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

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

School of Chemical and Biomolecular Engineering

Cornell University, Ithaca NY 14853

[1]School of Chemical Engineering

Stanford University, Stanford, California 94305

Running Title: Dynamic model of cell free protein synthesis

To be submitted: *Scientific Reports*

*Corresponding author:

Jeffrey D. Varner,

Professor, 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

Fill me in.

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

Introduction

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 [1]. 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 [2]. Shuler and coworkers generated many single cell kinetic models, including single cell models of eukaryotes [3, 4], minimal cell architectures [5], as well as DNA sequence based whole-cell models of *E. coli* [6]. Conversely, highly abstracted kinetic frameworks, such as the cybernetic framework, represented a paradigm shift, viewing cells as growth-optimizing strategists [7]. Cybernetic models have been highly successful at predicting metabolic choice behavior, e.g., diauxie behavior [8], steady-state multiplicity [9], as well as the cellular response to metabolic engineering modifications [10]. Unfortunately, traditional, fully structured cybernetic models also suffer from an identifiability challenge, as both the kinetic parameters and an abstracted model of cellular objectives must be estimated simultaneously. However, recent cybernetic formulations from Ramkrishna and colleagues have successfully treated this identifiability challenge through elementary mode reduction [11, 12].

In the post genomics world, large-scale stoichiometric reconstructions of microbial metabolism popularized by static, constraint-based modeling techniques such as flux balance analysis (FBA) have become standard tools [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 such as *E. coli* [15] or *B. subtilis* [16], are now available [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.

Cell-free systems offer many advantages for the study, manipulation and modeling of metabolism compared to *in vivo* processes. Central amongst these advantages is direct access to metabolites and the microbial biosynthetic machinery without the interference of a cell wall. This allows us to control as well as interrogate the chemical environment while the biosynthetic machinery is operating, potentially at a fine time resolution. Second, cell-free systems also allow us to study biological processes without the complications associated with cell growth. Cell-free protein synthesis (CFPS) systems are arguably the most prominent examples of cell-free systems used today [23]. However, CFPS is not new; CFPS in crude *E. coli* extracts has been used since the 1960s to explore fundamentally important biological mechanisms [24, 25]. Today, cell-free systems are used in a variety of applications ranging from therapeutic protein production [26] to synthetic biology [27]. Interestingly, many of the challenges confronting genome-scale kinetic modeling can potentially be overcome in a cell-free system. For example, there is no complex transcriptional regulation to consider, transient metabolic measurements are easier to

obtain, and we no longer have to consider cell growth. Thus, cell-free operation holds several significant advantages for model development, identification and validation. Theoretically, genome-scale cell-free kinetic models may be possible for industrially important organisms, such as *E. coli* or *B. subtilis*, if a simple, tractable framework for integrating allosteric regulation with enzyme kinetics can be formulated.

In this study, we present an effective biochemical network modeling framework for building dynamic cell-free metabolic models. The key innovation of our approach is the seamless integration of simple effective rules encoding complex regulation with traditional kinetic pathway modeling. This integration allows the description of complex regulatory interactions, such as time-dependent allosteric regulation of enzyme activity, in the absence of specific mechanistic information. The regulatory rules are easy to understand, easy to formulate and do not rely on overarching theoretical abstractions or restrictive assumptions. We tested our approach by modeling the time evolution of several hypothetical cell-free metabolic networks. In particular, we tested whether our effective modeling approach could describe classically expected enzyme kinetic behavior, and second whether we could simultaneously estimate kinetic parameters and regulatory connectivity, in the absence of specific mechanistic knowledge, from synthetic experimental data. Toward these questions, we explored five hypothetical cell-free networks. Each network shared the same enzymatic connectivity, but had different allosteric regulatory connectivity. We found that simple effective rules, when integrated with traditional enzyme kinetic expressions, captured complex allosteric patterns such as ultrasensitivity or non-competitive inhibition in the absence of mechanistic information. Second, when integrated into network models, these rules captured classical regulatory patterns such as product-induced feedback inhibition. Lastly, we showed, at least for the network architectures considered here, that we could simultaneously estimate kinetic parameters and allosteric connectivity from synthetic data starting from an unbiased collection of possible allosteric structures using

particle swarm optimization. However, when starting with an initial population that was heavily enriched with incorrect structures, our particle swarm approach could converge to an incorrect structure. While only an initial proof-of-concept, the framework presented here could be an important first step toward genome-scale cell-free kinetic modeling of the biosynthetic capacity of industrially important organisms.

The introduction has four paragraphs (introduction no longer than 3 pages). Follow the cell free paper from last year:

1. **First paragraph**: Introduce mathematical modeling, and its role in biochemical engineering.

2. **Second paragraph**: Contrast current static metabolic modeling approaches e.g., FBA with dynamic models.

3. **Third paragraph**: Introduce cell free protein synthesis.

4. **Fourth paragraph**: In this study, [Repeat the abstract with some additional detail]. Taken together, [killer statement].

Results

The results are presented in **past tense**. Each paragraph starts with a statement of the result in that paragraph in active voice. Each results paragraph ends with a Taken together type statement followed by a link statement e.g., Next we considered etc. When referring to figures, state what the figures shows (Fig. ZZ).

