Dynamic Modeling of Cell-Free Biochemical Networks using Effective Kinetic Models

Joseph A. Wayman, Adithya Sagar and Jeffrey D. Varner*

School of Chemical and Biomolecular Engineering

Cornell University, Ithaca NY 14853

Running Title: Effective models of metabolism

To be submitted: *Processes*

*Corresponding author:

Jeffrey D. Varner,

Associate 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

Cell-free systems offer many advantages for the study, manipulation and modeling of metabolism compared to *in vivo* processes. 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*, if a simple, tractable framework for integrating allosteric regulation with enzyme kinetics can be formulated. Toward this unmet need, we present an effective biochemical network modeling framework for building dynamic cell-free metabolic models. The key innovation of our approach is the integration of simple effective rules encoding complex allosteric regulation with traditional kinetic pathway modeling. We tested our approach by modeling the time evolution of several hypothetical cell-free metabolic networks. 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 classic 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. 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.

Keywords: Cell-free metabolism, Mathematical modeling

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. Recent cybernetic formulations from Ramkrishna and colleagues have successfully treated this identifiability challenge through elementary mode reduction, though the techniques replace detailed biological mechanism with an optimization heuristic [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 assump-

1

tion 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 connectiv-

ity from synthetic data. 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.

Results

Formulation and properties of effective cell-free metabolic models. We developed

two proof-of-concept metabolic networks to investigate the features of our effective bio-

chemical network modeling approach (Fig. 1). In both examples, substrate S was con-

verted to the end products P_1 and P_2 through a series of enzymatically catalyzed reac-

tions, including a branch point at hypothetical metabolite M_2. Several of these reactions

involved cofactor dependence (AH or A), and various allosteric regulatory mechanisms

modified the activity of pathway enzymes. Network A included feedback inhibition of the

initial pathway enzyme (E_1) by pathway end products P_1 and P_2 (Fig. 1A). On the other

hand, network B involved feedback inhibition of E_1 by P_2 and E_6 by P_1 (Fig. 1B). In both

networks, branch point enzymes E_3 and E_6 were subject to feed-forward activation by

reduced cofactor AH. Lastly, it is known experimentally that cell-free systems have a finite

operational lifespan. Loss of biosynthetic capability could be a function of many factors,

e.g., cofactor or metabolite limitations. We modeled the loss of biosynthetic capability as

a non-specific first-order decay of enzyme activity.

Allosteric regulation of enzyme activity was modeled by combining individual regula-

tory contributions to the activity of pathway enzymes into a control coefficient using an

integration rule (Fig. 2). This strategy is similar in spirit to the Constrained Fuzzy Logic

(cFL) approach of Lauffenburger and coworkers which has been used to effectively model

signal transduction pathways important in human health [28]. In our formulation, Hill-like

transfer functions $0 \leq f(\mathcal{Z}) \leq 1$ were used to calculate the influence of factor abundance

upon target enzyme activity. In this context, factors can be individual metabolite levels

or some function, e.g., the product of metabolite levels. However, more generally, factors

can also correspond to non-modeled influences, categorial variables or other abstract

quantities. In the current study, we simply let $\mathcal{Z}$ correspond to the abundance of individ-

ual metabolites, however in general this can be a complex function of both modeled and

5

unmodeled factors. When an enzyme was potentially sensitive to more than one regula-

tory input, logical integration rules were used to select which regulatory transfer function

influenced enzyme activity at any given time. Thus, our test networks involved important

features such as cofactor recycling, enzyme activity and metabolite dynamics, as well as

multiple overlapping allosteric regulatory mechanisms.

The rule-based regulatory strategy approximated the behavior of classical allosteric

activation and inhibition mechanisms (Fig. 3). We considered the enzyme catalyzed con-

version of substrate S to a product P, where the overall reaction rate was modeled as the

product of a Michaelis-Menten term and an effective allosteric control variable reflecting

the particular regulatory interaction. We first explored feed-forward substrate activation

of enzyme activity (for both positive and negative cooperativity). Consistent with classical

data, the rule-based strategy predicted a sigmoidal relationship between substrate abun-

dance and reaction rate as a function of the cooperativity parameter (Fig. 3A). For coop-

erativity parameters less than unity, increased substrate abundance *decreased* the max-

imum reaction rate. This was consistent with the idea that substrate binding *decreased*

at regulatory sites, which negatively impacted substrate binding at the active site. On the

other hand, as the cooperativity parameter increased past unity, the rate of conversion of

substrate S to product P by enzyme E approached a step function. In the presence of an

inhibitor, the rule-based strategy predicted non-competitive like behavior as a function of

the cooperativity parameter (Fig. 3B). When the control gain parameter, κ_{ij} in Eqn. (10),

was greater than unity, the inhibitory force was directly proportional to the cooperativity

parameter, η in Eqn. (10). Thus, as the cooperativity parameter increased, the maximum

reaction rate decreased (Fig. 3B). Interestingly, our rule-based approach was unable to

directly simulate competitive inhibition of enzyme activity. Taken together, the rule-based

strategy captured classical regulatory patterns for both enzyme activation and inhibition.

Thus, we are able to model complex kinetic phenomena such as ultrasensitivity, despite

an effective description of reaction kinetics.

End product yield was controlled by feedback inhibition, while product selectivity was controlled by branch point enzyme inhibition (Fig. 4). A critical test of our modeling approach was to simulate networks with known behavior. If we cannot reproduce the expected behavior of simple networks, then our effective modeling strategy, and particularly the rule-based approximation of allosteric regulation, will not be feasible for genome-scale cell-free problems. We considered two cases, control ON/OFF, for each network configuration. Each of these cases had identical kinetic parameters and initial conditions; the *only* differences between the cases were the allosteric regulation rules and the control parameters associated with these rules. As expected, end product accumulation was larger for network A when the control was OFF (no feedback inhibition of E_1 by P_1 and P_2), as compared to the ON case (Fig. 4A). We found this behavior was robust to the choice of underlying kinetic parameters, as we observed that same qualitative response across an ensemble of 100 randomized parameter sets, for fixed control parameters. The control ON/OFF response of network B was more subtle. In the OFF case, the behavior was qualitatively similar to network A. However, for the ON case, flux was diverted away from P_2 formation by feedback inhibition of E_6 activity at the M_2 branch point by P_1 (Fig. 4B). Lower E_6 activity at the M_2 branch point allowed more flux toward P_1 formation, hence the yield of P_1 also increased (Fig. 4C). Again, the control ON/OFF behavior of network B was robust to changes in kinetic parameters, as the same qualitative trend was conserved across an ensemble of 100 randomized parameters, for fixed control parameters. Taken together, these simulations suggested that the rule-based allosteric control concept could robustly capture expected feedback behavior for networks with uncertain kinetic parameters.

Estimating parameters and effective allosteric regulatory structures. A critical challenge for any dynamic model is the estimation of kinetic parameters. For metabolic pro-

7

cesses, there is also the added challenge of identifying the regulation and control structures that manage metabolism. Of course, these issues are not independent; any description of enzyme activity regulation will be a function of system state, which in turn depends upon the kinetic parameters. For cell free systems, regulated gene expression has been removed, however, enzyme activity regulation is still operational. We explored this linkage by estimating model parameters from synthetic data using both network structures. We generated synthetic measurements of the substrate S, intermediate M_5 and end product P_1 approximately every 20 min using network A. This data set is similar to published cell free studies both in terms of network coverage, and sampling frequency [23]. We then generated an ensemble of model parameter estimates by minimizing the difference between model simulations and the synthetic data using particle swarm optimization (PSO), starting from random initial parameter guesses. The estimation of kinetic parameters was sensitive to the choice of regulatory structure (Fig. 5). PSO identified an ensemble of parameters that bracketed the mean of the synthetic measurements in less than 1000 iterations when the control structure was correct (Fig. 5A and B). However, with control mismatch (network B simulated with network A parameters), model simulations were not consistent with the synthetic data (Fig. 5C and D). Taken together, these results suggested that we could perhaps simultaneously estimate both parameters and network control architectures, as incorrect control structures would be manifest as poor model fits.

