On the problem of identifying metabolic network dynamics from steady-state data: metrics and constraints

Elif Köksal, Eugenio Cinquemani

On the problem of identifying metabolic network dynamics from steady-state data: metrics and constraints

Elif Köksal and Eugenio Cinquemani

Abstract— **We address inference of dynamic metabolic network models from steady-state metabolite concentration and flux data. As a case study we rely on simulation of an example branched metabolic pathway. In general, steady-state data fitting does not guarantee appropriate dynamical performance. In order to tackle this problem, we propose tools for the assessment of the accuracy of dynamical model approximations, and suggest the use of constraints on the expected dynamical system behavior, in the form of reaction-rate constraints over a region of interest, to improve the dynamical properties of the model inferred. Numerical results demonstrate the potential of the approach.**

I. INTRODUCTION

The metabolism of a cell is the family of chemical reactions that govern the transformation of substrates (e.g. sugars) into products (ATPs, aminoacids, etc.) necessary for the functioning of the cell. Metabolic reactions are typically mediated by enzymes, proteins whose production via gene expression is controlled by regulatory mechanisms involving metabolic co-factors. Mathematical modeling of metabolism is thus essential for understanding cellular physiology and engineering modified cell functions.

Mathematical modeling of metabolic dynamics has a long history [1]. For a well-characterized network, reaction kinetics can be written in terms of nonlinear parametric models (ODEs), that may in turn be used to explore the dynamic response of metabolism and the sensitivity of steady-states to environmental perturbations, the role of variable enzyme concentration on the metabolic fluxes (i.e. reaction rates at steady state), etc. Even for well-characterized reaction networks, however, kinetic parameters are often unknown and need to be determined from experimental data.

Inference of metabolic models from experiments poses several challenges. From the experimental viewpoint, a popular approach is to observe the variation of the system steady-state in response to perturbations of enzyme or external metabolite concentrations (see e.g. [17]). Metabolite concentrations are measured in steady-state, and the corresponding fluxes are quantified by the help of flux balance analysis (see e.g. [13] and references therein). In this case, the parameters of a dynamical model are estimated from steady-state metabolite concentrations and associated fluxes, separately for every reaction [10], [6]. Alternatively, metabolite concentrations can be observed dynamically [20], but the corresponding reaction rates are typically unavailable. In this case, model inference exploits time course data of metabolite concentrations only (see e.g. [16]), leading to a single, but more complicated, estimation problem. Often, data are corrupted by outliers and missing entries, certain metabolites are not observed, and a model reduction step is needed to cope with this lack of information. Finally, full-blown models of metabolism are nonlinear.

All of the above motivates the interest in approximate models of metabolic dynamics such as linear, lin-log, power-law models [4], [6], [3], to name a few. Motivated by different mathematical or physical arguments, these model classes provide tractable approximations of metabolic dynamics at least in a region of interest, are amenable to biological interpretation and allow for analytic solution, identifiability analysis, and model reduction [2], [9]. Yet, the inherent discrepancy from the real system dynamics raises questions about the accuracy of models for dynamic predictions, especially if they are inferred from steady-state data. Simulation-based evaluation of approximate models has been performed e.g. in [14], [16], [15], showing the potential of different modelling formalisms but also emphasizing the limits of validity of the approximations and the importance of costraints at the inference stage.

In this paper, we address the problem of estimating dynamic models of metabolism from steady-state metabolite concentration and flux data. In general, accurate steady-state data fit does not imply good dynamic modeling overall. The objective of the paper is to propose tools for the quantitative assessment of dynamical approximations of metabolic networks, and methods for improving estimation of dynamic metabolic models from steady-state data.

We rely on a small, yet biologically relevant case study, the synthetic branched metabolic pathway of [11], and consider inference of approximate dynamic models from steady-state data obtained by *in-silico* perturbation of the system. We propose vector-field metrics for analyzing the dynamic properties of the models, and investigate optimization constraints as a means for ameliorating dynamic model inference from steady-state data. We propose constraints on vector-field properties, such as reaction rates away from steady-state, as an alternative to individual parameter constraints (see e.g. [5]) or steady-state flux constraints (see e.g. [13]), and show that they may lead to better dynamical model approximations overall. While the discussion is carried out with reference to lin-log and power-law models, most of the tools proposed and the considerations drawn apply equally well to different metabolic model classes.

E. Cinquemani and E. Köksal are with INRIA Grenoble – Rhône-Alpes, 655 Avenue de l'Europe, Montbonnot, St Ismier Cedex, France.

Corresponding author email eugenio.cinquemani@inria.fr.

II. MATHEMATICAL FRAMEWORK

A. Modeling Approaches

Kinetic models of biochemical reactions are described by ODEs [1]. The general formulation of metabolic network dynamics with enzyme-catalyzed reactions is

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}) = \mathbf{N} \cdot \mathbf{v} \tag{1}$$

$$v_h = e_h \cdot f_h(\mathbf{x}, \mathbf{u}, \mathbf{p}_h), \qquad h = 1, \ldots, m \tag{2}$$

where $\mathbf{x} \in \Re^{n_x}$ denotes the vector of internal metabolite concentrations of a n_x-dimensional system described by $\mathbf{F}(x)$, which can be written as a multiplication of the vector of reaction rates $\mathbf{v} \in \Re^m$ and a stoichiometry matrix $\mathbf{N} \in Z^{n_x \times m}$. For every reaction $h = 1 \ldots, m$, v_h is proportional to enzyme activity $e_h \in \Re$ and a nonlinear function f_h of internal metabolite concentrations $\mathbf{x}$, external metabolite concentrations $\mathbf{u} \in \Re^{n_u}$ and kinetic parameters $\mathbf{p}_h$.