1. **First section**:Description of the model biology
2. **Second section**:Estimation of the model parameters, and refinement of the model structure (inclusion of the AA degradation pathways)
3. **Third section**:Analysis of the flux distribution (over the ensemble?), sensitivity results (first parameters, then AA)

Discussion

The discussion has three (sometimes four) paragraphs:

1. **First paragraph**: Present a modified version of the last paragraph of the introduction. In this study, [...]. Taken together, [killer statement]

2. **Second paragraph**: Contrast the key findings of the study with other computational/experimental studies

3. **Third paragraph**: Present future directions. If you had more time, what would like to do? Highlight the key shortcomings of the approach and how will we address them in the future. In this case, we will have a scaling issue if we extend to genome scale. We should extend to dynamic cases, and we need to experimentally validate the findings.

Materials and Methods

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

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

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

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

The reaction rate was written as the product of a kinetic term ($\bar{r}_j$) and a control term (v_j), $r_j\left(\mathbf{x}, \mathbf{k}\right) = \bar{r}_j v_j$. In this study, we used either saturation or mass action kinetics. The control term $0 \leq v_j \leq 1$ depended upon the combination of factors which influenced rate process j. For each rate, we used a rule-based approach to select from competing control factors. If rate j was influenced by $1, \ldots, m$ factors, we modeled this relationship as $v_j = \mathcal{I}_j\left(f_{1j}\left(\cdot\right), \ldots, f_{mj}\left(\cdot\right)\right)$ where $0 \leq f_{ij}\left(\cdot\right) \leq 1$ denotes a regulatory transfer function quantifying the influence of factor i on rate j. The function $\mathcal{I}_j\left(\cdot\right)$ is an integration rule which maps the output of regulatory transfer functions into a control variable. Each regulatory

transfer function took the form:

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

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

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

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

Generation of model ensemble We generated an ensemble of 18,000 parameter sets via a downhill-only random walk Monte Carlo method [**?**]. Beginning with a single parameter set as a starting point, we calculated its cost function, equal to the sum-absolute-error between experimental data and model predicitons:

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

where $\mathcal{D}$ denotes the number of datasets, w_i denotes a weight, equal to 5 for the glucose, CAT, pyruvate, lactate, acetate, succinate, and malate datasets, and 1 elsewhere, $\mathcal{T}$ denotes the number of timepoints in the ith dataset, $t(j)$ denotes the jth timepoint, x_{ij}^{data} denotes the value of the ith dataset at the jth timepoint, and $x_i^{sim}|_{t(j)}$ dneotes the simulated value of the metabolite corresponding to the ith dataset, interpolated to the jth

timepoint. We then perturbed model parameters:

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

where $\mathcal{P}$ denotes the number of parameters, equal to 652, which includes 163 rate constants, 455 saturation constants, and 34 control parameters, k_i^{new} denotes the new value of the ith parameter, k_i denotes the current value of the ith parameter, a denotes a distribution variance, set to 0.03, and r denotes a random sample from the normal distribution. We stored the parameter set and calculated its cost; if it was less than the previous cost, we used the new parameter set to generate the following set. After generating 180,000 sets we defined the 18,000 sets with the lowest cost values as our ensemble, and the set with the lowest cost value as our best-fit set.

Global and local sensitivity analysis We conducted a global sensitivity analysis, using the variance-based method of Sobol, to estimate which of the experimentally controllable parameters affected the performance of the reduced order model [28]. This included the initial conditions of glucose, oxygen, amino acids, and enzymes. We computed the total sensitivity index of each parameter relative to a performance objective of area under the CAT curve (CAT production). We established the sampling bounds for each parameter from the value of that parameter in the set used to generate the ensemble. We used the sampling method of Saltelli *et al.* [29] to compute a family of $N\,(2d + 2)$ sets which obeyed our parameter ranges, where N was the number of trials, and d was the number of parameters in the model. In our case, N = 300 and d = 185, so the total sensitivity indices were computed from 111,600 model evaluations. The variance-based sensitivity analysis was conducted using the SALib module encoded in the Python programming language [30]. We conducted a local sensitivity analysis to estimate which of the other model parameters affected performance. This included the same parameters that were

varied in the ensemble: rate constants, saturation constants, and control parameters. The local sensitivity for each parameter was calculated across a sub-ensemble of 180 parameter sets, randomly chosen from the ensemble of 18,000 sets:

$$S_{ij} = \frac{p_{ij}}{AUC(p_{ij})} \frac{AUC(p_{ij} + \Delta p_{ij}) - AUC(p_{ij})}{\Delta p_{ij}} \qquad i = 1, 2, \ldots, \mathcal{E} \qquad j = 1, 2, \ldots, \mathcal{P}$$
$$\Delta p_{ij} = 0.001 \; p_{ij} \tag{7}$$

where $\mathcal{E}$ denotes the number of parameter sets in the sub-ensemble, equal to 180, $\mathcal{P}$ denotes the number of parameters, equal to 652, S_{ij} denotes the sensitivity of the jth parameter for the ith parameter set, p_{ij} denotes the value of the jth parameter for the ith parameter set, Δp_{ij} denotes the perturbation of the jth parameter for the ith parameter set, equal to 0.1% of the parameter value, and $AUC()$ denotes the area under the CAT curve. We then calculated the mean and