We modified our particle swarm identification strategy to simultaneously search over both kinetic parameters and putative control structures. In addition to our initial networks, we constructed three additional presumptive network models, each with the same enzymatic connectivity but different allosteric regulation of the pathway enzymes (Fig. 6). We then initialized a population of particles, each with one of the five potential regulatory programs and randomized kinetic parameters. Thus, we generated an initial population of particles that had *both* different kinetic parameters as well as different control structures.

We biased the distribution of the particle population according to our *a prior* belief of the correct regulatory program. To this end, we considered three different priors, a uniform distribution where each putative regulatory structure represented 20% of the population and two mixed distributions that were either positively or negatively biased towards the correct structure (network A). In both the positively biased and uniform cases the PSO clearly differentiated between the true or closely related structures and those that were materially different (Fig. 7). As expected, the positively biased population (40% of the initial particle population seeded with network A) gave the best results, where the correct structure was preferentially identified (Fig. 7A). On the other hand, when given a uniform distribution, the PSO approach identified a combination of network A and network C as the most likely control structures (Fig. 7B). Network A and C differ by the regulatory connection between the end product P_2 and enzyme E_1; in network A, end product P_2 was assumed to inhibit E_1, while in network C, end product P_2 activated E_1. Lastly, when the initial population was biased towards incorrect structures (initial population seeded with 90% incorrect structures), the particle swarm *misidentified* the correct allosteric structure (Fig. 7C). Interestingly, while each particle swarm identified parameter sets that minimized the simulation error, the estimated parameter values were not necessarily similar to the true parameters. The angle between the estimated and true parameters was not consistently small across the swarms (identical parameters would give an angle of zero). This suggested that our particle swarm approach identified a *sloppy* ensemble, i.e., parameter estimates that were individually incorrect but collectively exhibited the correct model behavior.

We calculated control program output and scaled metabolic flux for the positively, uniformly and negatively biased particle swarms (Fig. 8). Network A and network C models from the positively (Fig. 8A) and uniformly (Fig. 8B) biased particle swarms showed similar operational patterns, despite differences in kinetic parameters and control structures.

9

While models from the negatively biased population had error values similar to the correct structures in the previous swarms, they have different flux and control profiles (Fig. 8C). In all cases, regardless of network configuration or parameter values, the rate of enzyme decay was small compared to the other fluxes, and all networks had qualitatively similar trends for E_3 and E_6 control. Moreover, consistent with the correct model structure, production of end product P$_1$ was the preferred branch for all model configurations. However, there was variability in P$_2$ production flux across the population of models, especially for the uniform swarm when compared with the other cases. High P$_1$ branch flux resulted in end product inhibition of E_1 in both network A and network C, however in network D and E, high P$_1$ flux induced E_1 activation. These trends were manifested in different flux profiles, where the negatively biased population appeared more uniform across the population compared with the other swarms, and had higher E_1 specific activity. Interestingly, the behavior of network A and network C highlighted an artifact of our integration rule; both a positive or negative feedback connection from P$_2$ to E_1 were ignored because the P$_1$ inhibition of E_1 dominated. Thus, while theoretically distinct, network A and network C appeared operationally to the PSO algorithm to be the same network. On the other hand, networks B, D and E showed distinct behavior that was not consistent with the true network. These architectures exhibited either limited inhibition (network B) or activation (network D and E) of E_1 activity, resulting in significantly different metabolic flux profiles. However, the PSO was able to find low error parameter solutions, despite the mismatch in the control structures (error values similar, but not better than the best network A and network C estimates). Taken together, these results suggested that a uniform sampling approach could potentially yield an unbiassed estimate of both kinetic parameters and control structures. However, the negatively biased particle swarm results illustrated a potential shortcoming of the approach, namely convergence to a local error minimum despite a significantly incorrect control structure. This suggested that estimated

10

model structures will need to be further evaluated, for example by generating falsifiable

experimental designs which could distinguish between low error solutions.

Discussion

In this study, we presented an effective kinetic modeling strategy to dynamically simu-late cell-free biochemical networks. Our proposed strategy integrated traditional kinetic modeling with an effective rules based approach to dynamically describe metabolic reg-ulation and control. We tested this approach by developing kinetic models of hypotheti-cal cell-free metabolic networks. In particular, we tested whether our effective modeling approach could describe classically expected behavior, and second whether we could si-multaneously estimate kinetic parameters and regulatory connectivity, in the absence of specific mechanistic knowledge, from synthetic experimental data. Toward these ques-tions, we explored five hypothetical cell-free networks. In each network, a substrate S was converted to the end products P_1 and P_2 through a series of enzymatically catalyzed reactions, including a branch point at a hypothetical metabolite M_2. Each network also included the same cofactors and cofactor recycle architecture. However, while all five networks shared the same enzymatic connectivity, each had different allosteric regulatory connectivity. We found that simple effective rules, when integrated with traditional enzyme kinetic expressions, could capture complex allosteric patterns such as ultrasensitivity, or non-competitive inhibition in the absence of specific mechanistic information. Moreover, when integrated into network models, these rules captured classical regulatory patterns such as product-induced feedback inhibition. Lastly, we simultaneously estimated kinetic parameters and discriminated between competing regulatory structures, using synthetic data in combination with a modified particle swarm approach. If we considered all putative regulatory architectures to be equally likely, we were able to estimate a *sloppy* ensemble of models with the correct architecture and kinetic parameters.

The proposed modeling strategy shares features with other popular techniques, but also has several key differences. At its core, our effective modeling approach is similar to regulatory constraint-based methods, and to the cybernetic modeling paradigm devel-

12

oped by Ramkrishna and colleagues. Covert, Palsson and coworkers drastically improved the predictability of constraint-based approaches by integrating Boolean rules into the calculation of metabolic fluxes [29]. If the regulated intracellular flux problem is coupled with time-dependent extracellular balances, these models can predict complex behavior such as diauxie growth or the switch between aerobic and anaerobic metabolism. Another important feature of this approach is that it scales with biological complexity. For example, Covert *et al.* showed that a genome-scale model of *E. coli* augmented with a Boolean rule layer, correctly predicted approximately 80% of the outcomes of a high-throughput growth phenotyping experiment in *E. coli*. Further, they showed that they could learn new biology by iteratively refining the model and its associated rules [30]. However, while regulated flux balance analysis is a powerful technique, it does not easily allow the calculation of time-resolved metabolite abundance. Additionally, the Boolean rules which populate the regulatory layer are limited to ON/OFF decisions; for qualitative predictions of gene expression this is a reasonable limitation. However, Boolean rules will likely be less effective at capturing dynamic allosteric regulation in a cell-free metabolic system. On the other hand, the strength of cybernetic models is the integration of optimal metabolic control heuristics with traditional kinetic pathway modeling. Cybernetic models are highly predictive; they have successfully predicted mutant behavior from limited wild-type data [10, 31, 32], steady-state multiplicity [9], strain specific metabolic function [12] and have been used in bioprocess control applications [33]. However, cybernetic control heuristics are not mechanistic, instead they are the output of an optimal decision with respect to a set of hypothetical physiological objectives. Thus, they are abstractions which are difficult to translate into a specific biological mechanism. Our approach addresses the shortcomings of both regulatory constraint-based models and cybernetic models. First, similar to cybernetic models, the core of our approach is a kinetic model. Thus, we are able to directly calculate the time evolution of metabolism, for example the dynamic abundance of

13

network metabolites. Second, similar to regulatory flux balance analysis, our control laws describe specific mechanistic motifs, such as activation or inhibition of enzyme activity. However, our rules are continuous, thus they potentially allow a finer grained description of metabolic regulation and control mechanisms. Lastly, we can naturally incorporate unmodeled factors and categorical factors or combinations thereof into our control law formulations. Though requiring a more complex description of cellular metabolism, our approach may even be extended to simulate cell-based systems by incorporating the same control laws into transcription factor activation and gene expression regulation.

