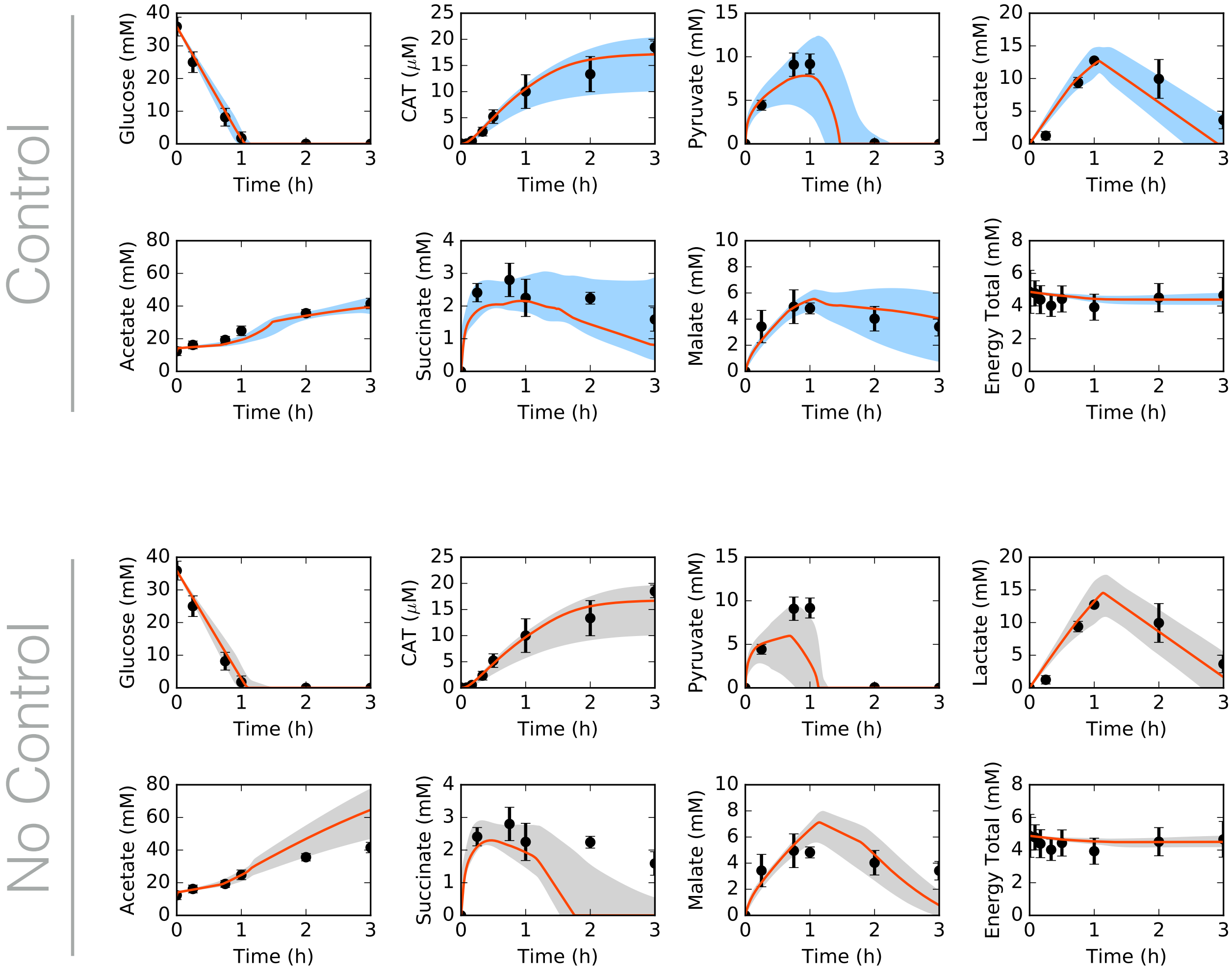

Fig. 1: Central carbon metabolism in the presence (top) and absence (bottom) of allosteric control, including glucose (substrate), CAT (product), and intermediates, as well as total concentration of energy species. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence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