A fully parametrized form of the functions f_h is often difficult to manipulate and to determine from experimental data. Therefore, approximate modeling formalisms are routinely used to describe the behavior of the network around reference conditions of interest. Among several approximate kinetic models of the nonlinear function f, lin-log and Generalized Mass Action (GMA) power-law models are widely used [2]. In lin-log kinetics, inspired by thermodynamic principles and in connection with metabolic control analysis [2], [6], the rate of every reaction is assumed proportional to a sum of logarithms of metabolite concentrations. That is, dropping index h for simplicity,

$$\frac{v}{e} = f(\mathbf{x}, \mathbf{u}, \mathbf{p}) = c + \sum_i a_i \ln(x_i) + \sum_j b_j \ln(u_j), \tag{3}$$

with parameters $\mathbf{p} = (\mathbf{a}, \mathbf{b}, c)$. In the GMA power-law approach, inspired by mass-action kinetics, reaction rates are proportional to a product of enzyme levels and metabolite concentrations with kinetic exponents (4),

$$\frac{v}{e} = f(\mathbf{x}, \mathbf{u}, \mathbf{p}) = \gamma \prod_i x_i^{\alpha_i} \prod_j u_j^{\beta_j}, \tag{4}$$

with parameters $\mathbf{p} = (\alpha, \beta, \gamma)$, where $\gamma > 0$. For every reaction, in (3) and (4), only metabolites that participate in the reaction are included. Both models are well defined as long as $\mathbf{x} > 0$ and $\mathbf{u} > 0$, whereas power-law may cope with null metabolite concentrations, and is more generally believed to be better suited than lin-log models for metabolite concentrations that may get small values [7]. In this work we will only consider these two formalisms, and assume that the enzyme concentrations $\mathbf{e}$ are known and fixed. Since their values can be absorbed into the model parameters, without loss of generality we set $\mathbf{e} = 1$ for all reactions, and write $v_h(\mathbf{x}, \mathbf{u}, \mathbf{p})$ without reporting the dependence of reaction rates on enzyme concentrations.

B. Estimation Problem

The problem we consider is that of reconstructing an approximate metabolic network model around a known reference point $\mathbf{v}^*, \mathbf{x}^*, \mathbf{u}^*$, based on simultaneous measurements of $\mathbf{x}$ and $\mathbf{v}$ in different steady states obtained by K separate constant perturbations of the inputs $\mathbf{u}$. We assume $\mathbf{N}$ to be known. Using the assumption that the models must satisfy (2) at $\mathbf{x}^*$, $\mathbf{u}^*$, $\mathbf{v}^*$, one gets the relations

$$v - v^* = f_{LL}(\mathbf{x}, \mathbf{u}, \mathbf{p}) \triangleq \sum_i a_i \ln(\frac{x_i}{x_i^*}) + \sum_j b_j \ln(\frac{u_j}{u_j^*}), \tag{5}$$

$$\ln(\frac{v}{v^*}) = f_{PL}(\mathbf{x}, \mathbf{u}, \mathbf{p}) \triangleq \sum_i \alpha_i \ln(\frac{x_i}{x_i^*}) + \sum_j \beta_j \ln(\frac{u_j}{u_j^*}), \tag{6}$$

holding for every individual reaction. Therefore, the problem becomes that of estimating the parameter vectors $\mathbf{a}$ and $\mathbf{b}$, for the lin-log model (5), or α and β, for the power-law model (5). Note that, for given $\mathbf{a}$ and $\mathbf{b}$ (resp. α and β), the value of c (resp. of γ) is determined by Eq. (3) (resp. (4)) evaluated at the reference point. Thus, for shortness, we will generally speak about estimation of $\mathbf{p}$. Given a set of steady-state data points $(\mathbf{x}^k, \mathbf{u}^k, \mathbf{v}^k)$, where $\mathbf{x}^k$ and $\mathbf{v}^k$ are obtained in response to the constant perturbation $\mathbf{u}^k$, with $k = 1, \ldots, K$, let us define $v_{LL}^k = v^k - v^*$ and $v_{PL}^k = \ln(v^k/v^*)$. We consider estimates of the lin-log and power-law parameters drawn by solving the optimization problems (one per reaction)

$$\min_{\mathbf{p}} \sum_{k=1}^{K} \left(v_{LL}^k - f_{LL}(\mathbf{x}^k, \mathbf{u}^k, \mathbf{p})\right)^2, \tag{7}$$

$$\min_{\mathbf{p}} \sum_{k=1}^{K} \left(v_{PL}^k - f_{PL}(\mathbf{x}^k, \mathbf{u}^k, \mathbf{p})\right)^2. \tag{8}$$

In absence of constraints, these are least-squares problems that can be solved explicitly.