There are several critical questions that should be explored following this proof-of-concept study. It is unclear how parameter identification will scale to genome-scale networks, and second it is unclear how we will identify allosteric connectivity at a genome-scale. The enzymatic connectivity for genome-scale cell-free networks can easily be established by stripping away the growth and cell wall machinery from whole cell genome reconstructions. Then metabolic fluxes can be transformed into kinetic expressions using heuristics such multiple saturation kinetics, which are then modified by our rule-based control variables. This leaves a large number of unknown kinetic constants that must be estimated from time-resolved metabolite measurements. Ensemble modeling is a well-established approach for parameter identification in large-scale deterministic models. Liao and coworkers developed a method that generates an ensemble of kinetic models that all approach the same steady-state, one determined by fluxomics measurements [34]. The best subpopulation of candidate models are selected based on their agreement with further measurements of genetically perturbed systems. Our work relies on heuristic search optimization to identify kinetic models consistent with steady-state and dynamic time-series measurements of cellular species [35–40]. Instead of estimating a single yet highly uncertain parameter set, both approaches estimate an ensemble of parameter sets whose model behavior recapitulates experimental measurements. Here, we showed that

14

particle swarm optimization quickly identified an ensemble of model parameters, at least for proof-of-concept metabolic networks using synthetic data. This suggested that we can expect reasonable model predictions, despite only partial parameter knowledge, as network size grows if we have properly designed experiments. Brown and Sethna showed in a model of signal transduction that good predictions were possible despite only order of magnitude estimates of parameter values [41]. Sethna and coworkers later showed that model performance is often controlled by only a few parameter combinations, a characteristic seemingly universal to multi-parameter models referred to as *sloppiness* [42]. We have also demonstrated *sloppy* behavior in a wide variety of signal transduction processes [35–40]. Thus, given our previous experience with models containing hundreds of unknown parameters, we expect parameter estimation to be a manageable challenge. On the other hand, a critical challenge will be the estimation of allosteric connectivity at a genome scale. The regulation of glycolytic enzymes, such as phosphofructokinase I, has been studied for many years [43, 44]. The allosteric regulation of metabolic enzymes can also be established from organism specific databases, such as EcoCyc [45], or more general allosteric databases, such as the AlloSteric Database [46]. However, for those enzymes that have not been well studied, we will need to infer allosteric interactions from experimental data. In general, the reverse engineering of regulatory network structure from data is a very difficult problem. Recently, Sauer and colleagues have developed a systematic, model-based approach for the identification of allosteric regulation *in vivo* [47]. They tested the effects of many putative allosteric protein-metabolite interactions on the performance of a kinetic model of glycolysis against dynamic metabolomic and fluxomic measurements. A method similar to this may be easily applied to cell-free systems in order to identify relevant *in vitro* allosteric interactions. Because omics measurements of cell-free environments are easy to obtain, identification of large-scale allosteric control structures may be possible. Also, there are many different approaches from the reverse

engineering of gene regulatory networks that perhaps could be adopted to this problem, however this remains an open question. Lastly, the choice of max/min integration rules or the particular form of the transfer functions could be generalized to include other rule types and functions. Theoretically, an integration rule is a function whose domain is a set of transfer function inputs, and whose range is $v \in [0, 1]$. Thus, integration rules other than max/min could be used, such as the mean or the product, assuming the range of the transfer functions is always $f \in [0, 1]$. Alternative integration rules such as the mean might have different properties which could influence model identification or performance. For example, a mean integration rule would be differentiable, which allows derivative-based optimization approaches to be used. The particular form of the transfer function could also be explored. We choose a Hill-like function because of its prominence in the systems and synthetic biology community. However, the only mathematical requirement for a transfer function is that it map a non-negative continuous or categorical variable into the range $f \in [0, 1]$. Thus, many types of transfer functions are possible.

Materials and Methods

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

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

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

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 $r_j(\mathbf{x}, \epsilon, \mathbf{k})$ 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 $\mathbf{x}(t_o) = \mathbf{x}_o$ and $\epsilon(t_o) = \mathbf{1}$ (initially we have 100% cell-free enzyme abundance).

Each reaction rate was written as the product of two terms, a kinetic term ($\bar{r}_j$) and a regulatory term (v_j):

$$r_j(\mathbf{x}, \epsilon, \mathbf{k}) = \bar{r}_j v_j \tag{3}$$

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 ac-

tivity which catalyzes reaction j, and K_{js} denotes the saturation constant for species s in reaction j. The product in Eqn. (4) was carried out over the set of *reactants* for reaction j (denoted as m_j^-).

The allosteric regulation term v_j depended upon the combination of factors which influenced the activity of enzyme i. For each enzyme, we used a rule-based approach to select from competing control factors (Fig. 2). If an enzyme was activated by m metabolites, we modeled this activation as:

$$v_j = \max \left(f_{1j} \left(\mathcal{Z} \right), \ldots, f_{mj} \left(\mathcal{Z} \right) \right) \tag{5}$$

where $0 \leq f_{ij} \left(\mathcal{Z} \right) \leq 1$ was a regulatory transfer function that calculated the influence of metabolite i on the activity of enzyme j. Conversely, if enzyme activity was inhibited by a m metabolites, we modeling this inhibition as:

$$v_j = 1 - \max \left(f_{1j} \left(\mathcal{Z} \right), \ldots, f_{mj} \left(\mathcal{Z} \right) \right) \tag{6}$$

Lastly, if an enzyme had both m activating and n inhibitory factors, we modeled the regulatory term as:

$$v_j = \min \left(u_j, d_j \right) \tag{7}$$

where:

$$u_j = \max_{j^+} \left(f_{1j} \left(\mathcal{Z} \right), \ldots, f_{mj} \left(\mathcal{Z} \right) \right) \tag{8}$$

$$d_j = 1 - \max_{j^-} \left(f_{1j} \left(\mathcal{Z} \right), \ldots, f_{nj} \left(\mathcal{Z} \right) \right) \tag{9}$$

The quantities j^+ and j^- denoted the sets of activating and inhibitory factors for enzyme j. If an enzyme had no allosteric factors, we set $v_j = 1$. There are many possible functional

forms for $0 \leq f_{ij}(\mathcal{Z}) \leq 1$. However, in this study, each individual transfer function took the form:

$$f_i(\mathbf{x}) = \frac{\kappa_{ij}^\eta \mathcal{Z}_j^\eta}{1 + \kappa_{ij}^\eta \mathcal{Z}_j^\eta} \qquad (10)$$

where $\mathcal{Z}_j$ denotes the abundance of the j factor (e.g., metabolite abundance), and κ_{ij} and η are control parameters. The κ_{ij} parameter represents a species gain parameter, while η is a cooperativity parameter (similar to a Hill coefficient). In the case $\eta > 1$, the allosteric interaction displays positive cooperativity. For $\eta < 1$, the interaction is negatively cooperative. Finally, if $\eta = 1$, the interaction displays no cooperativity. The effect of different values of η on reaction rate can be seen in Figure 3. The model equations were encoded using the Octave programming language and solved using the LSODE routine in Octave (v 3.8.1; www.octave.org). In some cases, metabolic fluxes (or other quantities) were scaled according to:

$$\hat{r}_j(t = \tau) = \left(\frac{r_j - \min \mathbf{r}}{\max \mathbf{r} - \min \mathbf{r}} \right) \bigg|_{t=\tau} \qquad (11)$$

where $0 \leq \hat{r}_j(t = \tau) \leq 1$ denotes the scaled value for flux j evaluated at time τ. We have used this scaling in a variety of other contexts [40, 48].