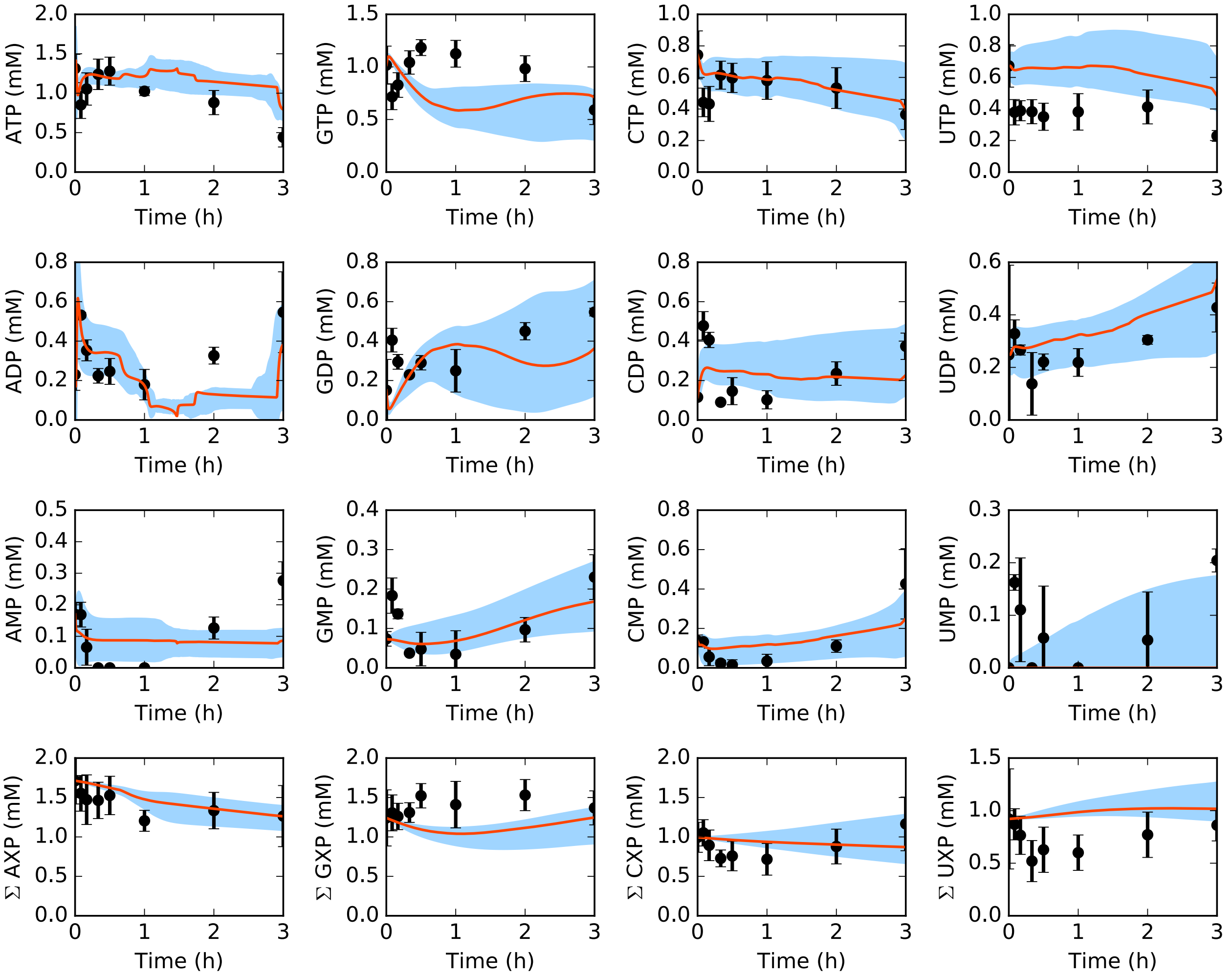

Fig. 2: Energy species and energy totals by base in the presence of allosteric control. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence interval (blue shaded region) over the ensemble of 100 sets.

