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

Nicholas Horvath, Michael Vilkhovoy, Joseph A. Wayman^a, Kara
Calhoun^b, James Swartz^b, Jeffrey D. Varner

*Robert Frederick Smith School of Chemical and Biomolecular Engineering, Cornell
University, Ithaca, NY 14853*

^a*School of Applied and Engineering Physics, Cornell University, Ithaca, NY 14853*

^b*School of Chemical Engineering, Stanford University, Stanford, CA 94395*

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 a training dataset for the cell-free production of a protein product, chloramphenicol acetyltransferase (CAT). The dataset consisted of measurements of glucose, organic acids, energy species, amino acids, and CAT. The ensemble accurately predicted these measurements, especially those of the central carbon metabolism. We then used the trained model to evaluate the optimality of protein production. CAT was produced with an energy efficiency of 12%, suggesting that the process 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, the translation rate is shown to be more important to productivity than the transcription rate. Finally, 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 use kinetic modeling to predict dynamic protein production in a cell-free *E. coli* system, and should provide a foundation for genome scale, dynamic modeling of cell-free *E. coli* protein synthesis.

Keywords: Biochemical engineering, cell-free protein synthesis, kinetic modeling

1 Introduction

2 Cell-free protein expression has become a widely used research tool in
 3 systems and synthetic biology, and a promising technology for personalized
 4 point of use biotechnology [1]. Cell-free systems offer many advantages for
 5 the study, manipulation and modeling of metabolism compared to *in vivo*
 6 processes. Central amongst these, is direct access to metabolites and the
 7 biosynthetic machinery without the interference of a cell wall, or complica-
 8 tions associated with cell growth. This allows us to interrogate (and po-
 9 tentially manipulate) the chemical microenvironment while the biosynthetic
 10 machinery is operating, potentially at a fine time resolution. Cell-free pro-
 11 tein synthesis (CFPS) systems are arguably the most prominent examples
 12 of cell-free systems used today [2]. However, CFPS is not new; CFPS in
 13 crude *E. coli* extracts has been used since the 1960s to explore fundamental
 14 biological mechanisms. For example, Matthaei and Nirenberg used *E. coli*
 15 cell-free extract in ground-breaking experiments to decipher the sequencing
 16 of the genetic code [3, 4]. Spirin and coworkers later improved protein pro-
 17 duction in cell free extracts by continuously exchanging reactants and prod-
 18 ucts; however, while these extracts could run for tens of hours, they could
 19 only synthesize a single product and were energy limited [5]. More recently,
 20 energy and cofactor regeneration in CFPS has been significantly improved;
 21 for example ATP can be regenerated using substrate level phosphorylation
 22 [6] or even oxidative phosphorylation [2]. Today, cell-free systems are used
 23 in a variety of applications ranging from therapeutic protein production [7]
 24 to synthetic biology [8, 1]. Moreover, there are also several CFPS technol-
 25 ogy platforms, such as the PANOx-SP and Cytomin platforms developed by
 26 Swartz and coworkers [9, 2], and the TX/TL platform of Noireaux [10]. How-
 27 ever, if CFPS is to become a mainstream technology for applications such as
 28 point of care biomanufacturing, we must first understand the performance
 29 limits of these systems, and eventually optimize their yield and productivity.
 30 A critical tool towards this goal is the development of a CFPS mathematical
 31 model.

32 Mathematical modeling has long contributed to our understanding of
 33 metabolism [11]. 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 [12]. 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 [13]. Shuler and coworkers generated many single cell kinetic models, including single cell models of eukaryotes [14, 15], minimal cell architectures [16], and DNA sequence based whole-cell models of *E. coli* [17]. More recent studies have extended the approach to integrate disparate models of cellular processes in *M. genitalium* [18], describe dozens of mutant strains in *E. coli* with a single kinetic model [19], and identify industrially useful target enzymes in *E. coli* to improve 1,4-butanediol production [20]. However, cell-free genome scale kinetic models of industrially 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 from an early glucose powered Cytomin *E. coli* cell-free extract. While cell-free technology has evolved considerably since these measurements were taken, developing a model using a previous generation CFPS platform offers several unique advantages. First and foremost, is the ability to directly compare the different improvements established by purely experimental means, to those estimated from a mathematical model. The CFPS model equations were formulated using the hybrid cell-free modeling framework of Wayman et al. [21], which integrates traditional kinetic modeling with a logical rule-based description of allosteric regulation. Model parameters were estimated from measurements of glucose, organic acids, energy species, amino acids, and the protein product, chloramphenicol acetyltransferase (CAT) over the course of a three hour protein synthesis reaction. A constrained Markov Chain Monte Carlo (MCMC) approach was used to minimize the squared difference between model simulations and experimental measurements, where a plausible range for each kinetic parameter was established from BioNumbers [22]. The ensemble of parameter sets described the training data with a median cost greater than two orders of magnitude smaller than a population of random parameter sets constructed using the same literature parameter constraints. We then used the ensemble of kinetic models to analyze the performance of the CFPS system, and to estimate the pathways most important to protein production. We calculated that CAT was produced with an energy efficiency of 12%, suggesting that much of the energy resources for

72 protein synthesis were diverted to non-productive pathways. By knocking
 73 out metabolic enzymes in groups, we showed that metabolism and protein
 74 production in particular depended upon oxidative phosphorylation and gly-
 75 colysis /gluconeogenesis. Taken together, this study provides a foundation
 76 for sequence specific genome scale, dynamic modeling of cell-free *E. coli* pro-
 77 tein synthesis.

78 Results

79 The cell-free *E. coli* metabolic network was constructed by removing
 80 growth associated reactions from the *iAF1260* reconstruction of K-12 MG1655
 81 *E. coli* [23], and by adding reactions describing chloramphenicol acetyltrans-
 82 ferase (CAT) biosynthesis (Fig. 1). In addition, reactions that were knocked
 83 out in the host strain used to prepare the extract were removed from the net-
 84 work (Δ speA, Δ tnaA, Δ sdaA, Δ sdaB, Δ gshA, Δ tonA, Δ endA). Lastly, we
 85 added the transcription and translation template reactions of Allen and Pals-
 86 son for the specific proteins of interest [24]. The metabolic network, which
 87 contained XX metabolites and YY reactions, is available in the supplemental
 88 materials. The dynamic CFPS model equations were formulated using the
 89 hybrid cell-free modeling framework of Wayman et al. [21]. An ensemble
 90 of model parameter sets ($N = 3,000$) was estimated from measurements of
 91 glucose, CAT, organic acids (pyruvate, lactate, acetate, succinate, malate),
 92 energy species (A(x)P, G(x)P, C(x)P, U(x)P), and 18 of the 20 proteinogenic
 93 amino acids [25] using a constrained Markov Chain Monte Carlo (MCMC)
 94 approach.