Several aspects of the problem deserve discussion. First, in a real context, data for both $\mathbf{x}^k$ and $\mathbf{v}^k$ can only be collected in steady state. On the other hand, one would like to use the resulting model also for dynamical analysis. It is a priori unclear if steady-state data may provide sufficient information for matching the real system dynamics with a necessarily approximate model, or what additional information can be used to enforce well-behaved estimates. Note that, for both modelling frameworks considered here and other model classes, the existence of nonnegative solutions of the ODEs (1)-(2) may not be guaranteed if the models are used to simulate the system far away from $\mathbf{x}^*, \mathbf{u}^*, \mathbf{v}^*$, e.g. for small concentrations or certain combinations of inputs. Finally, models should be biologically relevant, i.e. parameters should reflect the regulatory effect (enhancement or suppression) of the various chemical species on the system reactions. With the aid of a case study, the purpose of this work is to explore these aspects in connection with the expected behavior of the system, i.e. the a priori information that one may use while solving parameter estimation.

III. CASE STUDY

A. The System

We consider the Mendes-Kell synthetic metabolic network of [11], [6]. This branched pathway (Fig. 1) includes 8 in-

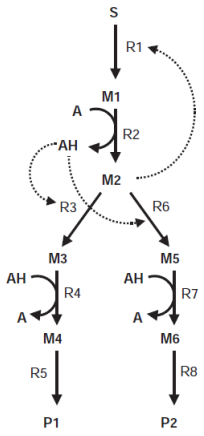

Fig. 1. Representation of the Mendes-Kell metabolic pathway from [6].

ternal, 3 external metabolites and 8 reversible reactions with allosteric interactions represented by Hill, ordered bi-bi and uni-uni kinetics equations. The detailed equations are given in [6]. Using the two conservation rules $[AH] = T_1 - [A]$ and $[M_5] = T_2 - [M_2] - [M_3] - [M_6]$, the system reduces to a 6-dimensional ODE, one per independent internal metabolite, with state $x = ([M_1], [M_2], [M_3], [M_4], [M_6], [A])$ and inputs $u = ([S], [P_1], [P_2])$. At the reference state $\mathbf{x}^* = (1.094, 0.170, 0.042, 0.101, 0.202, 0.019)$mM, obtained for $\mathbf{u}^* = (1.1, 0.1, 0.2)$mM with ($T_1 = 0.1$ and $T_2 = 0.3$), the (locally linearized) system has 6 negative real eigenvalues and is hence asymptotically stable. From now on, we will refer to this simulated system as the real system.

B. Inference of approximate models from simulated data

We consider the inference of a lin-log and a power-law model around the reference point $(\mathbf{x}^*, \mathbf{u}^*, \mathbf{v}^*)$ from steady-state perturbation data. Data consist in the steady-state values of $\mathbf{x}$ and $\mathbf{v}$ resulting from $K = 100$ different external metabolite concentrations $\mathbf{u}$, where each component of $\mathbf{u}$ is drawn at random according to a uniform distribution over $\pm 10\%$ of the corresponding entry of $\mathbf{u}^*$. In practice, data are obtained by simulating the system of Section III-A until steady-state is reached (all experiments we consider are such that reactions are never reversed). Models are estimated by solving problems (7) and (8) in the relevant parameters. The dynamic behavior of the resulting models starting from a state different from the steady-state $\mathbf{x}^*$ is compared to that of the real system in Fig. 2.

Results from this simulated inference enable already some considerations. First of all, steady-state data may allow for the identification of reasonable dynamic models, provided the data is sufficient to make estimation well-posed. Yet, the residual error in fitting the data says little about the goodness of the resulting model dynamics. This can be seen in Fig. 2, where comparable residuals of fit (reported in the caption) give rise to predictions of the system response to perturbed initial conditions of markedly different quality. In other words, an accurate fit may not correspond to an accurate model approximation away from the data points

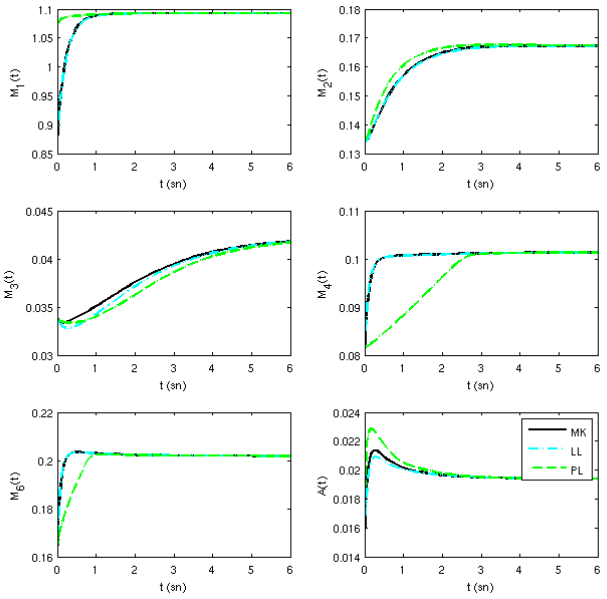

Fig. 2. Dynamic behavior of the original model (MK) and of the lin-log (LL) and power-law (PL) approximations starting from $\mathbf{x} = 0.8\mathbf{x}^*$. Parameters of the LL and PL models are the solution of (7) and (8), with fitting residuals 7.70E-5 and 1.10E-4, respectively.