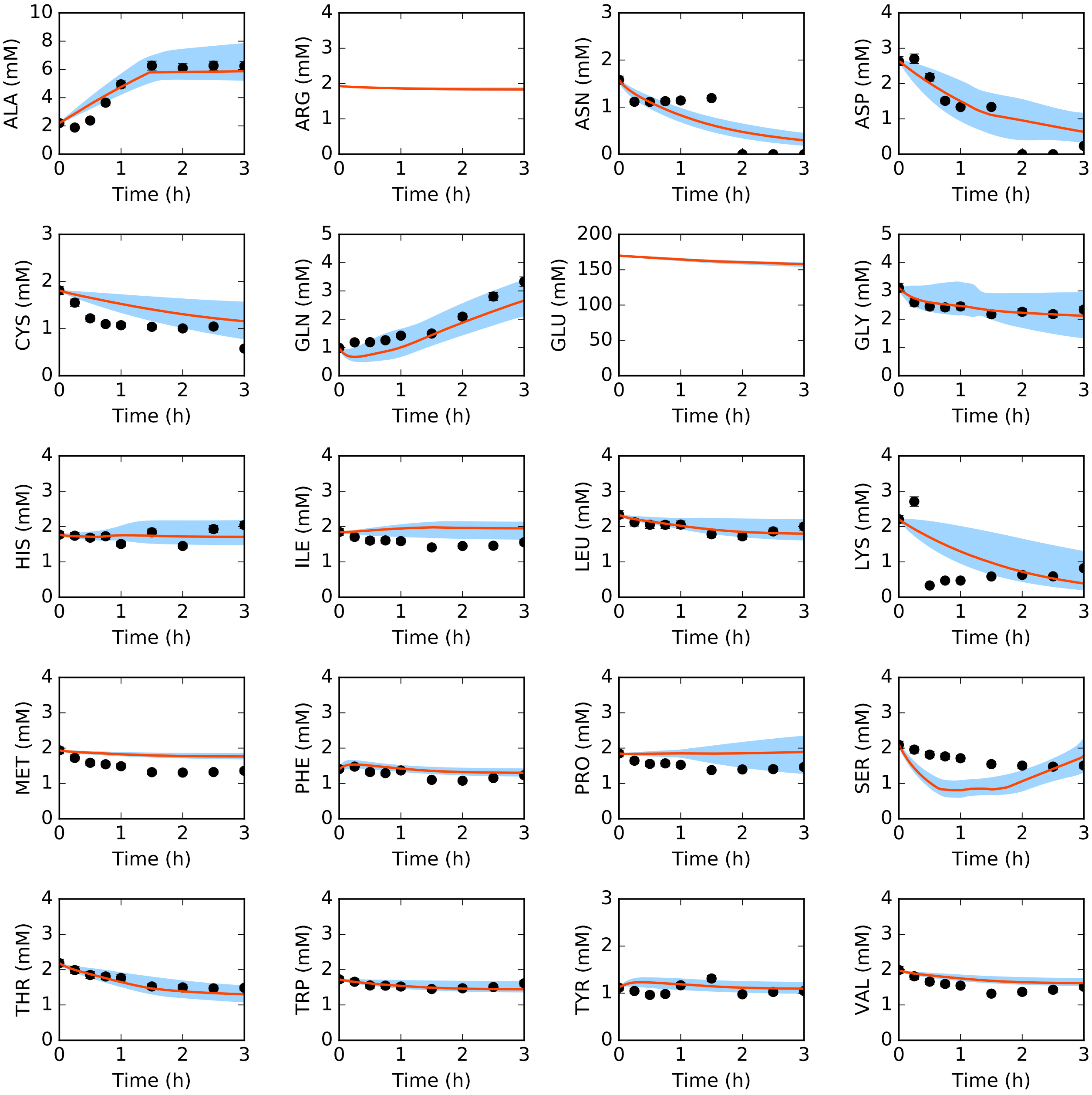

Fig. 3: Amino acids in the presence of allosteric control. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence interval (blue shaded region) over the ensemble of 100 sets.