Estimation of model parameters and structures from synthetic experimental data. Model parameters were estimated by minimizing the difference between simulations and synthetic experimental data (squared residual):

$$\min_{\mathbf{k}} \sum_{\tau=1}^{\mathcal{T}} \sum_{j=1}^{\mathcal{S}} \left(\frac{\hat{x}_j(\tau) - x_j(\tau, \mathbf{k})}{\omega_j(\tau)} \right)^2 \qquad (12)$$

where $\hat{x}_j(\tau)$ denotes the measured value of species j at time τ, $x_j(\tau, \mathbf{k})$ denotes the simulated value for species j at time τ, and $\omega_j(\tau)$ denotes the experimental measurement variance for species j at time τ. The outer summation is respect to time, while the in-

19

ner summation is with respect to state. We approximated a realistic model identification scenario, assuming noisy experimental data, limited sampling resolution (approximately 20 minutes per sample) and a limited number of measurable metabolites. We assumed a constant coefficient of variation of 10% for the synthetic data set.

We minimized the model residual using particle swarm optimization (PSO) [49]. PSO uses a *swarming* metaheuristic to explore parameter spaces. A strength of PSO is its ability to find the global minimum, even in the presence of potentially many local minima, by communicating the local error landscape experienced by each particle collectively to the swarm. Thus, PSO acts both as a local and a global search algorithm. For each iteration, particles in the swarm compute their local error by evaluating the model equations using their specific parameter vector realization. From each of these local points, a globally best error is identified. Both the local and global error are then used to update the parameter estimates of each particle using the rules:

$$\mathbf{\Delta}_i = \theta_1 \mathbf{\Delta}_i + \theta_2 \mathbf{r}_1 \left(\mathcal{L}_i - \mathbf{k}_i \right) + \theta_3 \mathbf{r}_2 \left(\mathcal{G} - \mathbf{k}_i \right) \tag{13}$$

$$\mathbf{k}_i = \mathbf{k}_i + \mathbf{\Delta}_i \tag{14}$$

where $\mathbf{\Delta_i}$ denotes the perturbation to the vector of parameters $\mathbf{k_i}$ for particle i. $(\theta_1, \theta_2, \theta_3)$ are adjustable parameters, $\mathcal{L}_i$ denotes the best local solution found by particle i, and $\mathcal{G}$ denotes the best solution found over the entire population of particles. The quantities r_1 and r_2 denote uniform random vectors with the same dimension as the number of unknown model parameters ($\mathcal{K} \times 1$). In this study, we used $(\theta_1, \theta_2, \theta_3) = (1.0, 0.05564, 0.02886)$. The quality of parameter estimates was measured using two criteria, goodness of fit (model residual) and angle between the estimated parameter vector $\mathbf{k}_j$ and the true parameter set $\mathbf{k}^*$:

$$\alpha_j = cos^{-1} \left(\frac{\mathbf{k}_j \cdot \mathbf{k}^*}{\|\mathbf{k}_j\| \, \|\mathbf{k}^*\|} \right) \tag{15}$$

20

If the candidate parameter set $\mathbf{k}_j$ were perfect, the residual between the model and synthetic data and the angle between $\mathbf{k}_j$ and the true parameter set $\mathbf{k}^*$ would be equal to zero.

We modified our PSO implementation to simultaneously search over kinetic parameters and putative model control structures. In the combined case, each particle potentially carried a different model realization in addition to a different kinetic parameter vector. We kept the update rules the same (along with the update parameters). Thus, each particle competed on the basis of goodness of fit, which allowed different model structures to contribute to the overall behavior of the swarm. We considered five possible model structures (A through E), where network A was the correct formulation (used to generate the synthetic data). We considered a population of 100 particles, where each particle in the swarm was assigned a model structure, and a random parameter vector. The PSO algorithm, model equations, and the objective function were encoded and solved in the Octave programming language (v 3.8.1; www.octave.org).

Acknowledgements

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-

22

473   bernetic models: application to recombinant yeast co-consuming glucose and xylose.

474   Biotechnol Bioeng 103: 984-1002.

475   12. Song HS, Ramkrishna D (2011) Cybernetic models based on lumped elementary

476   modes accurately predict strain-specific metabolic function. Biotechnol Bioeng 108:

477   127-40.

478   13. Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-

479   phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 10:

480   291-305.

481   14. Edwards JS, Palsson BO (2000) The escherichia coli mg1655 in silico metabolic

482   genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci U S

483   A 97: 5528-33.

484   15. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A genome-

485   scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for

486   1260 orfs and thermodynamic information. Mol Syst Biol 3: 121.

487   16. Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R (2007) Genome-scale re-

488   construction of metabolic network in bacillus subtilis based on high-throughput phe-

489   notyping and gene essentiality data. J Biol Chem 282: 28791-9.

490   17. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ (2009) Reconstruction of bio-

491   chemical networks in microorganisms. Nat Rev Microbiol 7: 129-43.

492   18. Ibarra RU, Edwards JS, Palsson BO (2002) Escherichia coli k-12 undergoes adaptive

493   evolution to achieve in silico predicted optimal growth. Nature 420: 186-9.

494   19. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions

495   for predicting intracellular fluxes in escherichia coli. Mol Syst Biol 3: 119.

496   20. Hyduke DR, Lewis NE, Palsson BØ (2013) Analysis of omics data with genome-scale

497   models of metabolism. Mol Biosyst 9: 167-74.

498   21. McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale

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

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

502 23. Jewett MC, Calhoun KA, Voloshin A, Wuu JJ, Swartz JR (2008) An integrated cell-
503   free metabolic platform for protein production and synthetic biology. Mol Syst Biol 4:
504   220.

505 24. MATTHAEI JH, NIRENBERG MW (1961) Characteristics and stabilization of dnaase-
506   sensitive protein synthesis in e. coli extracts. Proc Natl Acad Sci U S A 47: 1580-8.

507 25. NIRENBERG MW, MATTHAEI JH (1961) The dependence of cell-free protein syn-
508   thesis in e. coli upon naturally occurring or synthetic polyribonucleotides. Proc Natl
509   Acad Sci U S A 47: 1588-602.

510 26. Lu Y, Welsh JP, Swartz JR (2014) Production and stabilization of the trimeric influenza
511   hemagglutinin stem domain for potentially broadly protective influenza vaccines. Proc
512   Natl Acad Sci U S A 111: 125-30.

513 27. Hodgman CE, Jewett MC (2012) Cell-free synthetic biology: thinking outside the cell.
514   Metab Eng 14: 261-9.

515 28. Morris MK, Saez-Rodriguez J, Clarke DC, Sorger PK, Lauffenburger DA (2011) Train-
516   ing signaling pathway maps to biochemical data with constrained fuzzy logic: quan-
517   titative analysis of liver cell responses to inflammatory stimuli. PLoS Comput Biol 7:
518   e1001099.

519 29. Covert MW, Schilling CH, Palsson B (2001) Regulation of gene expression in flux
520   balance models of metabolism. J Theor Biol 213: 73-88.

521 30. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-
522   throughput and computational data elucidates bacterial networks. Nature 429: 92-6.

523 31. Varner JD (2000) Large-scale prediction of phenotype: concept. Biotechnol Bioeng
524   69: 664-78.

32. Song HS, Ramkrishna D (2012) Prediction of dynamic behavior of mutant strains from limited wild-type data. Metab Eng 14: 69-80.

33. Gadkar KG, Doyle FJ 3rd, Crowley TJ, Varner JD (2003) Cybernetic model predictive control of a continuous bioreactor with cell recycle. Biotechnol Prog 19: 1487-97.