95 The MCMC algorithm minimized the squared difference (residual) be-
 96 tween the training data and model simulations starting from an initial pa-
 97 rameter set assembled from literature and inspection. Bounds on permis-
 98 sible parameter values were established using studies from the BioNumbers
 99 database [22]. For each newly generated parameter set, we re-solved the bal-
 100 ance equations and calculated the cost function; all sets with a lower cost
 101 (and some with higher cost) were accepted into the ensemble. Parameter
 102 sets were also required to meet strict ordinary differential equation solver
 103 tolerances, to ensure numerical stability. Approximately $N = 3,000$ sets were
 104 accepted into an initial ensemble; $N = 100$ sets were then selected based
 105 upon error for the final ensemble. The final ensemble had a mean Pearson
 106 correlation coefficient of 0.78; this suggested parameter sets were not over
 107 sampled in the region of a local minimum. The median maximum reaction
 108 rate (V_{max}) across the ensemble was 11.6 mM/h, assuming a total cell-free
 109 enzyme concentration of 170 nM. This V_{max} corresponded to a median cat-
 110 alytic rate of 19 s^{-1} across the ensemble; this was in agreement with the 13.7
 111 s^{-1} median catalytic rate found by Bar-Even and coworkers [26]. The median
 112 enzyme activity decay constant was 0.0045 h^{-1} , corresponding to an enzyme
 113 activity half life of 6 days. The median saturation constant was 1.0 mM; this
 114 is within one order of magnitude of the 130 μM reported by Bar-Even and

coworkers. Lastly, both the median control gain parameter, and the control order parameter in the ensemble were order 1. While the maximum reaction rates of the ensemble were distributed evenly across the allowed range (Fig. S1A), the saturation constants were clustered around the upper and lower bounds (Fig. S1B).

The ensemble of kinetic CFPS models captured the time evolution of protein biosynthesis, and the consumption and production of organic acid, amino acid and energy species. Central carbon metabolites (Fig. 2, top), energy species (Fig. 4), 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 established from literature (Fig. 5); thus, we have confidence in the predictive capability of the estimated parameters. For 29 of the 37 measurements in the training dataset, the mean Akaike information criterion (AIC) of the ensemble was lower than that of the random sets, signifying a better fit of the data (Table 3). For the other 8 measurements, the random AIC was lower than the ensemble by an amount less than the standard deviation of either the random AIC or ensemble AIC (with the exception of isoleucine, which was quite close: $\sigma_{\text{AIC}}^{\text{Rand}} = 4.8$, $\mu_{\text{AIC}}^{\text{Rand}} - \mu_{\text{AIC}}^{\text{Ens}} = -5.0$). Taken together, these results suggested that the parameter ensemble modeled cell free metabolism and protein production, significantly better than if sampled randomly, not just overall but for the majority of individual measurements.

The model captured the biphasic time course of CAT production. During the first hour glucose powered protein production, and CAT was produced at 8 $\mu\text{M}/\text{h}$; subsequently, pyruvate and lactate reserves were consumed to power metabolism, and CAT was produced less quickly at 5 $\mu\text{M}/\text{h}$. Allosteric control was important to central carbon metabolism, especially pyruvate, acetate, and succinate (Fig. 2, bottom). The difference between the allosteric control and no-control cases was mostly seen in the second phase of CAT production, following glucose exhaustion. Specifically, pyruvate, succinate, and malate consumption and acetate accumulation increased following glucose exhaustion without the allosteric control mechanisms. The rate of acetate accumulation increased by 172%, while the rates of malate, pyruvate, and lactate consumption increased by 146%, 82%, and 9%, respectively. Succinate went from accumulating slightly in the second phase, in the presence of allosteric control, to being fully consumed. However, CAT production was robust to the removal of allosteric control, as seen in both the fits against

153 data and the metabolic fluxes (see supplementary information). While ATP
154 generation varied when allosteric control was removed, ATP expenditure to-
155 ward CAT production did not. Most of the fluxes that differed between the
156 two cases involved PEP and pyruvate, which directly participated in many of
157 the reactions modulated by allosteric control. Taken together, the ensemble
158 of kinetic models was consistent with time series measurements of the cell
159 free production of a model protein. Although the ensemble described the
160 experimental data, it was unclear which kinetic parameters and pathways
161 most influenced CAT production. To explore this question, we performed
162 reaction group knockout analysis.

163 The importance of CFPS pathways was estimated using pathway group
164 knockout analysis (Fig. 7). The metabolic network was divided into 19 re-
165 action groups, spanning central carbon metabolism, energetics, and amino
166 acid biosynthesis. The response in the productivity or overall system state
167 was calculated for single or pairwise deletion of each of these reaction groups.
168 Lastly, the overall effect of the deletion of a pathway was estimated by sum-
169 ming the single and pairwise effects (summation across the columns of the
170 response array). Glycolysis/gluconeogenesis and oxidative phosphorylation
171 had the greatest effect on both productivity and system state. This supports
172 previous studies that have suggested oxidative phosphorylation is occurring
173 in a cell-free system [2]; Jewett and coworkers observed a decrease in CAT
174 yield, ranging from 1.5-fold to 4-fold, when inhibiting oxidative phosphory-
175 lation reactions in the Cytomim cell-free platform, using both pyruvate and
176 glutamate as substrates. CAT productivity was also affected by two sectors
177 of amino acid biosynthesis: alanine/aspartate/asparagine, and glutamate/
178 glutamine biosynthesis. This was consistent with aspartate, glutamate, and
179 glutamine being key reactants in the biosynthesis of many other amino acids,
180 all of which are required for CAT synthesis. Meanwhile, the TCA cycle and
181 overflow metabolism (which included acetyl-coA/acetate reactions and the
182 interconversion of pyruvate and lactate) also had a significant effect on the
183 system state. These reactions directly impacted key system species: succi-
184 nate and malate in the TCA cycle, and acetate, pyruvate, and lactate in
185 the overflow metabolism. In addition, the relative influence of transcription
186 and translation were interrogated via Sobol sampling [27]. Productivity was
187 seen to have a sensitivity of 0.43 ± 0.06 with respect to the maximum re-
188 action rate of transcription, and 0.66 ± 0.08 for the maximum reaction rate
189 of translation. Thus, translation was the limiting step of cell free protein
190 synthesis.

191 The energy efficiency of CAT production, as well as the sources of en-
 192 ergy generation and consumption, were tracked for the best-fit set. Energy
 193 efficiency was calculated as the ratio of transcription and translation rates
 194 (weighted by the associated ATP costs of each step) to the amount of ATP
 195 generated by all sources. During the first phase of protein production, with
 196 glucose as the substrate, CAT was produced with a productivity of 8 $\mu\text{M}/\text{h}$
 197 and an energy efficiency of 10%. Oxidative phosphorylation accounted for
 198 greater than 50% of the ATP generated during the rapid phase of protein
 199 production (Table 1). The organic acids that accumulated in the first phase
 200 (with the exception of acetate) were then utilized as substrates in the second
 201 phase, once glucose was depleted. We assumed the second phase of CAT pro-
 202 duction was powered largely by pyruvate; although malate was consumed in
 203 the second phase, it only accounted for 11% of substrate consumption. Fur-
 204 thermore, lactate is connected in the stoichiometry only to pyruvate. Thus,
 205 it is reasonable to consider the second phase as pyruvate-driven production.
 206 Interestingly, while this mode of protein production was slower (5 $\mu\text{M}/\text{h}$),
 207 it exhibited a higher energy efficiency (14%). Of the ATP generated, about
 208 half was observed to come from oxidative phosphorylation in each of the two
 209 phases of production (Table 1, R_atp). Another 30% was generated by glycol-
 210 ysis during the first phase (R_pgk, R_pyk), which decreased to approximately
 211 20% following glucose exhaustion. However, glycolysis was also amongst the
 212 largest consumers of ATP during first phase of production (R_glk_atp, R_pfk)
 213 (Table 2). The TCA cycle (R_sucCD) contributed 3% of to the overall ATP
 214 generation in the first phase and 5% in the second. The hypothesis that
 215 pyruvate drives the second phase explains this; stores of accumulated pyru-
 216 vate can be converted to acetyl-CoA, as well as OAA (via PEP), and thus
 217 power the TCA cycle just as when glucose was available. Interestingly, ATP
 218 generation through acetate metabolism (R_ackA) increased from 12% in the
 219 first phase to 28% in the second. Amino acid degradation also contributes a
 220 negligible amount to energy production. While the efficiency of production
 221 was higher for the pyruvate-driven phase, it was still relatively low, sug-
 222 gesting that there is room for platform optimization. Taken together, this
 223 strengthens the importance of glycolysis and oxidative phosphorylation, and
 224 presents a trade-off between productivity and energy efficiency in CFPS.