Calculation of CAT yield The theoritical carbon yield of CAT was calculated using flux balance anaylsis (FBA) with a sequence-based analysis on CAT. The sequence specific FBA [31] problem was formulated as:

$$\max_{\boldsymbol{v}} \left(v_{obj} = \boldsymbol{c}^T \boldsymbol{v} \right) \qquad \alpha_i \leq v_i \leq \beta_i \qquad i = 1, 2, \ldots, \mathcal{R}$$
$$\text{Subject to : } \mathbf{Sv} = \mathbf{0} \tag{8}$$

where $\mathbf{S}$ denotes the stoichiometric matrix, $\mathbf{v}$ denotes the unknown flux vector, $\boldsymbol{c}$ denotes the objective selection vector, and α_i and β_i denote the lower and upper bounds on flux v_i, respectively. The stoichiometric matrix was expanded to include the transcription and translation reactions for producing CAT. The objective v_{obj} was to maximize the specific rate of CAT formation. The specific glucose, amino acids and oxygen uptake rates were constrained to allow a maximum flux of 10 mM/hr. The flux balance analysis problem was solved using the GNU Linear Programming Kit (v4.52) [32]. The solution flux vector was

used to calculate the theoritical carbon yield of CAT based on carbon consumed (glucose and amino acids). The carbon yield was formulated as:

$$Yield = \frac{C_{CAT} v_{CAT}}{\sum_{i=1}^{\mathcal{R}} C_i v_i}$$

where C_{CAT} and C_i denote the carbon number of CAT and substrate i, respectively, v_{CAT} and v_i denote the flux of CAT and substrate i, respectively, and $\mathcal{R}$ denotes the number of substrates consumed.

Acknowledgements

This study was supported by an award from [FILL ME IN].

References

1. Fredrickson AG (1976) Formulation of structured growth models. Biotechnol Bioeng 18: 1481-6.
2. Domach MM, Leung SK, Cahn RE, Cocks GG, Shuler ML (1984) Computer model for glucose-limited growth of a single cell of escherichia coli b/r-a. Biotechnol Bioeng 26: 203-16.
3. Steinmeyer D, Shuler M (1989) Structured model for Saccharomyces cerevisiae. Chem Eng Sci 44: 2017 - 2030.
4. Wu P, Ray NG, Shuler ML (1992) A single-cell model for cho cells. Ann N Y Acad Sci 665: 152-87.
5. Castellanos M, Wilson DB, Shuler ML (2004) A modular minimal cell model: purine and pyrimidine transport and metabolism. Proc Natl Acad Sci U S A 101: 6681-6.
6. Atlas JC, Nikolaev EV, Browning ST, Shuler ML (2008) Incorporating genome-wide dna sequence information into a dynamic whole-cell model of escherichia coli: application to dna replication. IET Syst Biol 2: 369-82.
7. Dhurjati P, Ramkrishna D, Flickinger MC, Tsao GT (1985) A cybernetic view of microbial growth: modeling of cells as optimal strategists. Biotechnol Bioeng 27: 1-9.
8. Kompala DS, Ramkrishna D, Jansen NB, Tsao GT (1986) Investigation of bacterial growth on mixed substrates: experimental evaluation of cybernetic models. Biotechnol Bioeng 28: 1044-55.
9. Kim JI, Song HS, Sunkara SR, Lali A, Ramkrishna D (2012) Exacting predictions by cybernetic model confirmed experimentally: steady state multiplicity in the chemostat. Biotechnol Prog 28: 1160-6.
10. Varner J, Ramkrishna D (1999) Metabolic engineering from a cybernetic perspective: aspartate family of amino acids. Metab Eng 1: 88-116.
11. Song HS, Morgan JA, Ramkrishna D (2009) Systematic development of hybrid cy-

bernetic models: application to recombinant yeast co-consuming glucose and xylose. Biotechnol Bioeng 103: 984-1002.

12. Song HS, Ramkrishna D (2011) Cybernetic models based on lumped elementary modes accurately predict strain-specific metabolic function. Biotechnol Bioeng 108: 127-40.
13. Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 10: 291-305.
14. Edwards JS, Palsson BO (2000) The escherichia coli mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci U S A 97: 5528-33.
15. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. Mol Syst Biol 3: 121.
16. Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R (2007) Genome-scale reconstruction of metabolic network in bacillus subtilis based on high-throughput phenotyping and gene essentiality data. J Biol Chem 282: 28791-9.
17. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7: 129-43.
18. Ibarra RU, Edwards JS, Palsson BO (2002) Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420: 186-9.
19. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in escherichia coli. Mol Syst Biol 3: 119.
20. Hyduke DR, Lewis NE, Palsson BØ (2013) Analysis of omics data with genome-scale models of metabolism. Mol Biosyst 9: 167-74.
21. McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale

metabolic network reconstructions of escherichia coli. Mol Syst Biol 9: 661.