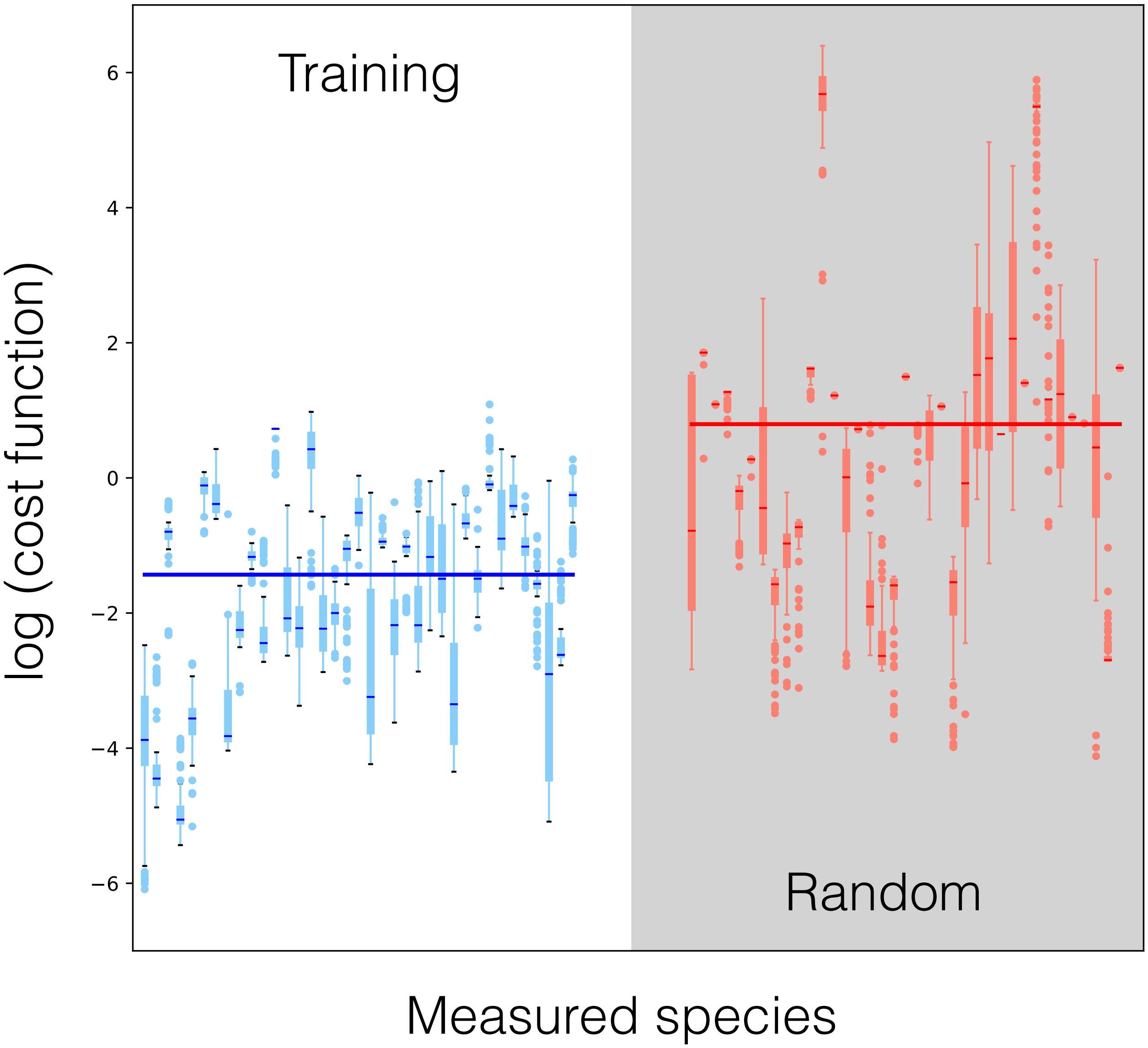

Fig. 4: Log of cost function across 37 datasets for data-trained ensemble (blue) and randomly generated ensemble (red, gray background). Median (bars), interquartile range (boxes), range excluding outliers (dashed lines), and outliers (circles) for each dataset. Median across all datasets (large bar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