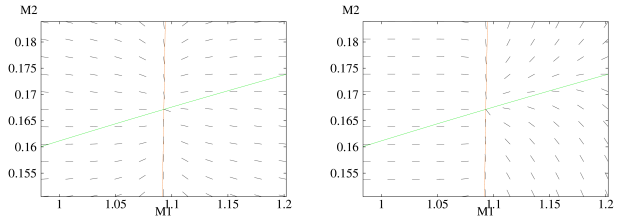

Fig. 3. Vector fields of lin-log (left) and power-law (right) models of Fig. (2). Segments indicate the local vector field direction, oriented toward the equilibrium point at the intersection of the nullclines of M_1 and M_2 (green and red lines, respectively). All other states are kept at the reference point. Vector field magnitude, not represented, increases much faster for the power-law model when moving away from the equilibrium.

themselves. As an example, the appearence of the vector field for M_1 and M_2 of the lin-log and power-law models in the vicinity of the reference point is shown in Fig. 3. The different behavior when moving away from the reference point is apparent.

The observations above are easily explained in terms of the relation between (approximate) modeling and limited informativeness of the data. If the parametric model class used for model inference contained the real system, then a well-posed estimation problem (data is enough to determine all unknown parameters) would return a model that is identical to the original system. Unfortunately, in general, the real system does not belong to these model classes. Even a well-posed estimation problem ensures that the model resembles the original system only in proximity of the data points. Away from the data points, the model

behavior is determined by the model structure and not by the original system. This is especially an issue if the data is limited (few or concentrated data points, or obeying implicit constraints, which may typically be the case for steady-state data [9]) and if the structure of the model class is not a good replica of the real system (a fact that, in general, cannot be determined a priori). Still, we argue that results can be improved significantly by a better statement of the estimation problem, even if the data remains limited and the model class suffers from limited approximation capabilities. The core of what follows is to show that introducing constraints on the expected dynamics of the system allows one to compensate for structural approximations and the loose exploration of the system provided by the data, thus making the estimated model better behaved at a negligible price in terms of data fit, and that in cases of interest the constraints can be formulated in terms of linear, hence tractable, optimization inequalities.

IV. IMPROVING MODEL INFERENCE

Model inference discussed so far was based purely on data fitting, and was shown to have uncertain implications on the model ability to replicate the original system dynamics. From now on, we address the dynamic performance of the models inferred from steady-state data more systematically. We introduce tools for quantifying the accuracy of the dynamical approximation of the system, consider a variety of common and original estimation constraints to account for prior information on the expected system dynamics, and discuss their effects on the estimated system dynamics.

A. Vector Field Metrics

Let $\mathbf{F}(\mathbf{x})$ be the vector field that determines the original system dynamics $\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x})$. Let $\hat{\mathbf{F}}(\mathbf{x})$ be the vector field of an approximating model, as it can be obtained by inferring e.g. a lin-log or power-law model from data. To quantify the difference between the original and approximating model at $\mathbf{x}$, we define several quantities. First we define the vector-length difference at $\mathbf{x}$

$$l(\mathbf{x}) = \|\mathbf{F}(\mathbf{x})\| - \|\hat{\mathbf{F}}(\mathbf{x})\|,$$

where $\|\cdot\|$ denotes the L_2 norm, and the cosine between vector fields at $\mathbf{x}$,

$$\theta(\mathbf{x}) = \frac{\mathbf{F}(\mathbf{x})^T \hat{\mathbf{F}}(\mathbf{x})}{\|\mathbf{F}(\mathbf{x})\| \cdot \|\hat{\mathbf{F}}(\mathbf{x})\|}.$$

Based on this, we quantify the difference between models over a region X of interest by the indices

$$L(X) = \frac{1}{vol(X)} \int_X |\, l(\mathbf{x}) \,| d\mathbf{x},$$

and

$$\Theta(X) = \frac{1}{vol(X)} \int_X \theta(\mathbf{x}) d\mathbf{x}.$$

By these definitions, $L(X) \geq 0$ and $\Theta(X) \in [-1, 1]$ quantify, in the order, the average discrepancy of the magnitude and the average alignment of the vector fields over X. The closer $L(X)$ is to 0 and the closer $\Theta(X)$ is to 1, the more similar the vector fields. $\hat{\mathbf{F}}$ is a perfect approximation of $\mathbf{F}$ over X if and only if $L(X) = 0$ and $\Theta(X) = 1$. In practice, an approximation of integrals $L(X)$ and $\Theta(X)$ shall be computed by gridding X. We note that a distance such as

$$\frac{1}{vol(X)} \int_X \frac{\|\mathbf{F}(\mathbf{x}) - \hat{\mathbf{F}}(\mathbf{x})\|}{\|\mathbf{F}(\mathbf{x})\| + \|\hat{\mathbf{F}}(\mathbf{x})\|} d\mathbf{x} \in [0, 1],$$

would suffice to quantify the difference between $\hat{\mathbf{F}}$ and $\mathbf{F}$ over X. However, decoupling vector-field magnitude and alignment information makes the analysis more informative, as we will see in Section V.