22. Zomorrodi AR, Suthers PF, Ranganathan S, Maranas CD (2012) Mathematical optimization applications in metabolic networks. Metab Eng 14: 672-86.
23. Jewett MC, Calhoun KA, Voloshin A, Wuu JJ, Swartz JR (2008) An integrated cell-free metabolic platform for protein production and synthetic biology. Mol Syst Biol 4: 220.
24. MATTHAEI JH, NIRENBERG MW (1961) Characteristics and stabilization of dnaase-sensitive protein synthesis in e. coli extracts. Proc Natl Acad Sci U S A 47: 1580-8.
25. NIRENBERG MW, MATTHAEI JH (1961) The dependence of cell-free protein synthesis in e. coli upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci U S A 47: 1588-602.
26. Lu Y, Welsh JP, Swartz JR (2014) Production and stabilization of the trimeric influenza hemagglutinin stem domain for potentially broadly protective influenza vaccines. Proc Natl Acad Sci U S A 111: 125-30.
27. Hodgman CE, Jewett MC (2012) Cell-free synthetic biology: thinking outside the cell. Metab Eng 14: 261-9.
28. Sobol I (2001) Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. Mathematics and Computers in Simulation 55: 271 - 280.
29. Saltelli A, Annoni P, Azzini I, Campolongo F, Ratto M, et al. (2010) Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. Computer Physics Communications 181: 259 - 270.
30. Herman JD. http://jdherman.github.io/salib/.
31. ALLEN TE, PALSSON BO (2003) Sequence-based analysis of metabolic demands for protein synthesis in prokaryotes. Journal of Theoretical Biology 220: 1 - 18.
32. (2016). GNU Linear Programming Kit, Version 4.52. URL http://www.gnu.org/software/glpk/glpk.html.

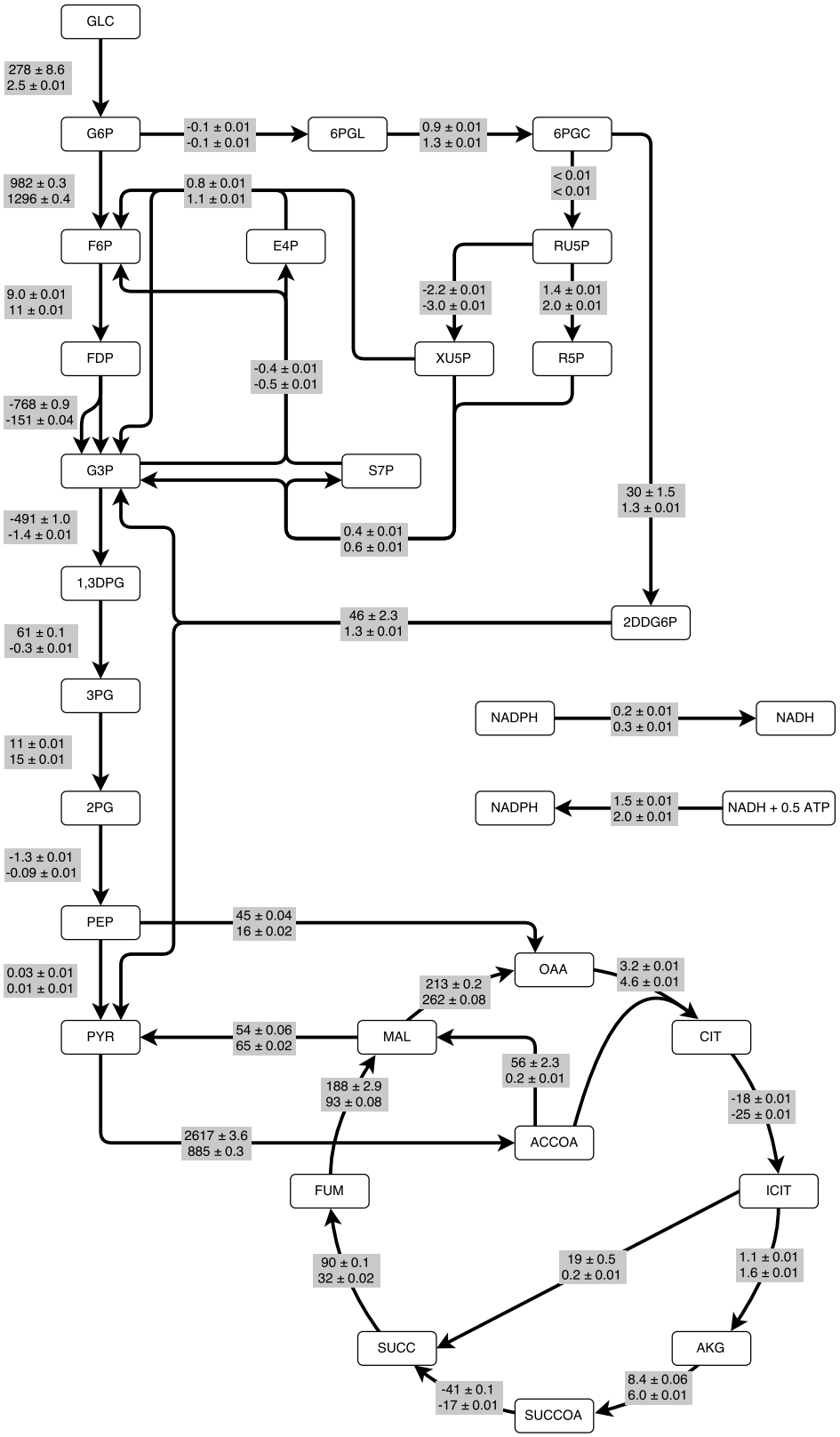

Fig. 1: Flux profile for glycolysis, pentose phosphate pathway, Entner-Doudoroff pathway, TCA cycle, and NADPH/NADH transfer. Mean ± standard error for the reaction flux at 1.5 hrs (top) and 3 hrs (bottom), normalized to CAT synthesis flux.