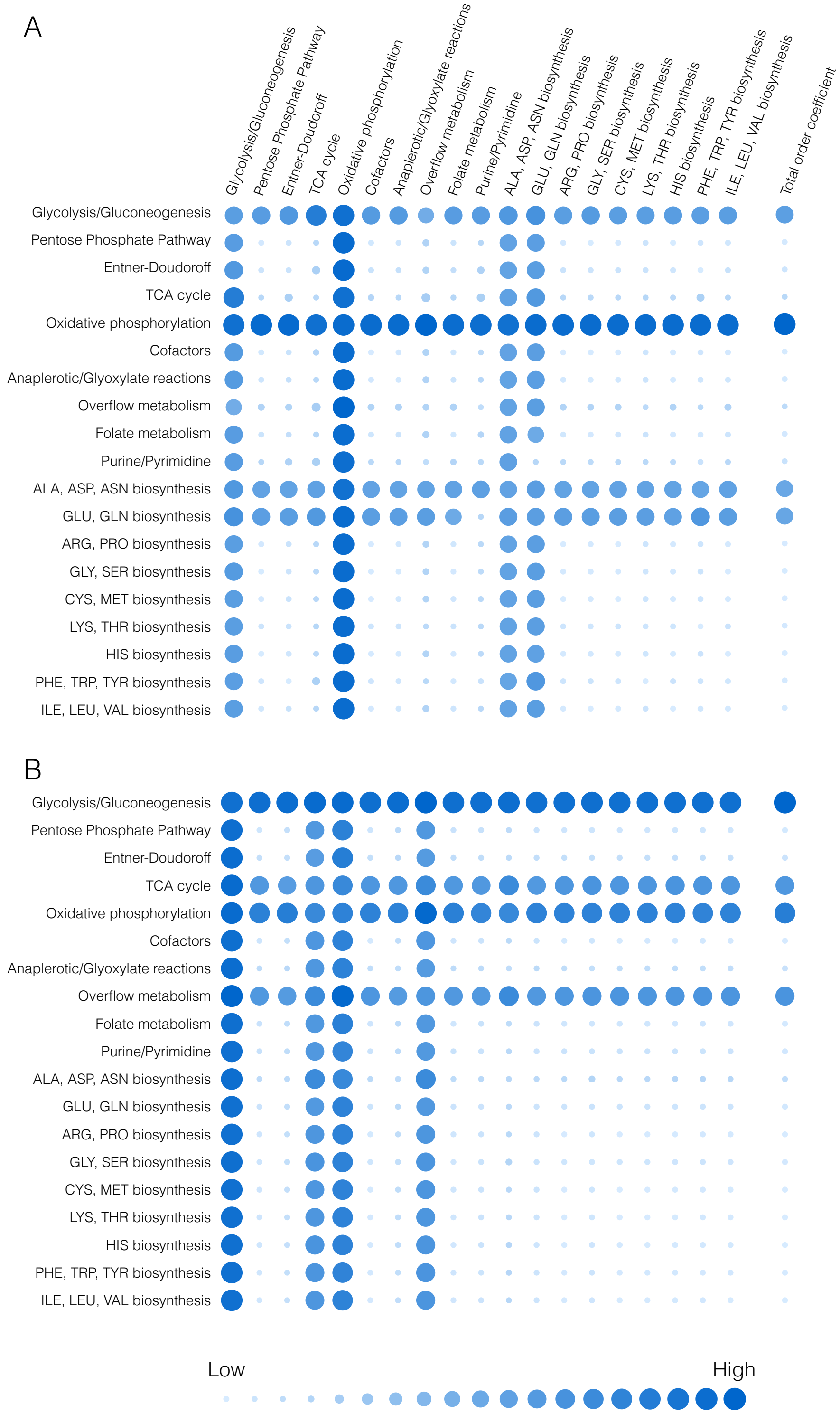

Fig. 5: Effect of group knockouts on system. A. Change in CAT productivity when one (diagonal) or two (off-diagonal) reaction groups are turned off. B. Change in system state (only species for which data exist) when one (diagonal) or two (off-diagonal) reaction groups are turned off. Total-order effect for each group calculated as the sum of first-order effect and all pairwise effects. Larger and darker circles represent greater effects.

Table 1: Reference values for reaction rate maxima (V$_{max}$) from literature. V$_{max}$ values calculated from original turnover numbers (K$_{cat}$) and characteristic enzyme concentration, extrapolated to cell-free using a dilution factor of 30. Characteristic rate maximum for all other reactions calculated as geometric mean of calculated rate maxima.

		Original value (μM)	**Cell-free equivalent (nM)**	**Reference**
Characteristic enzyme concentration		5	167	BNID 100735
Enzyme	**Reaction**	**K$_{cat}$ (min^{-1})**	**V$_{max}$ (mM/h)**	**Reference**
Serine dehydrase	R_ser_deg	10400	104	BNID 101119
Isocitrate dehydrogenase	R_icd	11900	119	BNID 101152
Lactate dehydrogenase	R_ldh	5800	58	BNID 101036
Aspartate transaminase	R_aspC, R_tyr, R_phe	25800	258	BNID 101108