34. Tran LM, Rizk ML, Liao JC (2008) Ensemble modeling of metabolic networks. Biophys J 95: 5606-17.

35. Luan D, Zai M, Varner JD (2007) Computationally derived points of fragility of a human cascade are consistent with current therapeutic strategies. PLoS Comput Biol 3: e142.

36. Song SO, Varner J (2009) Modeling and analysis of the molecular basis of pain in sensory neurons. PLoS One 4: e6758.

37. Tasseff R, Nayak S, Salim S, Kaushik P, Rizvi N, et al. (2010) Analysis of the molecular networks in androgen dependent and independent prostate cancer revealed fragile and robust subsystems. PLoS One 5: e8864.

38. Tasseff R, Nayak S, Song SO, Yen A, Varner JD (2011) Modeling and analysis of retinoic acid induced differentiation of uncommitted precursor cells. Integr Biol (Camb) 3: 578-91.

39. Nayak S, Siddiqui JK, Varner JD (2011) Modelling and analysis of an ensemble of eukaryotic translation initiation models. IET Syst Biol 5: 2.

40. Lequieu J, Chakrabarti A, Nayak S, Varner JD (2011) Computational modeling and analysis of insulin induced eukaryotic translation initiation. PLoS Comput Biol 7: e1002263.

41. Brown KS, Sethna JP (2003) Statistical mechanical approaches to models with many poorly known parameters. Phys Rev E Stat Nonlin Soft Matter Phys 68: 021904.

42. Machta BB, Chachra R, Transtrum MK, Sethna JP (2013) Parameter space compression underlies emergent theories and predictive models. Science 342: 604-7.

551 43. Berg JM, Tymoczko JL, Stryer L (2002) Biochemistry. W.H. Freeman.

552 44. Peskov K, Goryanin I, Demin O (2008) Kinetic model of phosphofructokinase-1 from
553 escherichia coli. J Bioinform Comput Biol 6: 843-67.

554 45. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, et al. (2013)
555 Ecocyc: fusing model organism databases with systems biology. Nucleic Acids Res
556 41: D605-12.

557 46. Huang Z, Mou L, Shen Q, Lu S, Li C, et al. (2014) Asd v2.0: updated content and
558 novel features focusing on allosteric regulation. Nucleic Acids Res 42: D510-6.

559 47. Link H, Kochanowski K, Sauer U (2013) Systematic identification of allosteric protein-
560 metabolite interactions that control enzyme activity in vivo. Nat Biotechnol 31: 357-
561 61.

562 48. Song SO, Chakrabarti A, Varner JD (2010) Ensembles of signal transduction models
563 using pareto optimal ensemble techniques (poets). Biotechnol J 5: 768-80.