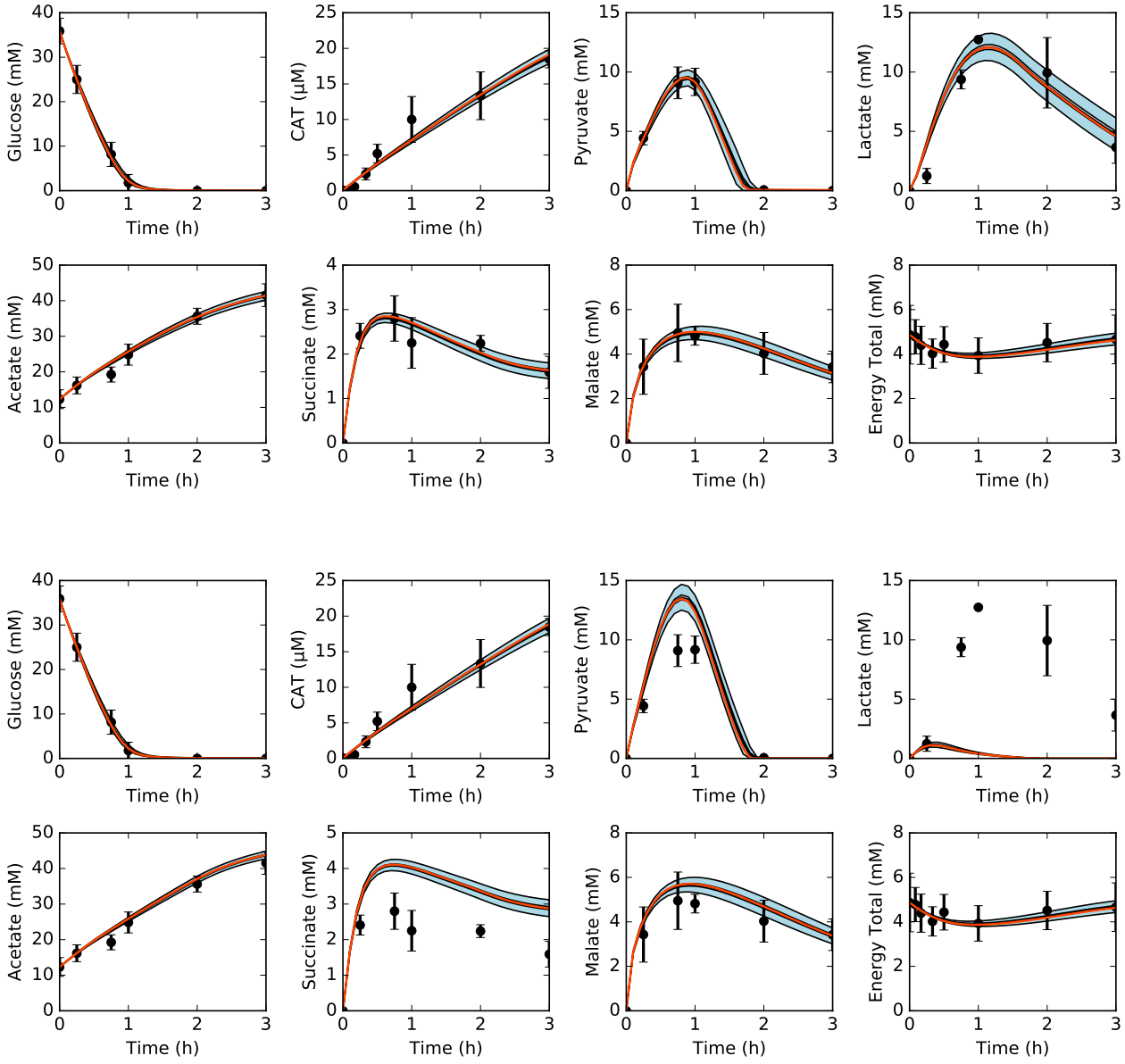

Fig. 2: A: Central carbon metabolism in the presence of allosteric control, including glucose (substrate), CAT (product), and intermediates, as well as total concentration of energy species. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence interval (blue shaded region) and 95% confidence interval of the mean (gray shaded region) over the ensemble of 18,000 sets. B: Central carbon metabolism in the absence of allosteric control.