Enolase	R_eno	13200	132	BNID 101028
Pyruvate kinase	R_pyk	25000	250	BNID 101029 BNID 101030
Malic enzyme	R_maeA, R_maeB	35400	354	BNID 101167
Phosphofructokinase	R_pfk	554400	5544	BNID 104955
Malate dehydrogenase	R_mdh	33000	330	BNID 101163
Citrate Synthase	R_gltA	42000	420	BNID 101149
6PG dehydrogenase	R_zwf, R_pgl, R_gnd	3200	32	BNID 101048
Succinate dehydrogenase	R_sdh	121	1.21	BNID 101162
Succinyl-coA synthetase	R_sucCD	4700	47	BNID 101158
3PGA dehydrogenase	R_gpm	1100	11	BNID 101135
PEP carboxylase	R_ppc	35400	354	BNID 101139
3PGA kinase	R_pgk	4300	43	BNID 101016
Characteristic rate maximum			110	
Transcription/Translation				
tRNA charging	0.03	14040	4212	BNID 104980

Table 2

	Volume (L)	Reference
Cellular volume	3.881e-16	XXX

	Concentration (mM)	Reference
CAT gene	5.19e-6	XXX
RNA polymerase	6.75e-5	[41]
Ribosome	2e-3	[41]
tRNA	1.6	BNID 108611
NAD	1.47	XXX
NADH	0.1	XXX
NADP	0.195	XXX
NADPH	0.062	XXX
G6P	3.48	XXX
F6P	0.6	XXX
FDP	0.272	XXX
DHAP	0.167	XXX
G3P	0.218	XXX
13DPG	0.008	XXX
3PG	2.13	XXX
2PG	0.399	XXX
6PGL	0.808	XXX
6PGC	0.808	XXX
RU5P	0.111	XXX
XU5P	0.138	XXX
R5P	0.398	XXX
S7P	0.276	XXX
E4P	0.098	XXX
NH3	0.052425	XXX
ARG	1.9326	XXX
H2O2	0.052425	XXX
HCO3	0.052425	XXX