564 49. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of the
565 International Conference on Neural Networks. pp. 1942 - 1948.
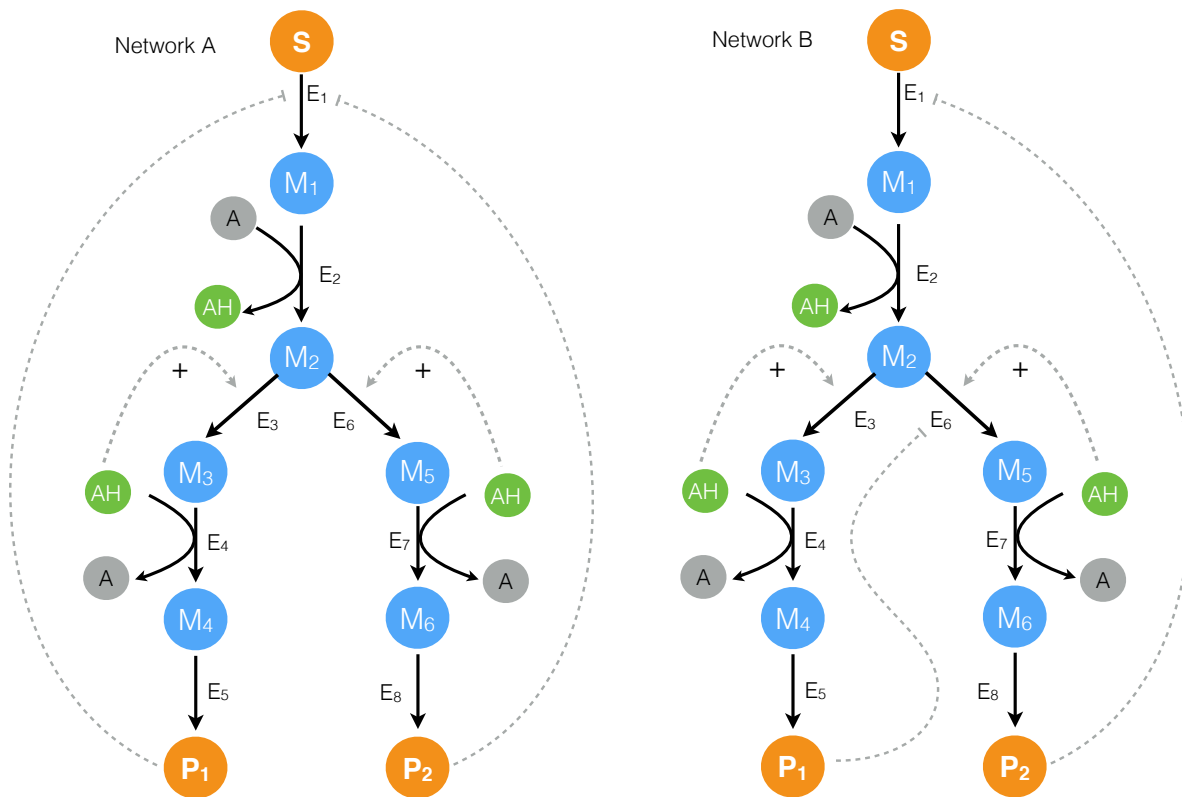
Fig. 1: Proof-of-concept cell-free metabolic networks considered in this study. Substrate S is converted to products P_1 and P_2 through a series of chemical conversions catalyzed by enzyme(s) E_j. The activity of the pathway enzymes is subject to both positive and negative allosteric regulation.
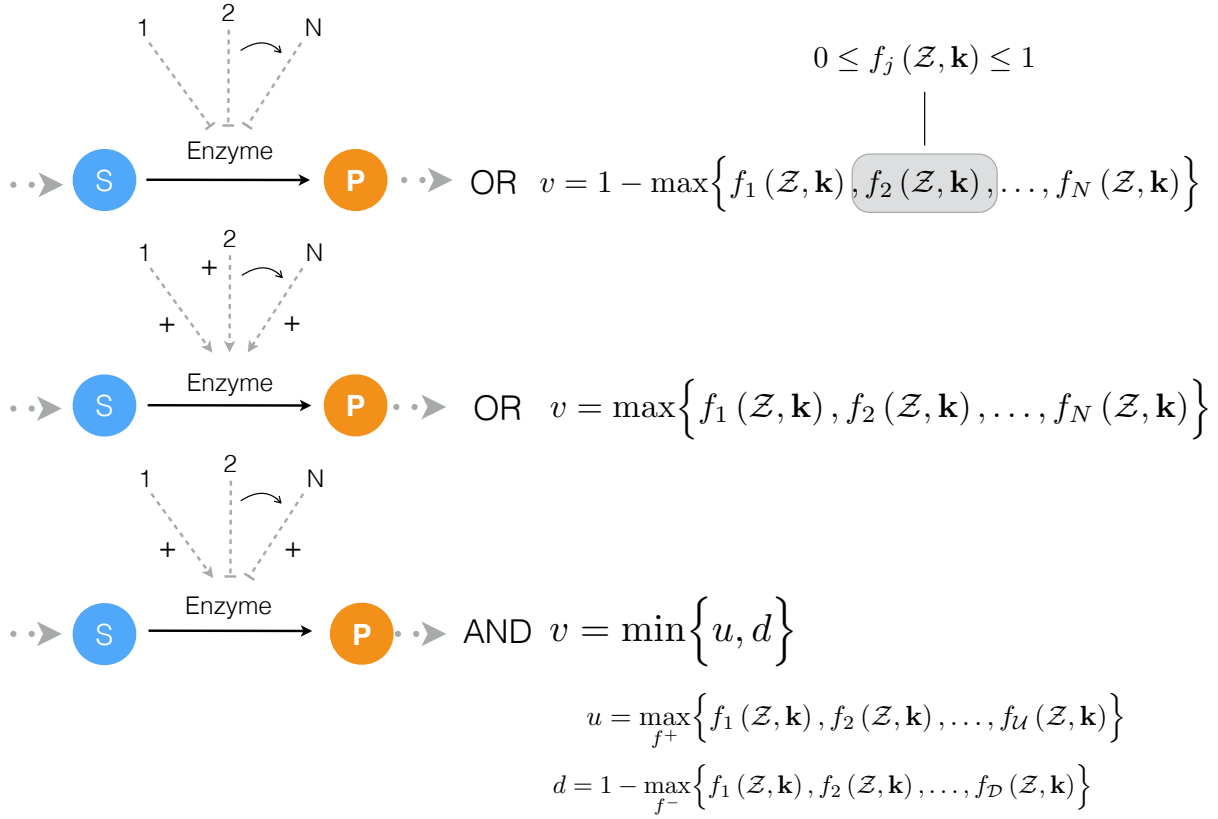
Fig. 2: Schematic of rule-based allosteric enzyme activity control laws. Traditional enzyme kinetic expressions, e.g., Michaelis–Menten or multiple saturation kinetics, are multiplied by an enzyme activity control variable $0 \le v_j \le 1$. Control variables are functions of many possible regulatory factors encoded by arbitrary functions of the form $0 \le f_j(\mathcal{Z}) \le 1$. At each simulation time step, the v_j variables are calculated by evaluating integration rules such as the max or min of the set of factors $f_1, \ldots$ influencing the activity of enzyme E_j.
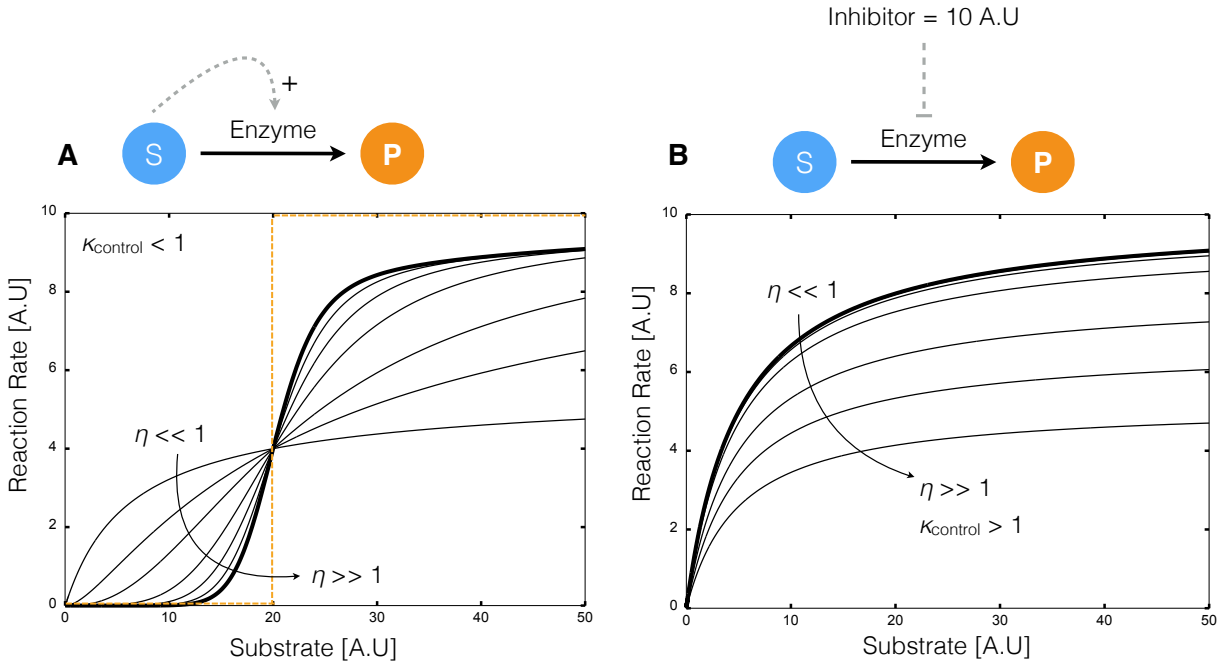
Fig. 3: Kinetics of simple transformations in the presence of activation and inhibition. **A**:The conversion of substrate S to product P by enzyme E was activated by S. For a fixed control gain parameter $\kappa_{control}$, the reaction rate approached a step for increasing cooperativity control parameter η. **B**:The conversion of substrate S to product P by enzyme E with inhibitor I. For a fixed control gain parameter $\kappa_{control}$, the reaction rate approximated non-competitive inhibition for increasing cooperativity control parameter η.
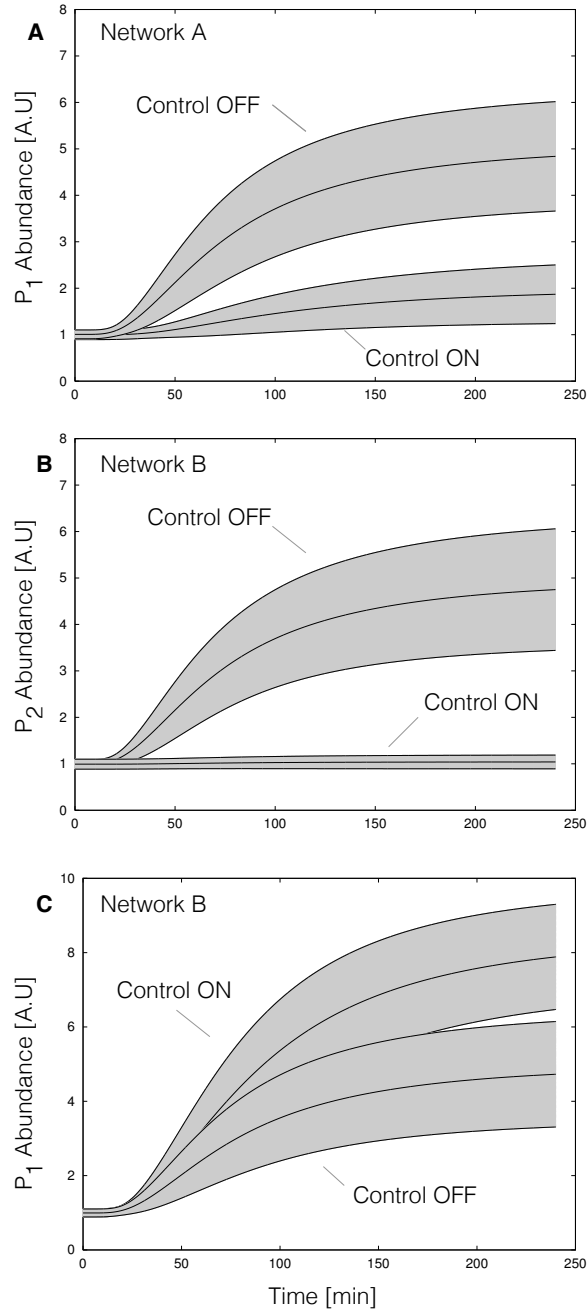