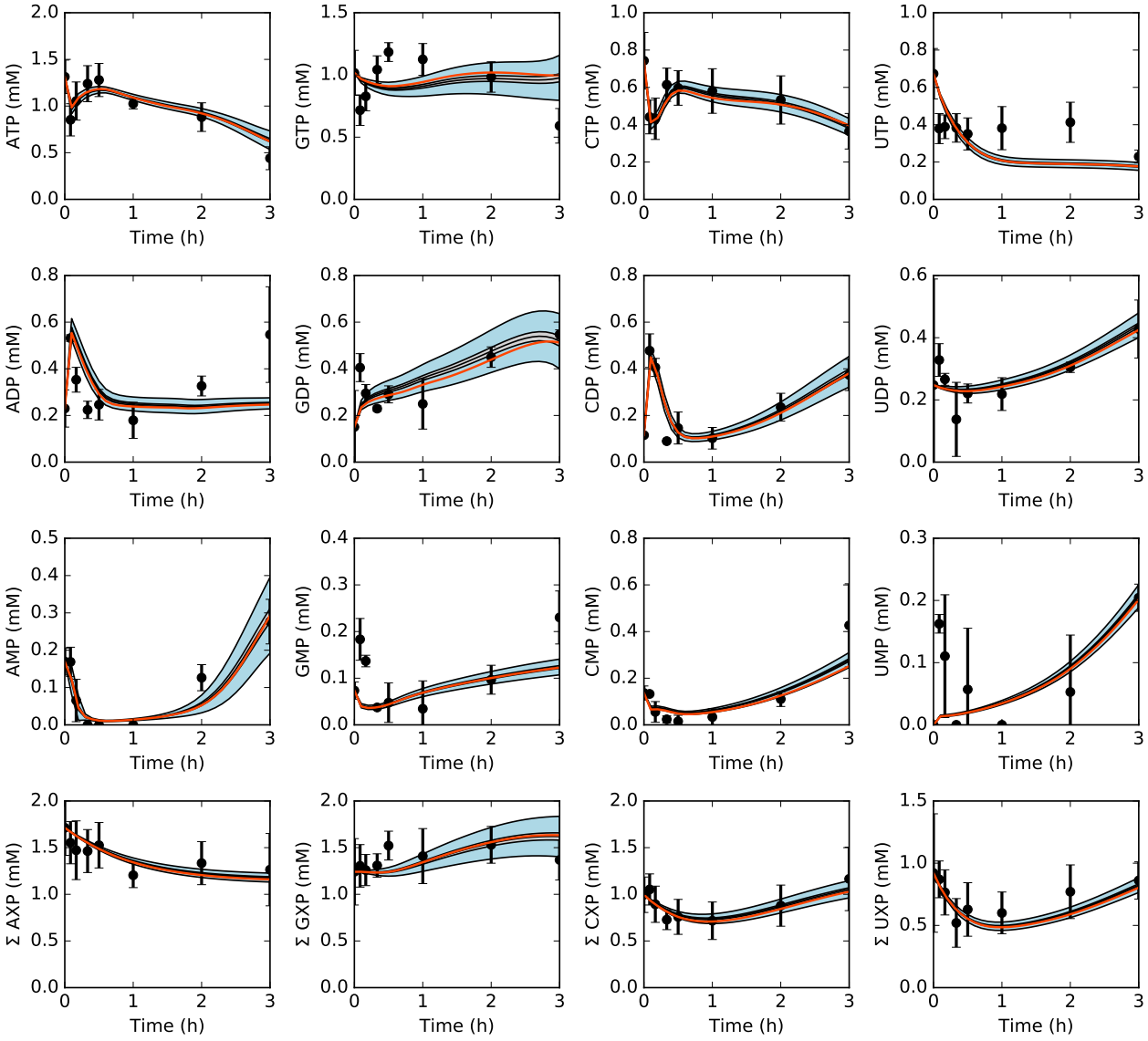

Fig. 3: Energy species and energy totals by base in the presence of allosteric control. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence interval (blue shaded region) and 95% confidence interval of the mean (gray shaded region) over the ensemble of 18,000 sets.