Discussion

In this study, we developed an ensemble of kinetic cell-free protein synthesis (CFPS) models using dynamic metabolite measurements from an early glucose powered Cytomin *E. coli* cell-free extract. 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, to describe the time evolution of the CFPS reaction. The ensemble captured dynamic metabolite measurements over 2-orders of magnitude better than random parameter sets generated in the same region of parameter space. The ensemble captured the biphasic time course of CAT production, relying on glucose during the first hour and pyruvate and lactate following glucose exhaustion. Allosteric control was essential to the description of the organic acid trajectories; without allosteric control, pyruvate, lactate, succinate, and malate were predicted to be consumed more quickly following glucose exhaustion, to power CAT synthesis. Interestingly, CAT production was robust to the removal of allosteric control; because the amino acids and energy species that are reactants for CAT synthesis were also not affected by allosteric control. We then used the ensemble of kinetic models to analyze the performance of the CFPS system, and to estimate the pathways most important to protein production. We calculated that CAT was produced with an energy efficiency of 12%, suggesting that much of the energy resources for protein synthesis were diverted to non-productive pathways. By knocking out metabolic enzymes in groups, we showed that metabolism and protein production in particular depended upon oxidative phosphorylation and glycolysis /gluconeogenesis. Using the Sobol sampling technique we demonstrated the greater importance of translation rate than transcription. Taken together, this study provides a foundation for sequence specific genome scale, dynamic modeling of cell-free *E. coli* protein synthesis.

The ensemble of models quantitatively described the dynamic time evolution of the cell-free protein production. Thus, the model could serve as a surrogate to rationally design cell-free production processes to optimize production rate and yield. In analyzing the effect of reaction groups on CAT production and the system state, the regions of metabolism associated with substrate utilization and energy generation were the most important. Oxidative phosphorylation was vital, since it provides most of the energetic needs of CFPS. While it is unknown how active oxidative phosphorylation is compared to that of *in vivo* systems, our modeling approach suggested it

was critical to CFPS performance. 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 continued to accumulate as a byproduct, the other organic acids were consumed as secondary substrates after glucose was no longer available. Glutamate also served as a substrate throughout both phases, powering amino acid synthesis. These results confirm experimental findings 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 productivity is significantly lower ($5 \mu\text{M/h}$, as opposed to $8 \mu\text{M/h}$). This is in accordance with literature, where pyruvate provided a relatively slow but continuous supply of ATP [28]. However, the energy efficiency is slightly higher (14% as compared with 10%). Taken together, this shows CFPS can be designed towards a specified application, either requiring a slow stable energy source or faster production.

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. Having captured the experimental data, we investigated if CFPS performance could be further improved. We showed that the model predicts CAT production with an energy efficiency of 10% under glucose and 14% under pyruvate. The accumulation of glycolytic intermediates and byproducts such as acetate and carbon dioxide were responsible for this sub-optimal performance. If fluxes could be balanced such that intermediates were fully utilized, CAT production would increase. Knocking out sections of network metabolism revealed that glycolysis/gluconeogenesis and

300 oxidative phosphorylation were the most important to CAT production and
301 the system as a whole. Productivity was also heavily dependent on the syn-
302 thesis reactions of alanine, aspartate, asparagine, glutamate, and glutamine,
303 while TCA cycle and overflow reactions affected the system state. Taken
304 together, these findings represent the first dynamic model of *E. coli* cell-
305 free protein synthesis, an important step toward a functional genome scale
306 description of cell free systems.

307 Materials and Methods

308 *Formulation and solution of the model equations.*

309 We used ordinary differential equations (ODEs) to model the time evo-
310 lution of metabolite (x_i), scaled enzyme activity (ϵ_i), transcription (m) and
311 translation ($\mathcal{P}$) in an *E. coli* cell free metabolic network:

$$\frac{dx_i}{dt} = \sum_{j=1}^{\mathcal{R}} \sigma_{ij} r_j(\mathbf{x}, \epsilon, \mathbf{k}) \quad i = 1, 2, \dots, \mathcal{M} \quad (1)$$

$$\frac{d\epsilon_i}{dt} = -\lambda_i \epsilon_i \quad i = 1, 2, \dots, \mathcal{E} \quad (2)$$

$$\frac{dm}{dt} = \bar{r}_T u - \bar{r}_d \quad (3)$$

$$\frac{d\mathcal{P}}{dt} = \bar{r}_X \quad (4)$$

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

323 Metabolic reaction rates were written as the product of a kinetic term
324 ($\bar{r}_j$) and a control term (v_j), $r_j(\mathbf{x}, \mathbf{k}) = \bar{r}_j v_j$. We used multiple saturation
325 kinetics to model the reaction term $\bar{r}_j$:

$$\bar{r}_j = V_j^{max} \epsilon_i \prod_{s \in m_j^-} \frac{x_s}{K_{js} + x_s} \quad (5)$$

326 where V_j^{max} denotes the maximum rate for reaction j , ϵ_i denotes the scaled
327 enzyme activity which catalyzes reaction j , K_{js} denotes the saturation con-
328 stant for species s , in reaction j and m_j^- denotes the set of *reactants* for
329 reaction j .

330 The control term $0 \leq v_j \leq 1$ depended upon the combination of factors
 331 which influenced rate process j . For each rate, we used a rule-based approach
 332 to select from competing control factors. If rate j was influenced by $1, \dots, m$
 333 factors, we modeled this relationship as $v_j = \mathcal{I}_j(f_{1j}(\cdot), \dots, f_{mj}(\cdot))$ where
 334 $0 \leq f_{ij}(\cdot) \leq 1$ denotes a transfer function quantifying the influence of factor
 335 i on rate j . The function $\mathcal{I}_j(\cdot)$ is an integration rule which maps the output
 336 of regulatory transfer functions into a control variable. We used hill-like
 337 transfer functions and $\mathcal{I}_j \in \{\min, \max\}$ in this study [21]. We included
 338 17 allosteric regulation terms, taken from literature, in the CFPS model.
 339 PEP was modeled as an inhibitor for phosphofructokinase [29, 30], PEP
 340 carboxykinase [29], PEP synthetase [29, 31], isocitrate dehydrogenase [29,
 341 32], and isocitrate lyase/malate synthase [29, 32, 33], and as an activator for
 342 fructose-biphosphatase [29, 34, 35, 36]. AKG was modeled as an inhibitor
 343 for citrate synthase [29, 37, 38] and isocitrate lyase/malate synthase [29, 33].
 344 3PG was modeled as an inhibitor for isocitrate lyase/malate synthase [29, 33].
 345 FDP was modeled as an activator for pyruvate kinase [29, 39] and PEP
 346 carboxylase [29, 40]. Pyruvate was modeled as an inhibitor for pyruvate
 347 dehydrogenase [29, 41, 42] and as an activator for lactate dehydrogenase
 348 [43]. Acetyl CoA was modeled as an inhibitor for malate dehydrogenase [29].
 349 The symbol $\bar{r}_T$ denotes the transcription rate, u denotes a promoter spe-
 350 cific activation model, and $\bar{r}_d$ denotes the transcript degradation rate. The
 351 transcription rate was modeled as:

$$\bar{r}_T = k_{cat}^T \cdot R_T \left(\frac{G_P}{K_G^T + G_P} \right) \prod_{s \in m_T^-} \frac{x_s}{K_s^T + x_s} \quad (6)$$

352 where k_{cat}^T denotes the maximum transcription rate, R_T denotes the RNA
 353 polymerase concentration, G_P denotes the gene concentration, K_G^T denotes
 354 the gene saturation constant, K_s^T denotes the saturation constant for species
 355 s , and m_T^- denotes the set of *reactants* for transcription: ATP, GTP, CTP,
 356 UTP, and water. In this study, we considered only the T7 promoter; we have
 357 previously estimated $u \simeq 0.95$ for a T7 [REF-MIKE]. While transcription was
 358 modeled as saturating with respect to gene concentration, the gene was not
 359 considered a reactant in the stoichiometry as it was not consumed. Transcript
 360 degradation was modeled as first-order in transcript:

$$\bar{r}_d = k_d \cdot m \quad (7)$$

361 where k_d denotes the transcript degradation rate constant.