Fig. 4: ON/OFF control simulations for Network A and Network B for an ensemble of 100 kinetic parameter sets versus time. For each case, simulations were conducted using kinetic and initial conditions generated randomly from a hypothetical true parameter set. The gray area represents $\pm$ one standard deviation surrounding the mean. Control parameters were fixed during the ensemble calculations. **A**: End product P_1 abundance versus time for Network A. The abundance of P_1 decreased with end product inhibition of E_1 activity (Control-ON) versus the no inhibition case (Control-OFF). **B**: End product P_2 abundance versus time for Network B. Inhibition of branch point E_6 by end product P_1 decreased P_2 abundance (Control-ON) versus the no inhibition case (Control-OFF). **C**: End product P_1 abundance versus time for Network A. Inhibition of branch point E_6 by end product P_1 decreased P_1 abundance (Control-ON) versus the no inhibition case (Control-OFF).
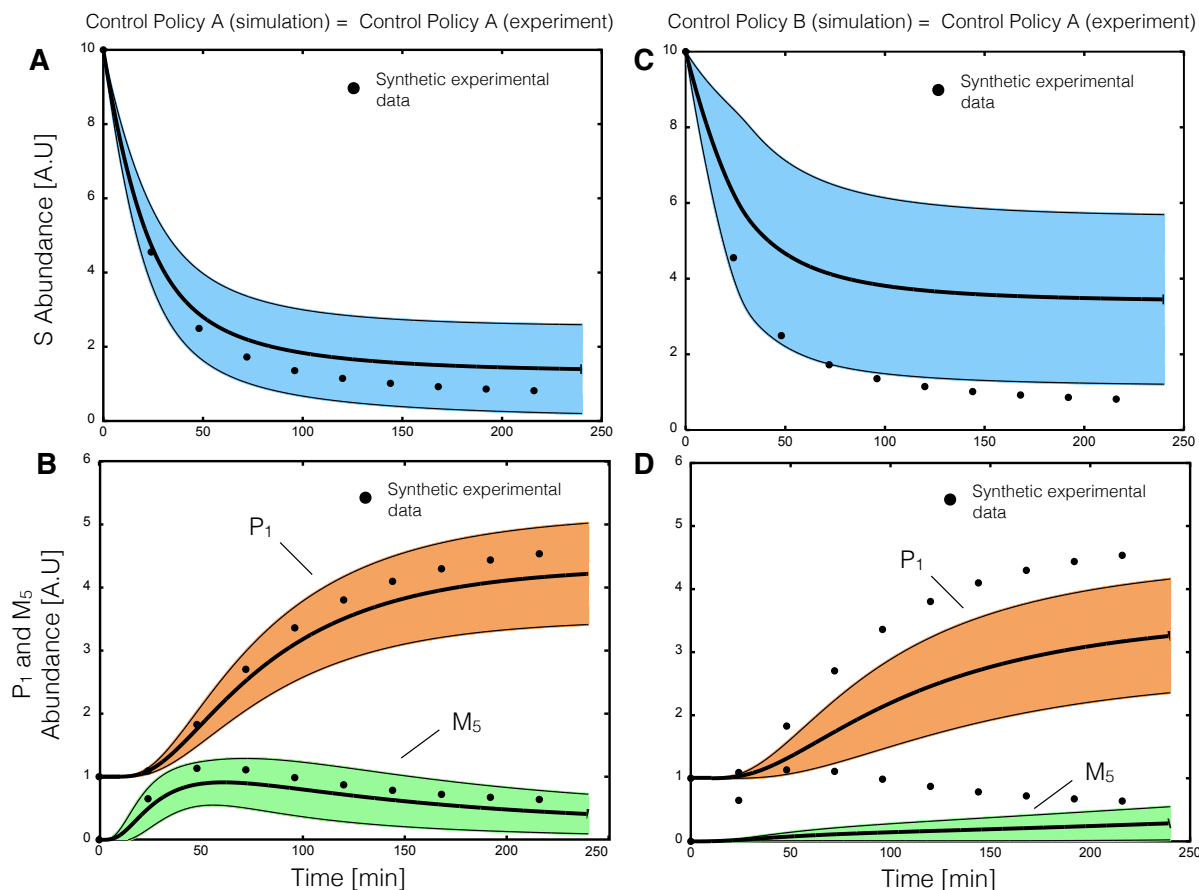
Fig. 5: Parameter estimation from synthetic data for the same and mismatched allosteric control logic using particle swarm optimization (PSO). Synthetic experimental data was generated from a hypothetical parameter set using Network A, where substrate S, end product P_1 and intermediate M_5 were sampled approximately every 20 minutes. For cases **A,B** 20 particles were initialized with randomized parameters and allowed to search for 300 iterations. **A,B**: PSO estimated an ensemble of 20 parameters sets consistent with the synthetic experimental data assuming the correct enzymatic and control connectivity starting from randomized initial parameters. **C,D**: In the presence of control mismatch (Network B control policy simulated with Network A kinetic parameters) the ensemble of models did not describe the synthetic data. The synthetic data plotted here was unperturbed by noise. However, we assumed a coefficient of variation of 10% for the synthetic data during parameter estimation.
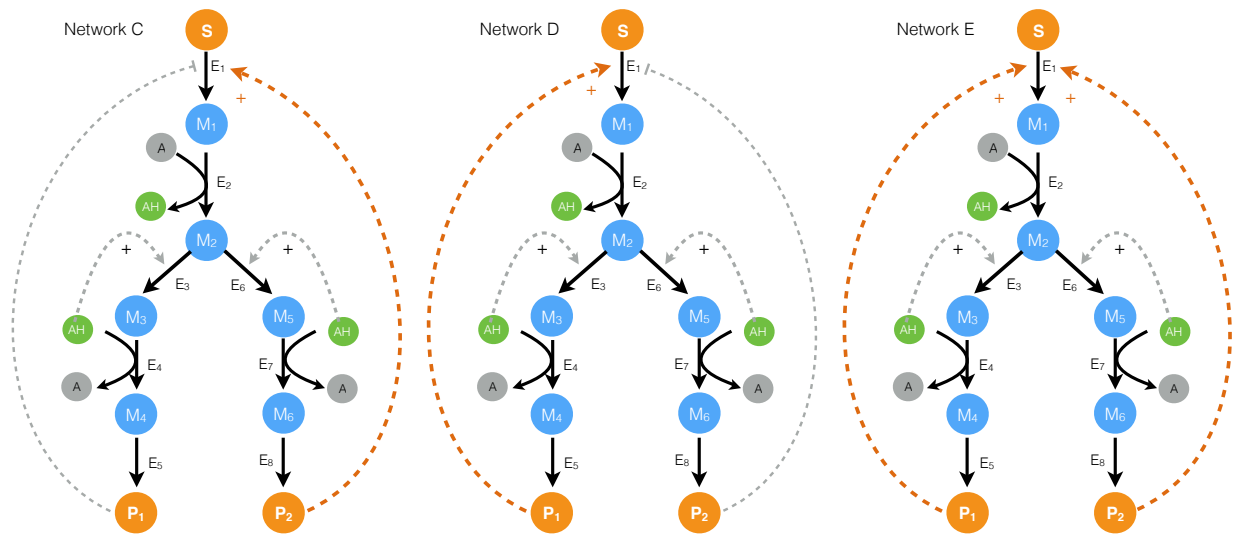
31