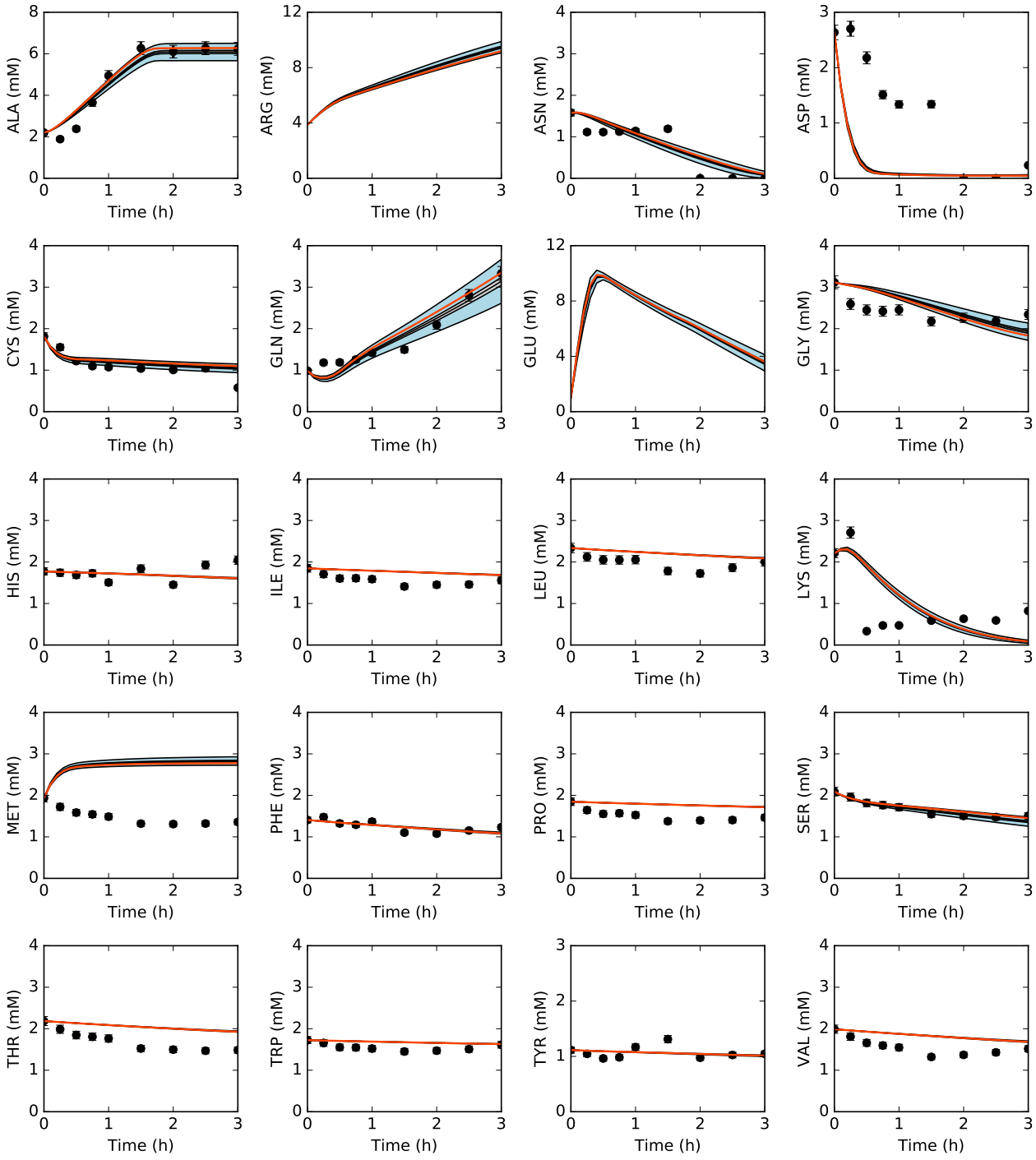

Fig. 4: Amino acids in the presence of allosteric control. Best-fit parameter set (orange line) versus experimental data (points). 95% confidence interval (blue shaded region) and 95% confidence interval of the mean (gray shaded region) over the ensemble of 18,000 sets.