362 The symbol $\bar{r}_X$ denotes the translation rate, which was modeled as:

$$\bar{r}_X = k_{cat}^X \cdot R_X \left(\frac{m}{K_{mRNA}^X + m} \right) \prod_{s \in m_X^-} \frac{x_s}{K_s^X + x_s} \quad (8)$$

363 where k_{cat}^X denotes the maximum translation rate, R_X denotes the ribo-
 364 some concentration, m denotes the transcript concentration, K_{mRNA}^X denotes
 365 the transcript saturation constant, K_s^X denotes the saturation constant for
 366 species s , and m_X^- denotes the set of *reactants* for translation: GTP, wa-
 367 ter, and the 20 species representing tRNA charged with amino acids. While
 368 translation was modeled as saturating with respect to transcript concentra-
 369 tion, the transcript was not considered a reactant in the stoichiometry as it
 370 is not consumed.

371 *Estimation of kinetic model parameters.*

372 We estimated an ensemble of kinetic parameter sets using a constrained
 373 Markov Chain Monte Carlo (MCMC) random walk strategy. We have used
 374 this technique previously to estimate numerically stable low-error parameter
 375 sets for signal transduction models [44, 45]. Starting from a small number
 376 of parameter sets estimated by inspection and literature, we calculated the
 377 cost function, equal to the sum-squared-error between experimental data and
 378 model predictions:

$$\text{cost} = \sum_{i=1}^{\mathcal{D}} \left[\frac{w_i}{\mathcal{Y}_i^2} \sum_{j=1}^{\mathcal{T}_i} \left(y_{ij} - x_{i|t(j)} \right)^2 \right] \quad (9)$$

379 where $\mathcal{D}$ denotes the number of datasets ($\mathcal{D} = 37$), w_i denotes the weight
 380 of the i^{th} dataset, $\mathcal{T}_i$ denotes the number of timepoints in the i^{th} dataset,
 381 $t(j)$ denotes the j^{th} timepoint, y_{ij} denotes the measurement value of the i^{th}
 382 dataset at the j^{th} timepoint, and $x_{i|t(j)}$ denotes the simulated value of the
 383 metabolite corresponding to the i^{th} dataset, interpolated to the j^{th} timepoint.
 384 Lastly, the cost function was scaled by the maximum experimental value in
 385 the i^{th} dataset, $\mathcal{Y}_i = \max_j (y_{ij})$. We then perturbed each model parameter
 386 between an upper and lower bound that varied by parameter type:

$$k_i^{new} = \min(\max(k_i \cdot \exp(a \cdot r_i), l_i), u_i) \quad i = 1, 2, \dots, \mathcal{P} \quad (10)$$

387 where $\mathcal{P}$ denotes the number of parameters ($\mathcal{P} = 815$), which includes 204
 388 maximum reaction rates (V^{max}), 204 enzyme activity decay constants, 548

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 collected using the BioNumbers database [22]. Transcription, translation, and mRNA degradation were bounded within a factor of two of their reference values. A characteristic cell-free enzyme concentration of 170 nM was calculated by diluting the one-tenth maximal concentration of *lacZ* (5 μ M, BNID 100735) by a cell-free dilution factor of 30. This enzyme level was then used to calculate rate maxima from turnover numbers for various enzymes from BioNumbers (Table 4). Rate maxima were bounded within one order of magnitude of the reference value where available; all other rate maxima were bounded within two orders of magnitude of the geometric mean of the available values. Enzyme activity decay constants were bounded between 0 and 1 h⁻¹, corresponding to half lives of 42 minutes and infinity. Saturation constants were bounded between 0.0001 and 10 mM. Control gain parameters were bounded between 0.05 and 10 (dimensionless), while order parameters were bounded between 0.02 and 10 (dimensionless).

For each newly generated parameter set, we re-solved the balance equations and calculated the cost function. All sets with a lower cost were accepted into the ensemble. Sets with a higher cost were also accepted into the ensemble, if they satisfied the acceptance constraint:

$$\mathcal{R}_{0,1}^{uniform} < \exp\left(-\alpha \cdot \frac{\text{cost}_{\text{new}} - \text{cost}}{\text{cost}}\right) \quad (11)$$

where $\mathcal{R}_{0,1}^{uniform}$ denotes a random number taken from a uniform distribution between 0 and 1, **cost** denotes the cost of the current parameter set, **cost_{new}** denotes the cost of the new parameter set, and α denotes a tunable parameter to control the tolerance to high-error sets. A total of 3,875 sets were accepted into the initial ensemble, from which we selected $N = 100$ with minimal error for the final ensemble.

Lastly, a random ensemble of 100 parameter sets was generated within the same parameter bounds as the trained ensemble. The randomized parameter sets were generated using a Monte Carlo approach: each parameter was taken from a uniform distribution constructed between its upper and lower

423 bounds. The model equations were then solved and the cost function, and
 424 the Akaike information criterion (AIC) were calculated for each of the 37
 425 separate experimental datasets.

426 *Reaction group knockouts.*

427 The metabolic network was divided into 19 reaction groups: glycoly-
 428 sis/gluconeogenesis, pentose phosphate, Entner-Doudoroff, TCA cycle, ox-
 429 idative phosphorylation, cofactor reactions, anaplerotic/glyoxylate reactions,
 430 overflow metabolism, folate synthesis, purine/pyrimidine reactions, alanine/
 431 aspartate/asparagine synthesis, glutamate/glutamine synthesis, arginine/proline
 432 synthesis, glycine/serine synthesis, cysteine/methionine synthesis, threonine/
 433 lysine synthesis, histidine synthesis, tyrosine/tryptophan/phenylalanine syn-
 434 thesis, and valine/leucine/isoleucine synthesis. Each reaction group and pair
 435 of reaction groups were removed and the model was re-solved; the CAT pro-
 436 ductivity was then calculated and subtracted from that of the base case (no
 437 knockouts):

$$P_{ii} = |\Delta\text{CAT} - \Delta\text{CAT}_{\Delta R_i}| \quad (12)$$

$$P_{ij} = |\Delta\text{CAT} - \Delta\text{CAT}_{\Delta R_i \Delta R_j}| \quad (13)$$

$$P_i^{\text{total}} = P_{ii} + \sum_j P_{ij} \quad (14)$$

438 where P_{ii} denotes the first-order productivity knockout effect for reaction
 439 group i , P_{ij} denotes the pairwise productivity knockout effect for reaction
 440 groups i and j , P_i^{total} denotes the total-order productivity knockout effect for
 441 reaction group i , ΔCAT denotes the base case CAT productivity, $\Delta\text{CAT}_{\Delta R_i}$ de-
 442 notes the CAT productivity when reaction group i is knocked out, $\Delta\text{CAT}_{\Delta R_i \Delta R_j}$
 443 denotes the CAT productivity when reaction groups i and j are knocked out,
 444 and $|x|$ denotes the absolute value of x . The system state, defined as the
 445 model predictions for all species for which experimental data exists, was also
 446 recorded for each knockout and compared to the base case:

$$S_{ii} = \|\mathbf{x}^{\text{data}} - \mathbf{x}_{\Delta R_i}^{\text{data}}\|_2 \quad (15)$$

$$S_{ij} = \|\mathbf{x}^{\text{data}} - \mathbf{x}_{\Delta R_i \Delta R_j}^{\text{data}}\|_2 \quad (16)$$

$$S_i^{\text{total}} = S_{ii} + \sum_j S_{ij} \quad (17)$$

where S_{ii} denotes the first-order system state knockout effect for reaction group i , S_{ij} denotes the pairwise system state knockout effect for reaction groups i and j , S_i^{total} denotes the total-order system state knockout effect for reaction group i , $\mathbf{x}^{data}$ denotes the base-case system state, $\mathbf{x}_{\Delta R_i}^{data}$ denotes the system state when reaction group i is knocked out, $\mathbf{x}_{\Delta R_i \Delta R_j}^{data}$ denotes the system state when reaction groups i and j are knocked out, and $\|x\|_2$ denotes the l^2 norm of x . In order to not dominate the colorbar, the total-order knockout effects were normalized to the same ranges as the main arrays (first-order and pairwise effects).

Sensitivity of CAT productivity to transcription and translation.

The catalytic rates of transcription and translation were sampled within one order of magnitude on each side from the best-fit values. The parameter bounds were set as the base-10 logarithms of the upper and lower bound for each rate; then, 10 was taken to the power of each parameter sample to obtain the catalytic rates:

$$k_{cat}^{T,sample} \in \left[\log_{10} \left(k_{cat}^{T,bf} / 10 \right), \log_{10} \left(k_{cat}^{T,bf} * 10 \right) \right] \quad (18)$$

$$k_{cat}^{X,sample} \in \left[\log_{10} \left(k_{cat}^{X,bf} / 10 \right), \log_{10} \left(k_{cat}^{X,bf} * 10 \right) \right] \quad (19)$$

$$\Delta CAT = f \left(10^{k_{cat}^{T,sample}}, 10^{k_{cat}^{X,sample}} \right) \quad (20)$$

where $k_{cat}^{T,sample}$ denotes the sample of the transcription catalytic rate, $k_{cat}^{X,sample}$ denotes the sample of the translation catalytic rate, $k_{cat}^{T,bf}$ denotes the best-fit value of the transcription catalytic rate, and $k_{cat}^{X,bf}$ denotes the best-fit value of the translation catalytic rate. The sampling was performed using the Sensitivity Analysis Library in Python (Numpy) with 3000 samples [46].

Calculation of energy efficiency.

Energy efficiency was calculated as the ratio of transcription and translation (weighted by the appropriate energy species coefficients) to ATP gen-

eration:

$$\text{Efficiency} = \frac{\Delta_{\tau}\text{mRNA} \cdot \alpha_T + \Delta_{\tau}\text{CAT} \cdot \alpha_X}{\sum_{j \in \{R_{ATP}\}} \int_{\tau} \sigma_j^{ATP} \bar{r}_j} \quad (21)$$

$$\alpha_T = 2 \cdot (\text{ATP}_T + \text{CTP}_T + \text{GTP}_T + \text{UTP}_T) \quad (22)$$

$$\alpha_X = 2 \cdot \text{ATP}_X + \text{GTP}_X \quad (23)$$

where $\Delta_{\tau}\text{mRNA}$ denotes the net accumulation of mRNA in phase τ (first, second, or overall), $\Delta_{\tau}\text{CAT}$ denotes the net accumulation of protein in phase τ , α_T denotes the energy cost of transcription, α_X denotes the energy cost of translation, R_{ATP} denotes the set of ATP-producing reactions, and σ_j^{ATP} denotes the ATP coefficient for reaction j . ATP_T , CTP_T , GTP_T , UTP_T denote the stoichiometric coefficients of each energy species for transcription, and ATP_X , GTP_X denote the stoichiometric coefficients of ATP and GTP for translation. During transcription and tRNA charging, triphosphate molecules are consumed with monophosphates as byproducts; this is the reason for the factors of 2 on ATP_T , CTP_T , GTP_T , UTP_T , and ATP_X .

Availability of model code.

The cell free model equations, and the parameter estimation procedure, were implemented in the Julia programming language. The model equations were solved using the CVODE solver of the SUNDIALS suite [47], with an absolute tolerance and relative tolerance of $1e^{-9}$; any sets exhibiting CVODE errors were discarded. Thus, the numerical stability of all parameters in the ensemble was ensured. The model code and parameter ensemble is freely available under an MIT software license and can be downloaded from <http://www.varnerlab.org>.

490 **Competing interests**

491 The authors declare that they have no competing interests.

492 **Author's contributions**

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

497 **Acknowledgements**

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

500 **Funding**

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

- 506 [1] K. Pardee, S. Slomovic, P. Q. Nguyen, J. W. Lee, N. Donghia, D. Bur-
507 rill, T. Ferrante, F. R. McSorley, Y. Furuta, A. Vernet, M. Lewandowski,
508 C. N. Boddy, N. S. Joshi, J. J. Collins, Portable, on-demand biomolec-
509 ular manufacturing, *Cell* 167 (2016) 248–59.e12.
- 510 [2] M. C. Jewett, K. A. Calhoun, A. Voloshin, J. J. Wu, J. R. Swartz,
511 An integrated cell-free metabolic platform for protein production and
512 synthetic biology, *Mol Syst Biol* 4 (2008) 220.
- 513 [3] J. H. Matthaei, M. W. Nirenberg, Characteristics and stabilization of
514 dnaase-sensitive protein synthesis in e. coli extracts, *Proc Natl Acad Sci*
515 *U S A* 47 (1961) 1580–8.
- 516 [4] M. W. Nirenberg, J. H. Matthaei, The dependence of cell-free protein
517 synthesis in e. coli upon naturally occurring or synthetic polyribonu-
518 cleotides, *Proc Natl Acad Sci U S A* 47 (1961) 1588–602.
- 519 [5] A. Spirin, V. Baranov, L. Ryabova, S. Ovodov, Y. Alakhov, A contin-
520 uous cell-free translation system capable of producing polypeptides in
521 high yield, *Science* 242 (1988) 1162–1164.
- 522 [6] D.-M. Kim, J. R. Swartz, Regeneration of adenosine triphosphate from
523 glycolytic intermediates for cell-free protein synthesis, *Biotechnology*
524 *and Bioengineering* 74 (2001) 309–316.
- 525 [7] Y. Lu, J. P. Welsh, J. R. Swartz, Production and stabilization of the
526 trimeric influenza hemagglutinin stem domain for potentially broadly
527 protective influenza vaccines, *Proc Natl Acad Sci U S A* 111 (2014)
528 125–30.
- 529 [8] C. E. Hodgman, M. C. Jewett, Cell-free synthetic biology: thinking
530 outside the cell, *Metab Eng* 14 (2012) 261–9.
- 531 [9] M. C. Jewett, J. R. Swartz, Mimicking the escherichia coli cytoplasmic
532 environment activates long-lived and efficient cell-free protein synthesis,
533 *Biotechnology and Bioengineering* 86 (2004) 19–26.
- 534 [10] J. Garamella, R. Marshall, M. Rustad, V. Noireaux, The all e. coli tx-tl
535 toolbox 2.0: A platform for cell-free synthetic biology, *ACS Synth Biol*
536 5 (2016) 344–55.