B. Use of Prior Information

In addition to experimental data, depending on the physical or biological interpretation of the model, a priori belief on the model parameters may be available. Parameter signs may be fixed based on supposed reaction kinetics (a typical requirement for power-law models [3]) and/or elasticities [8], i.e. the sensitivity of reaction rates to parameter variations (most relevant to lin-log models [2]),

$$(-1)^{s_\ell} p_\ell \geq 0, \tag{9}$$

where p_ℓ denotes the generic ℓth parameter and $s_\ell \in \{0, 1\}$. In addition, in power-law models, typical values for the exponents α and β are suggested in the literature, inspired by kinetic reaction orders of biochemical system theory. Here kinetic orders are mostly within the range of $[-1, 3]$ [5], but exceptions are possible (see e.g. [12]). Similar constraints can be considered for lin-log model parameters (though e.g. in [6] only sign constraints are considered). Boundaries on parameter values are captured by constraints of the form

$$p_\ell \in [\underline{p}_\ell, \overline{p}_\ell]. \tag{10}$$

These constraints are linear, hence tractable, but focus on the parameter properties and not on the dynamics that may result from them. We propose that constraint on the maximal reaction rates may provide a more explicit information about what the system dynamics should look like. For a given reaction rate v_h, consider constraints of the form

$$\frac{v_h(\mathbf{x}, \mathbf{u}, \mathbf{p}_h)}{v_h^*} \in [\underline{r}_h, \overline{r}_h], \quad \forall (\mathbf{x}, \mathbf{u}) \in S, \tag{11}$$

where $S = [\underline{s}_1, \overline{s}_1] \times \ldots \times [\underline{s}_{n_x+n_u}, \overline{s}_{n_x+n_u}]$ is a hyper-rectangular region. In particular, one may choose S such that $(\mathbf{x}^*, \mathbf{u}^*) \in S$ and $1 \in [\underline{r}_h, \overline{r}_h]$, so that $\underline{r}_h$ and $\overline{r}_h$ express maximal allowable reaction rate deviations from the reference flux v^* over S. We have the following result, where we drop h from the notation for simplicity.

Proposition 1: For both lin-log and power-law models, constraints (11) are equivalent to the finite set of linear constraints on $\mathbf{p}$ given by

$$\underline{r} \leq \frac{v(\mathbf{x}, \mathbf{u}, \mathbf{p})}{v^*} \leq \overline{r} \quad \forall (\mathbf{x}, \mathbf{u}) \in V(S), \tag{12}$$

where $V(S)$ is the ($2^{n_x+n_u}$-dimensional) vertex set of S.

Proof: For lin-log models, for every fixed $\mathbf{x}$ and $\mathbf{u}$, constraint (12) is already linear in the parameters. For power-law models, note that v/v^* is always nonnegative. Hence we may assume $\underline{r} \geq 0$ without loss of generality, and, thanks to monotonicity of the logarithm function, (12) is equivalent to

$$\ln \underline{r} \leq \ln \frac{v(\mathbf{x}, \mathbf{u}, \mathbf{p})}{v^*} \leq \ln \overline{r} \quad \forall (\mathbf{x}, \mathbf{u}) \in V(S), \qquad (13)$$

which is again linear in the parameters. The fact that (11) implies (12) is obvious. The reverse implication follows from the monotonicity of (3) and (4) in each of the entries of $\mathbf{x}$ and $\mathbf{u}$. ∎

Therefore, all constraints in this section can be easily integrated into the optimization problems (7) and (8) in the form of a finite set of linear constraints. Analogous results can be established for a variety of modelling formalisms (e.g. all pseudo-linear models).

V. Model inference using prior information: A numerical study

We perform a simulated study of the estimation of approximate models. We consider 10 datasets of steady-state measurements. Each dataset comprises 100 steady-state values $\mathbf{x}^k$ and $\mathbf{v}^k$ computed by solving the original system of Section III-A from the initial condition $\mathbf{x}^*$ for 100 corresponding input perturbations $\mathbf{u}$, the entries of which are drawn at random according to independent uniform distributions within $\pm 10\%$ of their nominal values $\mathbf{u}^*$. For every dataset, we estimate different lin-log and power-law models by solving unconstrained least squares problems (7) and (8) as well as linearly constrained versions of them based on various choices of the constraints (9)-(12). Namely, we consider lin-log models inferred without constraints (LL) and with sign constrains (LL$_S$), as well as power-law models inferred without constrains (PL), with sign-constraints (PL$_S$), with sign and parameter constraints (PL$_{SP}$), as well as with sign and reaction-rate constraints (PL$_{SR}$). Sign constrains are chosen in accordance with the choice of [9], reflecting the positive or negative effect of metabolites in the various reaction rates of the original model. For the constraints on the exponents of power-law models, inspired by [19], [18], we set $[\underline{p}_\ell, \overline{p}_\ell] = [-2, 5]$ for all parameters. Finally, loose rate constraints were placed over the region S given by $(x^*, u^*) \pm 10\%$ (all possible variations of internal and external metabolites up to 10% of their reference values). Writing the constraints in the form (13), for $h = 1, \ldots, m$, we took $\ln \underline{r}_h = -\infty$ (no lowerbound) and $\ln \overline{r}_h = 1.5 \cdot \ln(\overline{v}_h / v_h^*)$, where $\overline{v}_h$ is the hth-reaction maximal rate value of the real (simulated) system over S.