Fig. 6: Schematic of the alternative allosteric control programs used in the structural particle swarm computation. Each network had the same enzymatic connectivity, initial conditions and kinetic parameters, but alternative feedback control structures for the first enzyme in the pathway.
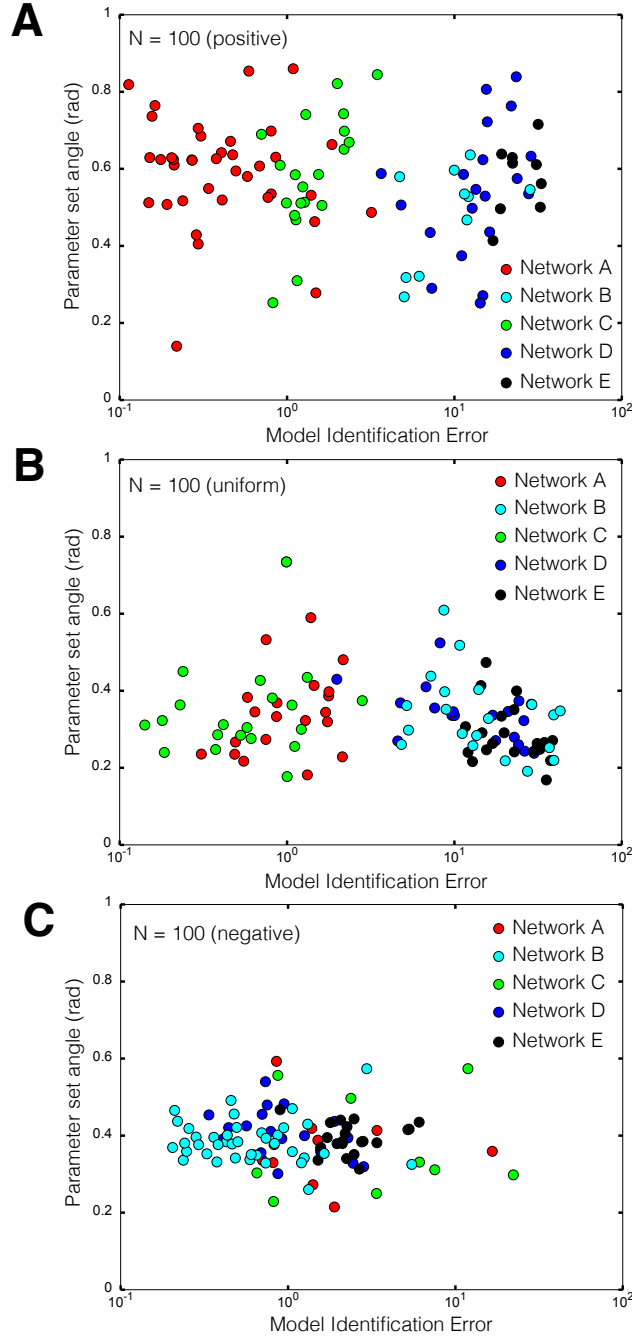
Fig. 7: Combined control and kinetic parameter search using modified particle swarm optimization (PSO). A population of 100 particles was initialized with randomized kinetic parameters and one of five possible control configurations (Network A - E). Simulation error was minimized for a synthetic data set (S, end product P_1 and intermediate M_5 sampled approximately every 20 min) generated using Network A. **A**: Simulation error versus parameter set angle for 100 particles biased toward the correct regulatory program (A,B,C,D,E) = (40%, 10%, 20%, 20% and 10%). **B**: Simulation error versus parameter set angle for 100 uniformly distributed particles (A,B,C,D,E) = (20%, 20%, 20%, 20% and 20%). **C**: Simulation error versus parameter set angle for 100 negatively biased particles (A,B,C,D,E) = (10%, 40%, 10%, 20% and 20%). Network A (the correct structure) was preferentially identified for positively and uniform biased particle distributions, but misidentified in the presence of a large incorrect bias.
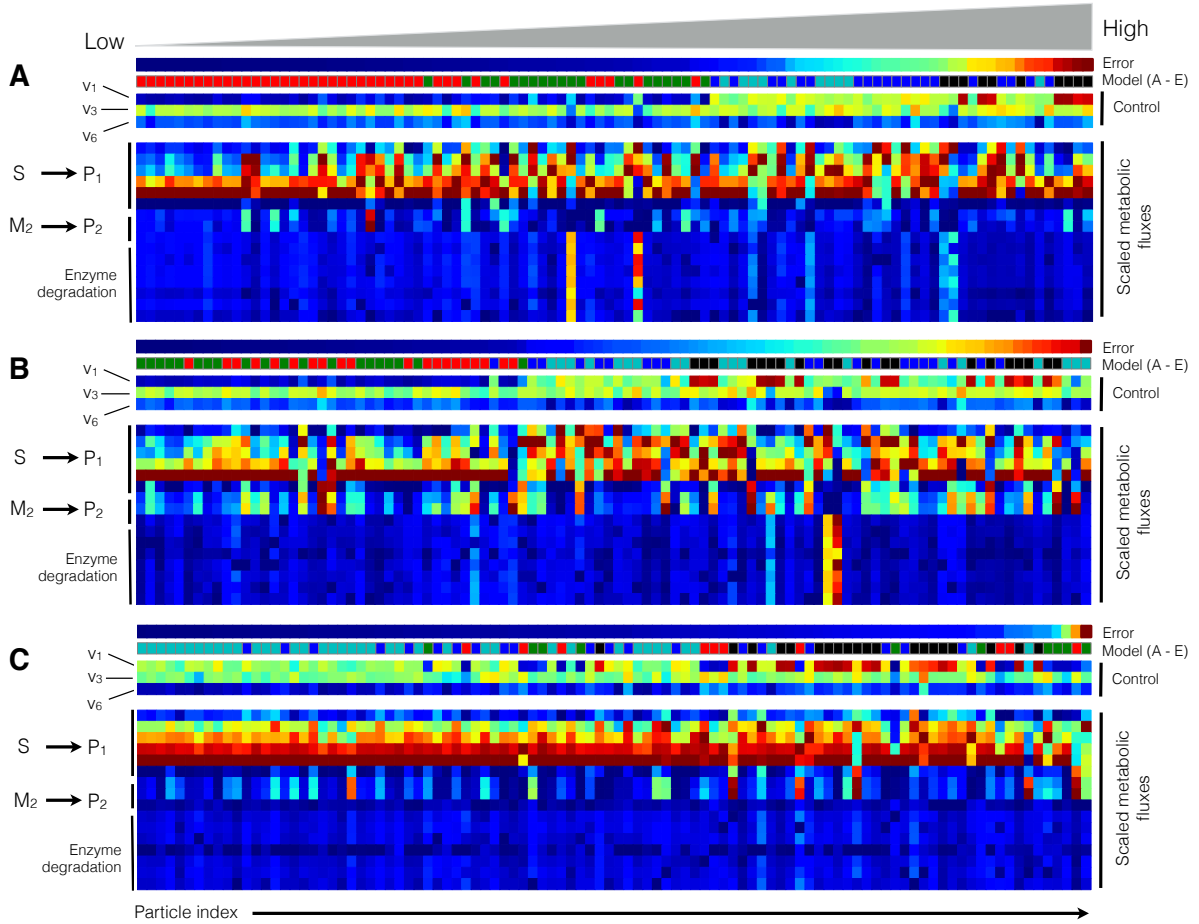
Fig. 8: Metabolic flux and control variables as a function of network type and particle index at t = 100 min. The control variables governing E_1, E_3 and E_6 activity and the scaled metabolic flux were calculated for the positively (top), uniformly (middle) and negatively (bottom) biased particle swarms (N = 100). Blue denotes a low value, while red denotes a high value for the respective quantity being plotted. The particles from each swarm were sorted based upon simulation error (low to high error). **A**: Model performance for the positively biased particle swarm as a function of particle index. **B**: Model performance for the uniformly biased particle swarm as a function of particle index. **C**: Model performance for the negatively biased particle swarm as a function of particle index. Models with significant control mismatch showed distinct control and flux patterns versus those models with the correct or closely related control policies. In particular, models with the correct control policy showed stronger inhibition of E_1 activity, leading to decreased flux from S→P_1. Conversely, models with significant mismatch had increased E_1 activity, leading to an altered flux distribution. This is especially apparent in the negatively biased particle swarm.