- 537 [11] J. A. Wayman, J. D. Varner, Biological systems modeling of metabolic
538 and signaling networks, *Curr Opin Chem Eng* 2 (2013).
- 539 [12] A. G. Fredrickson, Formulation of structured growth models, *Biotechnol*
540 *Bioeng* 18 (1976) 1481–6.
- 541 [13] M. M. Domach, S. K. Leung, R. E. Cahn, G. G. Cocks, M. L. Shuler,
542 Computer model for glucose-limited growth of a single cell of *escherichia*
543 *coli* b/r-a, *Biotechnol Bioeng* 26 (1984) 203–16.
- 544 [14] D. Steinmeyer, M. Shuler, Structured model for *Saccharomyces cere-*
545 *visiae*., *Chem Eng Sci* 44 (1989) 2017–30.
- 546 [15] P. Wu, N. G. Ray, M. L. Shuler, A single-cell model for *cho* cells, *Ann*
547 *N Y Acad Sci* 665 (1992) 152–87.
- 548 [16] M. Castellanos, D. B. Wilson, M. L. Shuler, A modular minimal cell
549 model: purine and pyrimidine transport and metabolism, *Proc Natl*
550 *Acad Sci U S A* 101 (2004) 6681–6.
- 551 [17] J. C. Atlas, E. V. Nikolaev, S. T. Browning, M. L. Shuler, Incorporating
552 genome-wide dna sequence information into a dynamic whole-cell model
553 of *escherichia coli*: application to dna replication, *IET Syst Biol* 2 (2008)
554 369–82.
- 555 [18] J. R. Karr, J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs,
556 B. Bolival, N. Assad-Garcia, J. I. Glass, M. W. Covert, A whole-cell
557 computational model predicts phenotype from genotype, *Cell* 150 (2012)
558 389–401.
- 559 [19] A. Khodayari, C. D. Maranas, A genome-scale *Escherichia coli* ki-
560 netic metabolic model *k-ecoli457* satisfying flux data for multiple mutant
561 strains, *Nat Commun* 7 (2016) 13806.
- 562 [20] S. Andreozzi, A. Chakrabarti, K. C. Soh, A. Burgard, T. H. Yang,
563 S. Van Dien, L. Miskovic, V. Hatzimanikatis, Identification of metabolic
564 engineering targets for the enhancement of 1,4-butanediol production in
565 recombinant *E. coli* using large-scale kinetic models, *Metab. Eng.* 35
566 (2016) 148–159.

- 567 [21] J. A. Wayman, A. Sagar, J. D. Varner, Dynamic modeling of cell-free
568 biochemical networks using effective kinetic models, *Processes* 3 (2015)
569 138.
- 570 [22] R. Milo, P. Jorgensen, U. Moran, G. Weber, M. Springer, Bionumbers–
571 the database of key numbers in molecular and cell biology, *Nucleic Acids*
572 *Res* 38 (2009) 750–3.
- 573 [23] A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce,
574 P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, B. Ø. Palsson, A
575 genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655
576 that accounts for 1260 ORFs and thermodynamic information, *Mol Syst*
577 *Biol* 3 (2007) 121.
- 578 [24] T. E. Allen, B. Ø. Palsson, Sequence-based analysis of metabolic de-
579 mands for protein synthesis in prokaryotes, *J Theor Biol* 220 (2003)
580 1–18.
- 581 [25] M. Vilkhovoy, N. Horvath, C.-h. Shih, J. Wayman, K. Calhoun,
582 J. Swartz, J. Varner, Sequence specific modeling of *e. coli* cell-free pro-
583 tein synthesis, *bioRxiv* (2017).
- 584 [26] A. Bar-Even, E. Noor, Y. Savir, W. Liebermeister, D. Davidi, D. S.
585 Tawfik, R. Milo, The moderately efficient enzyme: Evolutionary and
586 physicochemical trends shaping enzyme parameters, *Biochemistry* 50
587 (2011).
- 588 [27] I. Sobol, Global sensitivity indices for nonlinear mathematical models
589 and their monte carlo estimates., *Math Comput Simulat* 55 (2001) 271–
590 80.
- 591 [28] J. Swartz, A pure approach to constructive biology, *Nat Biotechnol* 19
592 (2001) 732–3.
- 593 [29] O. Kotte, J. B. Zaugg, M. Heinemann, Bacterial adaptation through
594 distributed sensing of metabolic fluxes, *Mol Syst Biol* 6 (2010) 355.
- 595 [30] R. Cabrera, M. Baez, H. M. Pereira, A. Caniuguir, R. C. Garratt,
596 J. Babul, The crystal complex of phosphofructokinase-2 of *Escherichia*
597 *coli* with fructose-6-phosphate: kinetic and structural analysis of the
598 allosteric ATP inhibition, *J Biol Chem* 286 (2011) 5774–83.

- 599 [31] M. Chulavatnatol, D. E. Atkinson, Phosphoenolpyruvate synthetase
600 from *Escherichia coli*. Effects of adenylate energy charge and modifier
601 concentrations, *J Biol Chem* 248 (1973) 2712–5.
- 602 [32] T. Ogawa, K. Murakami, H. Mori, N. Ishii, M. Tomita, M. Yoshin,
603 Role of phosphoenolpyruvate in the NADP-isocitrate dehydrogenase and
604 isocitrate lyase reaction in *Escherichia coli*, *J Bacteriol* 189 (2007) 1176–
605 8.
- 606 [33] C. MacKintosh, H. G. Nimmo, Purification and regulatory properties
607 of isocitrate lyase from *Escherichia coli* ML308, *Biochem J* 250 (1988)
608 25–31.
- 609 [34] J. L. Donahue, J. L. Bownas, W. G. Niehaus, T. J. Larson, Purification
610 and characterization of glpX-encoded fructose 1, 6-bisphosphatase, a
611 new enzyme of the glycerol 3-phosphate regulon of *Escherichia coli*, *J*
612 *Bacteriol* 182 (2000) 5624–7.
- 613 [35] J. K. Hines, H. J. Fromm, R. B. Honzatko, Novel allosteric activation
614 site in *Escherichia coli* fructose-1,6-bisphosphatase, *J Biol Chem* 281
615 (2006) 18386–93.
- 616 [36] J. K. Hines, H. J. Fromm, R. B. Honzatko, Structures of activated
617 fructose-1,6-bisphosphatase from *Escherichia coli*. Coordinate regulation
618 of bacterial metabolism and the conservation of the R-state, *J Biol Chem*
619 282 (2007) 11696–704.
- 620 [37] D. S. Pereira, L. J. Donald, D. J. Hosfield, H. W. Duckworth, Active
621 site mutants of *Escherichia coli* citrate synthase. Effects of mutations on
622 catalytic and allosteric properties, *J Biol Chem* 269 (1994) 412–7.
- 623 [38] M. S. Robinson, R. A. Easom, M. J. Danson, P. D. Weitzman, Cit-
624 rate synthase of *Escherichia coli*. Characterisation of the enzyme from a
625 plasmid-cloned gene and amplification of the intracellular levels, *FEBS*
626 *Lett* 154 (1983) 51–4.
- 627 [39] T. Zhu, M. F. Bailey, L. M. Angley, T. F. Cooper, R. C. Dobson, The
628 quaternary structure of pyruvate kinase type 1 from *Escherichia coli* at
629 low nanomolar concentrations, *Biochimie* 92 (2010) 116–20.