For each type of model (LL, LL$_S$, PL, PL$_S$, PL$_{SP}$ and PL$_{SR}$), from the 10 different estimation results (one per dataset) we compute mean and standard deviations of the approximation indices $L(X)$ and $\Theta(X)$, where X is the hyperrectangle within $\pm 10\%$ of $\mathbf{x}^*$. In practice these indices are computed by sampling X on a uniform grid with coordinates $x_i = \{x_i^* \pm 2\%,\ x_i^* \pm 6\%,\ x_i^* \pm 10\%\}$, $i = 1, \ldots, n_x$. These statistics are reported in Table I. In the same table, we also report analogous statistics on the normalized residuals of fit $K^{-1} \sum_k \| \left(\mathbf{v}^k - \hat{\mathbf{v}}(\mathbf{x}^k, \mathbf{u}^k, \mathbf{p})\right) / \mathbf{v}^* \|$, where rate predictions $\hat{\mathbf{v}}$ depend on the model inferred, and the normalized Mean Steady-State Error (MSSE) of the corresponding models, defined as $L^{-1} \sum_l \|(\mathbf{x}^l - \hat{\mathbf{x}}^l)/\mathbf{x}^l\|$, where $\mathbf{x}^l$ and $\hat{\mathbf{x}}^l$ are the steady-state response of the true system and the estimated model, respectively, to the lth of $L = 50$ input perturbations u drawn uniformly from $[0.5, 5] \times [0.05, 1.5] \times [0.1, 0.3]$. An example of the dynamic response to input perturbations of the models obtained from one of the 10 datasets is reported in Fig. 4.

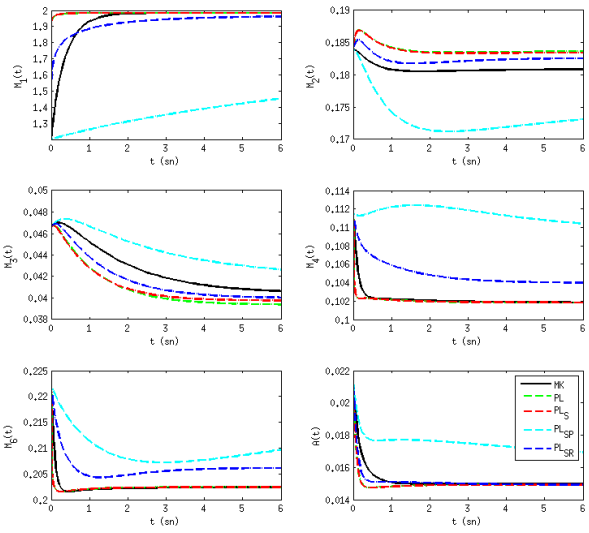

Fig. 4. Dynamic response to input $S = 2$ (other inputs unchanged) of the original system (black) and of the models PL$_S$, PL$_{SP}$ and PL$_{SR}$ inferred from one steady-state dataset (colors in the legend; fitting residuals are $4.76E-004$, 0.003 and 0.0026, in the same order).

The first observation concerning lin-log models is that using sign constraints generally improves the approximation obtained in all respects, although the improvement is limited and both models (LL and LL$_S$) provide rather accurate approximations of the vector field of the original system over the region considered. Of course, imposing constraints on the parameter signs can only worsen the fit of the data. On the other hand, the use of appropriate constraints returns a model that is biologically consistent, in the sense that parameter signs agree by construction with the system features. Interestingly, the constrained model also improves the approximation of the steady-state system response to input perturbations, as can be observed in the last column of the table.

Let us now turn to power-law models, whose structure appeared to be less suited for the case study under consideration. The steady-state performance (MSSE) of PL$_S$ is marginally better than that of PL, and both compare well with LL and LL$_S$ in this respect. Yet, the error in vector-field strength $L(X)$ is huge for both PL and PL$_S$. Further inspection of the results reveals that this is due to large values

TABLE I

STATISTICS OF THE MODELS ESTIMATED USING STEADY-STATE DATA FROM 10% PERTURBATION OF $\mathbf{u}^*$. TABLE ENTRIES REPORT MEAN±STANDARD DEVIATION OF THE QUANTITIES INDICATED.

Model	Residual	$L(X)$	$\Theta(X)$	MSSE
LL	7.43E-5±6.21E-6	0.0083±3.52E-3	0.9996±2.40E-4	0.0294±0.0009
LL_S	4.98E-4±3.25E-5	0.0053±2.54E-3	0.9998±5.17E-5	0.027±0.0011
PL	1.05E-4±1.46E-5	1.53E+04±1.09E+4	0.7224±8.13E-4	0.0328±0.0014
PL_S	4.94E-4±3.24E-5	6.35E+04±1.53E+3	0.7249±6.18E-4	0.0296±.0013
PL_{SP}	3.12E-3±1.52E-4	0.3190±4.13E-4	0.4697±3.14E-3	0.1036±.0018
PL_{SR}	2.71E-3±1.32E-4	0.2491±2.59E-4	0.8812±1.59E-3	0.0392±.0013

estimated for certain parameters in optimizing the data fit. This makes the reaction rate values and hence the vector-field diverge fast (with high exponents) when moving away from $(\mathbf{x}^*, \mathbf{u}^*)$. The excessively fast (and similar) dynamics of PL and PL_S are apparent from the example of Fig. 4. The alignment $\Theta(X)$ is similar for PL and PL_S and worse than LL and LL_S.