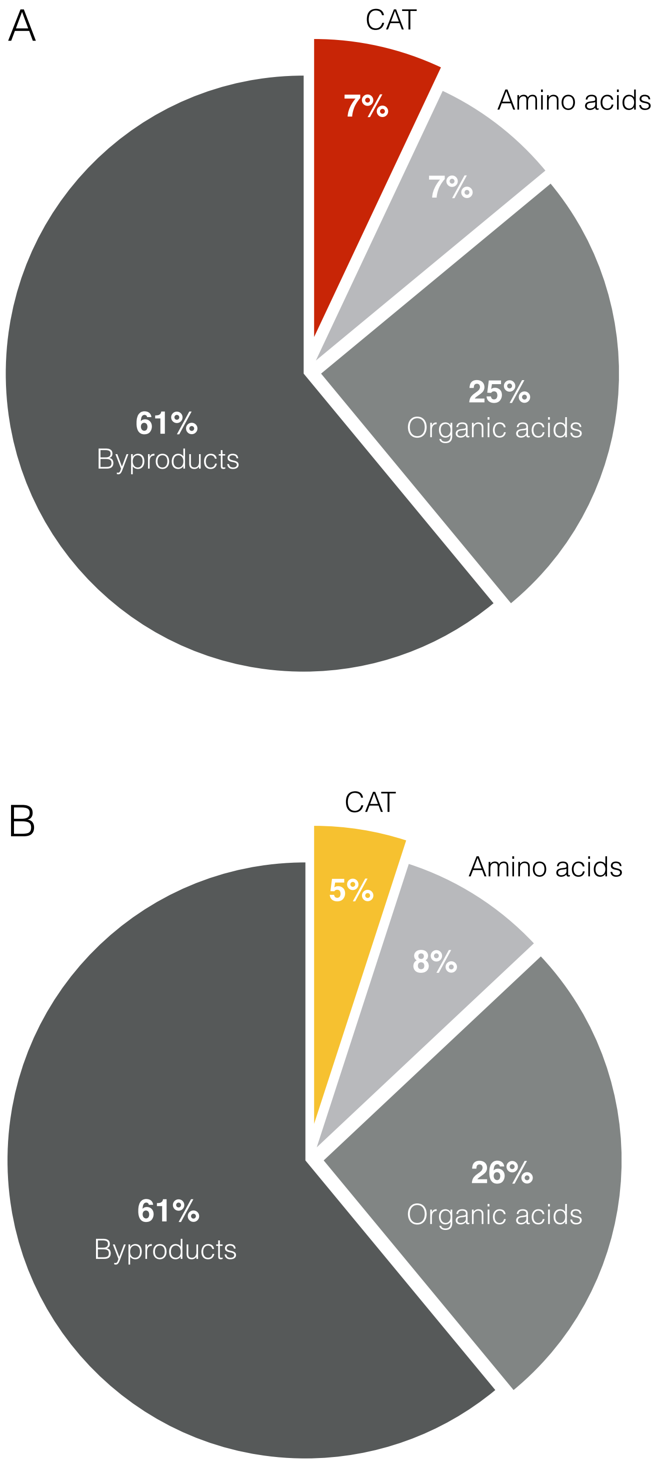

Fig. 6: Carbon and energy balances for the best-fit set. A. Carbon moles produced as CAT, amino acids (alanine and glutamine), organic acids (lactate, acetate, succinate, and malate), and other byproducts including carbon dioxide, as percentages of total carbon consumption (glucose and all other amino acids). B. Energy cost of CAT production, accumulation of amino acids (alanine and glutamine), accumulation of organic acids (lactate, acetate, succinate, and malate), and other byproducts, as percentages of total energy utilization from glucose. Energy costs calculated in terms of equivalent ATP molecules.