- 630 [40] R. C. Wohl, G. Markus, Phosphoenolpyruvate carboxylase of Es-
631 cherichia coli. Purification and some properties, J Biol Chem 247 (1972)
632 5785–92.
- 633 [41] S. Kale, P. Arjunan, W. Furey, F. Jordan, A dynamic loop at the
634 active center of the Escherichia coli pyruvate dehydrogenase complex E1
635 component modulates substrate utilization and chemical communication
636 with the E2 component, J Biol Chem 282 (2007) 28106–16.
- 637 [42] P. Arjunan, N. Nemeria, A. Brunskill, K. Chandrasekhar, M. Sax,
638 Y. Yan, F. Jordan, J. R. Guest, W. Furey, Structure of the pyruvate de-
639 hydrogenase multienzyme complex E1 component from Escherichia coli
640 at 1.85 Å resolution, Biochemistry 41 (2002) 5213–21.
- 641 [43] S. Okino, M. Suda, K. Fujikura, M. Inui, H. Yukawa, Production of D-
642 lactic acid by Corynebacterium glutamicum under oxygen deprivation,
643 Appl Microbiol Biotechnol 78 (2008) 449–54.
- 644 [44] R. Tasseff, S. Nayak, S. Salim, P. Kaushik, N. Rizvi, J. D. Varner, Anal-
645 ysis of the molecular networks in androgen dependent and independent
646 prostate cancer revealed fragile and robust subsystems, PLoS One 5
647 (2010) e8864.
- 648 [45] R. Tasseff, S. Nayak, S. O. Song, A. Yen, J. D. Varner, Modeling and
649 analysis of retinoic acid induced differentiation of uncommitted precursor
650 cells, Integr Biol (Camb) 3 (2011) 578–91.
- 651 [46] J. D. Herman, <http://jdherman.github.io/salib/>, ????
- 652 [47] A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban,
653 D. E. Shumaker, C. S. Woodward, SUNDIALS: Suite of nonlinear and
654 differential/algebraic equation solvers, ACM T Math Software (TOMS)
655 31 (2005) 363–396.
- 656 [48] T. Kigawa, Y. Muto, S. Yokoyama, Cell-free synthesis and amino acid-
657 selective stable isotope labeling of proteins for NMR analysis, J Biomolec
658 NMR 6 (1995) 129–134.

Figure 1: Schematic of the core portion of the cell-free *E. coli* metabolic network. Metabolites of glycolysis, pentose phosphate pathway, Entner-Doudoroff pathway, and TCA cycle are shown. Metabolites of oxidative phosphorylation, amino acid biosynthesis and degradation, transcription/translation, chorismate metabolism, and energy metabolism are not shown.

Figure 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 or gray shaded region) over the ensemble of 100 sets.

Figure 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.

Figure 4: 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.

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

Figure 6: Key reaction fluxes of the network, in the first (gray boxes, top row) and second (gray boxes, bottom row) phases of metabolism. A. Fluxes of ATP generation and consumption, and GTP consumption toward protein synthesis. B. Fluxes of glycolysis and lactate and acetate metabolism. Fluxes are normalized to the first-phase glucose uptake rate. For PEP and pyruvate, accumulation (normalized to glucose uptake) is also shown.

Figure 7: 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: Breakdown of ATP generation. Flux through ATP-generating pathways in the first and second phases as percentages of total ATP generation in that phase.

Name	Index	Reaction	Phase 1	Phase 2
R_pgk	12	$13\text{DPG} + \text{ADP} \rightarrow 3\text{PG} + \text{ATP}$	14%	21%
R_pyk	18	$\text{ADP} + \text{PEP} \rightarrow \text{ATP} + \text{PYR}$	16%	<1%
R_sucCD	45	$\text{ADP} + \text{P}_i + \text{SUCCOA} \rightarrow \text{ATP} + \text{COA} + \text{SUCC}$	3%	5%
R_atp	55	$\text{ADP} + \text{P}_i + 4 \text{H}_e \rightarrow \text{ATP} + 4 \text{H} + \text{H}_2\text{O}$	54%	46%
R_ackA	68	$\text{ACTP} + \text{ADP} \rightarrow \text{AC} + \text{ATP}$	12%	28%
R_asn_deg	102	$\text{ASN} + \text{AMP} + \text{PP}_i \rightarrow \text{NH}_3 + \text{ASP} + \text{ATP}$	<1%	<1%
R_thr_deg3	109	$\text{THR} + \text{P}_i + \text{ADP} \rightarrow \text{NH}_3 + \text{FOR} + \text{ATP} + \text{PROP}$	<1%	<1%

Table 2: Breakdown of ATP consumption. Flux through ATP-consuming pathways in the first and second phases as percentages of total ATP consumption in that phase.

Name	Index	Reaction	Phase 1	Phase 2
R_glk_atp	1	$\text{ATP} + \text{GLC} \rightarrow \text{ADP} + \text{G6P} + \text{H}$	22%	<1%
R_pfk	4	$\text{ATP} + \text{F6P} \rightarrow \text{ADP} + \text{FBP}$	24%	<1%
R_pps	22	$\text{ATP} + \text{H}_2\text{O} + \text{PYR} \rightarrow \text{AMP} + \text{PEP} + \text{P}_i$	1%	1%
R_acs	70	$\text{AC} + \text{ATP} + \text{COA} \rightarrow \text{ACCOA} + \text{AMP} + \text{PP}_i$	8%	19%
R_glnA	86	$\text{GLU} + \text{ATP} + \text{NH}_3 \rightarrow \text{GLN} + \text{ADP} + \text{P}_i$	1%	2%
R_atp_amp	152	$\text{ATP} + \text{H}_2\text{O} \rightarrow \text{AMP} + \text{PP}_i$	6%	13%
R_udp_utp	160	$\text{UDP} + \text{ATP} \rightarrow \text{UTP} + \text{ADP}$	3%	6%
R_cdp_ctp	161	$\text{CDP} + \text{ATP} \rightarrow \text{CTP} + \text{ADP}$	4%	8%
R_gdp_gtp	162	$\text{GDP} + \text{ATP} \rightarrow \text{GTP} + \text{ADP}$	3%	4%
R_atp_ump	163	$\text{ATP} + \text{UMP} \rightarrow \text{ADP} + \text{UDP}$	1%	3%
R_atp_cmp	164	$\text{ATP} + \text{CMP} \rightarrow \text{ADP} + \text{CDP}$	2%	3%
R_adk_atp	166	$\text{AMP} + \text{ATP} \rightarrow 2 \text{ ADP}$	18%	35%
tRNA charging	185-204	$\text{AA} + \text{tRNA} + \text{ATP} + \text{H}_2\text{O} \rightarrow \text{AA}\cdot\text{tRNA} + \text{AMP} + \text{PP}_i$	2%	2%
Other			4%	4%

Table 3: Mean and standard deviation of Akaike information criterion (AIC), by measurement, for the ensemble and random ensemble.