Imposing boundaries on individual parameter values is a potential remedy. Indeed, in PL_{SP}, the vector-field strength error $L(X)$ is improved overall. However, the alignment index $\Theta(X)$ gets worse. From the simulated example of Fig. 4, dynamics now appear too slow and in qualitative disagreement with the original system. The steady-state performance captured by the MSSE is also unsatisfactory. Clearly, a different choice of parameter boundaries could improve the situation. In particular, the largest fitting residual observed for PL_{SP} suggests that the problem is overconstrained. However, it is a priori unclear what should drive this choice. In fact, despite the kinetic order interpretation of the power-law exponent, the original system also depends on conservation laws that spoil this interpretation. In addition, phenomenological laws whose underlying reactions are not known in detail may be better captured by parameters that share little with biochemical kinetics, which makes the interpretation of model parameters and associated boundaries rather difficult.

Among power-law models, best approximation performance overall is obtained with PL_{SR}. Despite a data fitting error residual much larger than PL and PL_S and only slightly smaller than PL_{SP}, imposing rate constraints led to the smallest error $L(X)$ and the largest alignment values $\Theta(X)$, although the MSSE is slighly worse than for PL and PL_S. Relative to PL_{SP}, index $L(X)$ is similar, but the alignment $\Theta(X)$ is much improved. For one of the 10 datasets, a more detailed comparison of the vector field properties of the two approximate models is given in Fig. 5, in the form of histograms of $l(x)$ and $\theta(x)$ over X. Alignment $\theta(x)$ of PL_{SR} is clearly better concentrated near 1. Index $l(x)$ is concentrated on the positive semi-axis for both models, i.e. vector-field magnitude is smaller than for the original system, but that of PL_{SR} is generally closer to 0. This suggests overall slow dynamics for both, but slower dynamics for PL_{SP}, which is in substatial agreement with the example simulation of Fig. 4. Of course, different choices of reaction rate constraints may either improve or worsen performance.

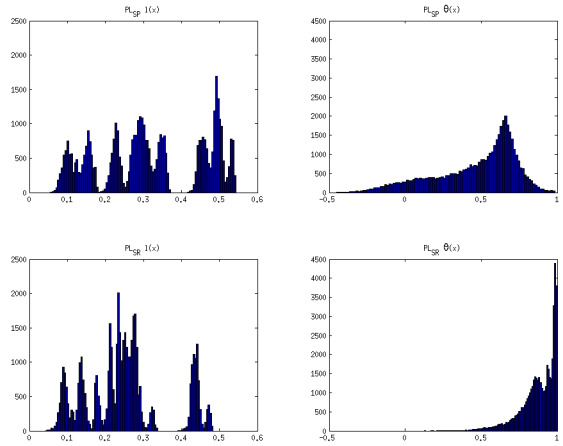

Fig. 5. Histograms of $l(x)$ (left) and $\theta(x)$ (right) over X for PL_{SP} (above) and PL_{SR} (below).

Similar to individual parameter constraints, the choice is not trivial. Still, it is intriguing to see how performance improves from PL_{SP} to PL_{SR} without much change in data fitting residuals, which suggests that the tightness of the constraints is in fact similar, but their shape in the parameter space is significantly different. Moreover, from a conceptual viewpoint, it should be noted that, contrary to individual parameter constraints, rate constraints are placed on expected properties of the system in dynamical conditions. The specific choice of model class only transfers these constraints to the parameters of the model in accordance with its structure.

As a final note, to test robustness of the models to perturbed conditions, we reported in Table II the existence of a valid (numerical) solution for different dynamic models inferred from one dataset subject to perturbations of inputs or of the system equilibrium. Invalid solutions (indicated with "-") are returned if the system would move to negative metabolite concentrations or the numerical solver is unable to provide a solution due e.g. to stiffness. From the table we observe that, contrary to common belief, power-law models are generally not well defined from a dynamic viewpoint for small concentrations, unless appropriate constraints are imposed at the model inference stage. A similar observation applies for the response of the models to large inputs. On

TABLE II

ROBUSTNESS OF DIFFERENT POWER-LAW MODELS INFERRED FROM ONE DATASET TO PERTURBED STATES (COLUMNS 2-8) OR INPUTS (COLUMNS 9-12). IN EACH COLUMN, SIMULATION IS PERFORMED IN REFERENCE CONDITIONS EXCEPT FOR THE INDICATED PERTURBATION. SUCCESSFUL SIMULATIONS ARE INDICATED WITH OK.

	Different values of initial condition $x(0)$							Different values of input S			
Model	$\mathbf{x}^*/10^5$	$\mathbf{x}^*/10^4$	$\mathbf{x}^*/10^3$	$\mathbf{x}^*/10^2$	$\mathbf{x}^*/10$	$0.2\mathbf{x}^*$	$1.3\mathbf{x}^*$	10	11	15	20
LL	-	-	-	OK	OK	OK	OK	OK	OK	OK	OK
LL_S	-	-	-	OK	OK	OK	OK	OK	OK	OK	OK
PL	-	-	-	-	OK	OK	OK	OK	OK	OK	-
PL_S	-	-	-	-	-	OK	OK	OK	-	-	-
PL_{SP}	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
PL_{SR}	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK

the other hand, lin-log models are invalid for concentrations limiting to zero, since reaction rates diverge by definition.