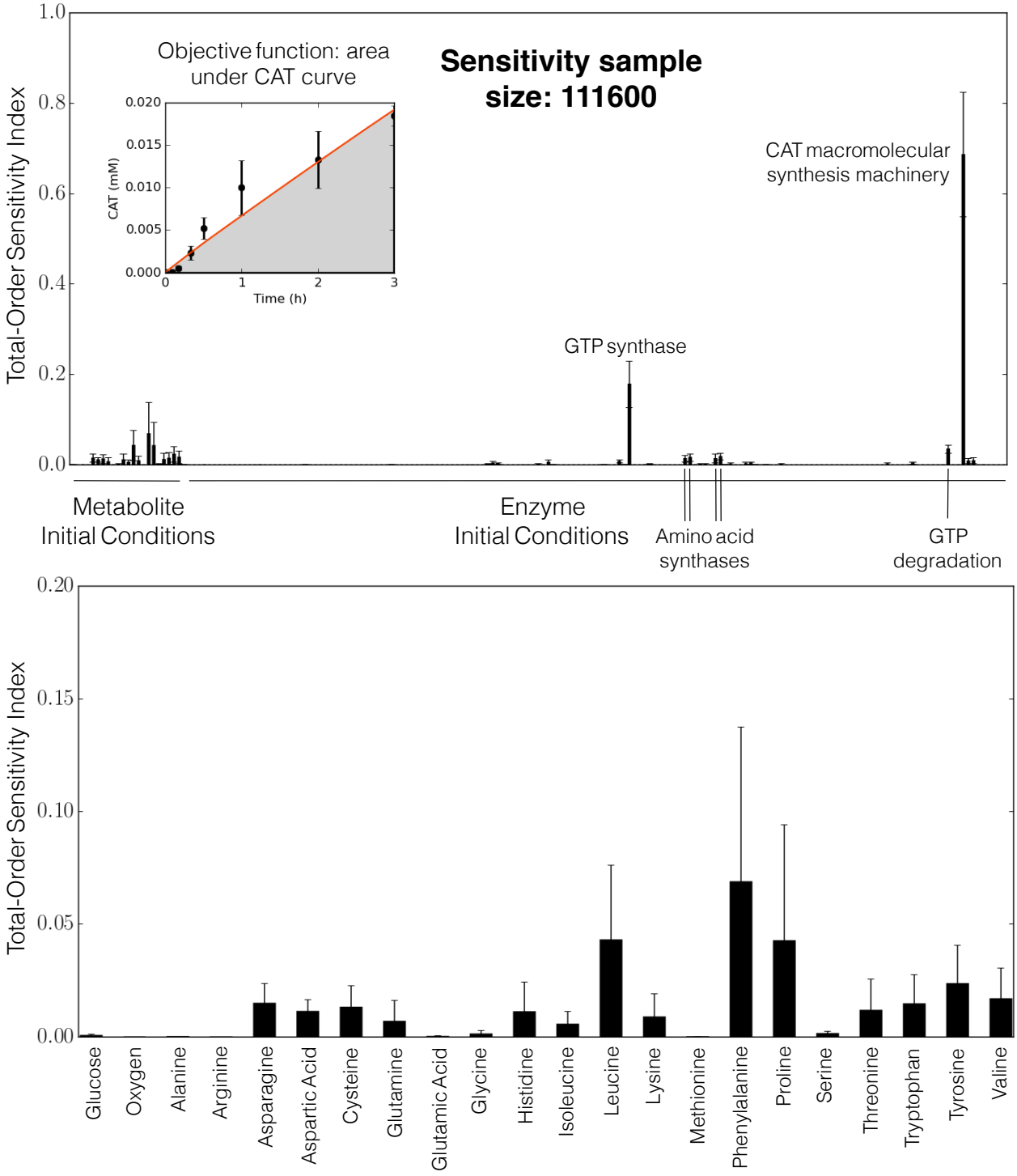

Fig. 5: Total-order global sensitivities for experimentally controllable initial conditions, including glucose, oxygen, amino acids, and enzymes.

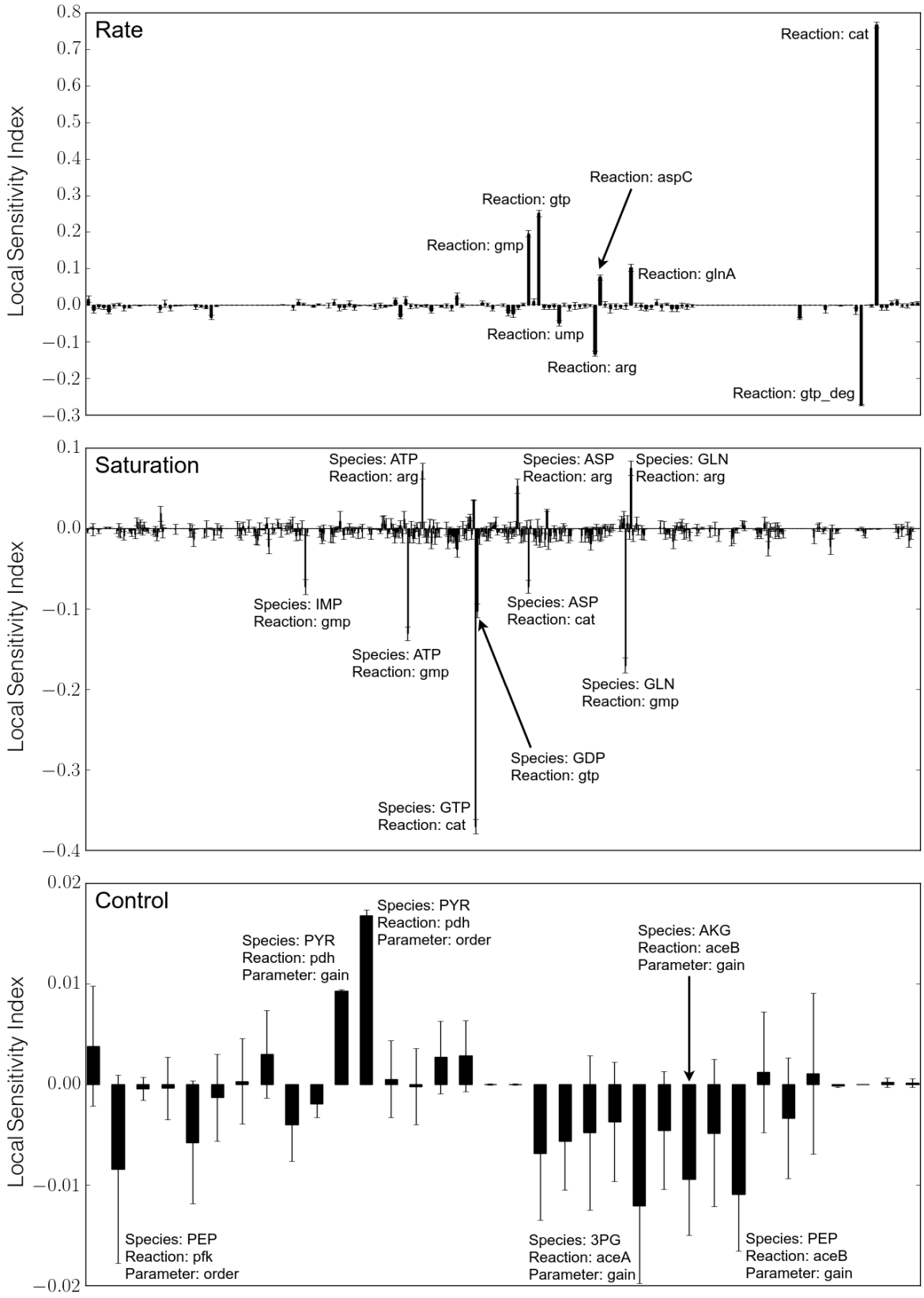

Fig. 6: Mean and standard error of local sensitivities of rate constants (top), saturation constants (middle), and control parameters (bottom).