Measurement	$\mu_{\text{AIC}}^{\text{Ens}}$	$\sigma_{\text{AIC}}^{\text{Ens}}$	$\mu_{\text{AIC}}^{\text{Rand}}$	$\sigma_{\text{AIC}}^{\text{Rand}}$	$\mu_{\text{AIC}}^{\text{Rand}} - \mu_{\text{AIC}}^{\text{Ens}}$
GLC	65.4	2.1	103.9	0.6	38.5
CAT	-23.0	10.5	-5.2	<0.1	17.8
PYR	64.8	10.3	84.7	0.7	19.9
LAC	70.7	4.5	88.9	<0.1	18.2
AC	79.4	6.0	96	2.1	16.6
SUCC	59.6	3.4	55.5	4.1	-4.1
MAL	60.8	4.1	71.6	6.3	10.8
ATP	51.1	3.3	69.1	<0.1	18.0
ADP	39.8	3.7	53.2	4.7	13.4
AMP	32.9	1.5	75.1	5.7	42.2
GTP	53.4	1.6	68.2	<0.1	14.8
GDP	45.7	2.9	43.6	9.5	-2.1
GMP	46.5	4.2	46.1	12.5	-0.4
CTP	44.9	2.6	58.5	<0.1	13.7
CDP	38.8	1.6	50.7	8.2	11.8
CMP	32.1	4.0	51.9	9.1	19.8
UTP	55.6	5.2	53	<0.1	-2.7
UDP	28.2	4.6	51.9	11.5	23.6
UMP	35.3	3.3	72.3	7.3	36.9
ALA	66.4	4.4	100.5	1.1	34.1
ASN	53.7	1.5	67.6	3.8	13.8
ASP	65.9	2.5	79.5	<0.1	13.6
CYS	60.5	3.1	74	<0.1	13.5
GLN	54.3	5.6	84.7	<0.1	30.4
GLY	47.2	12.7	75.5	11.7	28.3
HIS	46.3	6.2	43.2	3.2	-3.2
ILE	53.3	3.8	48.4	4.8	-5.0
LEU	41.5	6.5	52.5	4.6	10.9
LYS	68.4	2.0	73.9	0.2	5.5
MET	55.9	1.0	57.4	4	1.5
PHE	43.4	5.9	57.7	8.3	14.3
PRO	54.4	2.8	47.9	6.7	-6.5
SER	65.9	4.1	81.4	<0.1	15.6
THR	28.2	5.5	63.2	14.9	35.0
TRP	31.2	5.7	79.9	1.4	48.6
TYR	39.3	2.0	36.7	5.4	-2.6
VAL	51.3	3.1	55.5	4.6	4.1

Table 4: Reference values for reaction rate maxima (V_{max}) from BioNumbers. V_{max} values calculated from turnover numbers (k_{cat}) from BioNumbers, and a characteristic enzyme concentration of 170 nM. Characteristic rate maximum for all other reactions calculated as geometric mean of calculated rate maxima.

Enzyme	Reaction	k_{cat} (min^{-1})	V_{max} (mM/h)	BNID#
Serine dehydrase	R_ser_deg	10400	104	101119
Isocitrate dehydrogenase	R_icd	11900	119	101152
Lactate dehydrogenase	R_ldh	5800	58	101036
Aspartate transaminase	R_aspC	25800	258	101108
	R_tyr			
	R_phe			
Enolase	R_eno	13200	132	101028
Pyruvate kinase	R_pyk	25000	250	101029
				101030
Malic enzyme	R_maeA	35400	354	101167
	R_maeB			
Phosphofructokinase	R_pfk	554400	5544	104955
Malate dehydrogenase	R_mdh	33000	330	101163
Citrate Synthase	R_gltA	42000	420	101149
6PG dehydrogenase	R_zwf	3200	32	101048
	R_pgl			
	R_gnd			
Succinate dehydrogenase	R_sdh	121	1.21	101162
Succinyl-coA synthetase	R_sucCD	4700	47	101158
3PGA dehydrogenase	R_gpm	1100	11	101135
PEP carboxylase	R_ppc	35400	354	101139
3PGA kinase	R_pgk	4300	43	101016
Characteristic V_{max}			110	

Table 5: Reference values for transcription, translation, and mRNA degradation from literature. Transcription rate calculated from elongation rate, mRNA length, and promoter activity level. Translation rate calculated from elongation rate, protein length, and polysome amplification constant. mRNA degradation rate calculated from mRNA degradation time.

Description	Parameter	Value	Units	Reference
T7 RNA polymerase concentration	R_T	1.0	μM	
Ribosome concentration	R_X	2	μM	[10]
Transcription saturation coefficient	K_T	100	nM	estimated
Translation saturation coefficient	K_X	45	μM	estimated
Transcription elongation rate	$\dot{v}_T$	25	nt/s	[10]
CAT mRNA length	l_G	660	nt	[48]
Promoter activity level	u	0.9		estimated
Transcription rate	$k_{cat}^T = \left(\frac{\dot{v}_T}{l_G}\right) u$	123	h^{-1}	calculated
Translation elongation rate	$\dot{v}_X$	1.5	aa/s	[10]
CAT protein length	l_P	219	aa	[48]
Polysome amplification constant	K_P	10		estimated
Translation rate	$k_{cat}^X = \left(\frac{\dot{v}_X}{l_P}\right) K_P$	247	h^{-1}	calculated
mRNA degradation time	$t_{1/2}$	8	min	BNID 106253
mRNA degradation rate	$k_{deg} = \frac{\ln(2)}{t_{1/2}}$	5.2	h^{-1}	calculated
ATP transcription coefficient	ATP_T	176		calculated
CTP transcription coefficient	CTP_T	144		calculated
GTP transcription coefficient	GTP_T	151		calculated
UTP transcription coefficient	UTP_T	189		calculated
ATP tRNA charging coefficient	ATP_X	219		calculated
GTP translation coefficient	GTP_X	438		calculated

Table 6: Nomenclature

Symbol	Compound name
GLC	alpha-D-Glucose
G6P	Glucose 6-phosphate
F6P	Fructose 6-phosphate
FBP	Fructose 1,6-diphosphate
T3P	Dihydroxyacetone phosphate
13DPG	1,3-bis-Phosphoglycerate
3PG	3-Phosphoglycerate
2PG	2-Phosphoglycerate
PEP	Phosphoenolpyruvate
PYR	Pyruvate
LAC	D-Lactate
6PG	6-Phospho-D-glucono-1,5-lactone; 6-Phospho-D-gluconate
RU5P	D-Ribulose 5-phosphate
XU5P	D-Xylulose 5-phosphate
R5P	Ribose 5-phosphate
S7P	sedo-Heptulose 7-phosphate
G3P	Glyceraldehyde 3-phosphate
E4P	Erythrose 4-phosphate
2DDG6P	2-Dehydro-3-deoxy-D-gluconate 6-phosphate
COA	Coenzyme A
ACCOA	Acetyl coenzyme A
AC	Acetate
CIT	Citrate
ICIT	Isocitrate
AKG	alpha-Ketoglutarate

Continued on next page

659

Symbol	Compound name
SUCCOA	Succinyl coenzyme A
SUCC	Succinate
FUM	Fumarate
MAL	Malate
OAA	Oxaloacetate
FOR	Formate
PROP	Propanoate
ALA	Alanine
ARG	Arginine
ASP	Aspartate
ASN	Asparagine
CYS	Cysteine
GLU	Glutamate
GLN	Glutamine
GLY	Glycine
HIS	Histidine
ILE	Isoleucine
LEU	Leucine
LYS	L-Lysine
MET	Methionine
PHE	Phenylalanine
PRO	Proline
SER	Serine
THR	Threonine
TRP	Tryptophan
TYR	Tyrosine

Continued on next page

Symbol	Compound name
VAL	Valine
AA	Amino acid
AA·tRNA	Aminoacyl tRNA
ATP	Adenosine triphosphate
ADP	Adenosine diphosphate
AMP	Adenosine monophosphate
CTP	Cytidine triphosphate
CDP	Cytidine diphosphate
CMP	Cytidine monophosphate
GTP	Guanosine triphosphate
GDP	Guanosine diphosphate
GMP	Guanosine monophosphate
UTP	Uridine triphosphate
UDP	Uridine diphosphate
UMP	Uridine monophosphate
CAT	Chloramphenicol acetyltransferase

661

Figure S1: Histograms of model parameters, across the ensemble of 100 sets. A. Histogram of rate maxima. B. Histogram of saturation constants.