VI. CONCLUSIONS

In this paper we have addressed the problem of modelling metabolic dynamics from steady-state data. We showed that data fitting per se does not generally guarantee accurate dynamic performance. We have proposed metrics for a quantitative analysis of the dynamic model approximation, and discussed optimization constraints as a means to compensate for the limited informativity of the data and the discrepancy between the model structure of choice and the real system dynamics. Based on a simulated study on a simple yet informative system, we showed that the use of suitable constraints on the reaction rates driving the system dynamics allows one to improve the dynamical behavior of the models inferred from steady-state data. Although we focused on approximate kinetic models, several aspects of the work can be extended and generalized to various other model classes. In particular, we noted that robustness of the model away from the reference state is challenged by reaction rates involving powers of metabolite concentrations, which appear in many common model classes. For all these models, the risk of overfitting steady-state data by selecting large exponents, resulting in unsuitable dynamical behavior, can be avoided by applying the constraints discussed in this paper. In many cases, this will still result in tractable (linear, convex) optimization problems. On the other hand, the tools for quantifying the accuracy of dynamical model approximations apply equally well to any ODE model. Finally, generalization of reaction-rate constraints to other system properties, such as rates of change of metabolite concentrations, are plausible and deserve further investigation.

ACKNOWLEDGMENTS

The authors would like to thank H.de Jong for invaluable discussions and comments.

REFERENCES

[1] R. Heinrich and S. Schuster, *The Regulation of Cellular Systems*, Champman & Hall, NY; 1996.

[2] J.J. Heijnen, Approximate Kinetic Formats Used in Metabolic Network Modelling, *Biotechnol Bioeng.*, vol. 91 (5), 2005, pp 534-545.

[3] M.A. Savageau, *Biochemical System Analysis: A Study of Function and Design in Molecular Biology*, MA, Addison & Wesley, 1976.

[4] X. Delagado and J. Liao, Metabolic Control Analysis Using Transient Metabolite Concentrations, *Biochem. J.*, vol. 285,1992, pp 965-972.

[5] E.O. Voit et.al, Regulation of Glycolysis in Liactococcus Lactis: An Unfinished Systems Biological Case Study, *IEE Proc.-Syst. Biol.*, vol. 153 (4), 2006, pp 286-298.

[6] D. Visser and J. J. Heijnen, Dynamic Simulation and Metabolic Redesign of a Branched Pathway Using Lin-log Kinetics, *Metab Eng*, vol. 5, 2003, pp 164-176.

[7] R.C.H. del Rosario, E. Mendoza and E.O. Voit, Challenges in Lin-Log Modelling of Glycolysis in Lactococcus Lactis, *IET Syst. Biol.*, vol. 2 (3), pp 136-149.

[8] J. Nielsen, Metabolic Control Analysis of Biochemical Pathways Based on Thermokinetic Description of Reaction Rates, *Biochem. J.*, vol. 321, 1997, pp 133-138.

[9] S. Berthoumieux, M. Brilli, D. Kahn, H. de Jong and E. Cinquemani, On the identifiability of Metabolic Network Models, *J. Math. Biol.*, 2012. in press.

[10] S. Berthoumieux, M. Brilli, H. de Jong, D. Kahn and E. Cinquemani, Identification of metabolic network models from incomplete high-throughput datasets, *Bioinformatics*, vol. 27 (13), 2011, pp i186-i195.

[11] P. Mendes and D.B. Kell, Non-linear Optimization of Biochemical Pathways: Applications to Metabolic Engineering and Parameter Estimation, *Bioinformatics*, vol. 14, 1998, pp 869-883.

[12] M.A. Savageau, Michaelis-Menten Mechanism Reconsidered: Implications of Fractal Kinetics, *J. Theor. Biol.*, vol. 176, 1995, pp 115-124.

[13] K. Smallbone, E. Simeonidis, N. Swainston and Pedro Mendes, Towards a genome-scale kinetic model of cellular metabolism, *BMC Systems Biology*, 2010, vol. 4 (6).

[14] F. Hadlich, S. Noack and W. Wiechert, Translating biochemical network models between different kinetic formats, *Metabolic Engineering*, vol. 11, 2009, pp 87-100.

[15] D. Visser, J.W. Schmid, K. Mauch, M. Reuss, and J.J. Heijnen, Optimal re-design of primary metabolism in Escherichia coli using linlog kinetics, *Metabolic Engineering*, vol. 6 (4), 2004, pp 378-390.

[16] R.S. Costa, D. Machado, I. Rocha, and E.C. Ferreira, Hybrid dynamic modeling of *Escherichia coli* central metabolic network combining Michaelis-Menten and approximate kinetic equations, *Biosystems*, vol. 100 (2), 2010, pp 150-8.

[17] N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P.Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori and M. Tomita, Multiple high-throughput analyses monitor the response of *E. coli* to perturbations, *Science*, vol. 316(5824), 2007, pp 593–597.

[18] C.-L. Ko and E. Voit and F.-S.Wang, Estimating parameters for generalized mass action models with connectivity information, *BMC Bioinformatics*, vol. 10 (1), 2009, pp 140.

[19] G. Jia, G. Stephanopoulos and R. Gunawan, Incremental parameter estimation of kinetic metabolic network models, *BMC Systems Biology*, vol. 6 (1), 2012, pp 142.

[20] C. Chassagnole, N. Noisommit-Rizzi, J.-W. Schmid, K. Mauch and M. Reuss, Dynamic modeling of the central carbon metabolism of *Escherichia coli*, *Biotechnol Bioeng.*, vol. 79 (1), 2002, pp 53-73.