Practical Identification and Experimental Design for Parameter Estimation in Kinetic Models of Metabolism

Shyam Srinivasan[a], William R. Cluett[a] and Radhakrishnan Mahadevan[*,a,b]

a - Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON, Canada.

b - Institute for Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON, Canada.

* Corresponding author

Abstract

1 Introduction:

The use of metabolic engineering spans a wide variety of applications. Some notable examples include the design of microorganisms for the biosynthesis of commodity and specialty chemicals (Andreozzi, Chakrabarti, ET AL. 2016), engineering mammalian cells as therapeutic targets for cures to some ailments affecting humans (Di Filippo, ET AL. 2016; Apaolaza, ET AL. 2017), and changing the constituents of the human gut microbial community to cure related diseases (Zerfaß, Chen, AND Soyer 2018). These applications require us to understand the numerous complex interactions, their roles in cell function, and sometimes even the mechanisms behind these interactions. Computational models offer a systematic way to integrate available experimental data, and to study and understand these interactions through mathematical representations of the biological systems in which these interactions occur (Bordbar, Monk, ET AL. 2014; Saa AND Nielsen 2017). They are also used to predict changes in cell function based on changes in the type and nature of the modeled interactions (Andreozzi, Chakrabarti, ET AL. 2016), or aid in the identification of therapeutic targets for drug discovery and development (Bordbar, McCloskey, ET AL. 2015; Chandrasekaran, ET AL. 2017)

Constraint-based models (CBMs) of metabolism are used to improve our understanding of metabolism by representing it as a stoichiometric network of reactions ~~that operate under a pseudo-steady state assumption~~ (Bordbar, Monk, ET AL. 2014). The ability of CBMs to shine light on the nonintuitive interactions that

govern cellular metabolism is leveraged to engineer and asses the impact of designs that alter the ability of a cell to grow, or produce a desired metabolite (Maia, M. Rocha, AND I. Rocha 2016). However, in CBMs, metabolism is assumed to operate under a pseudo steady state. Consequently, the metabolite concentrations within the metabolic network are assumed to be constant, and changes in metabolite concentrations are not modeled. Furthermore, since CBMs represent metabolism using only the stoichiometry of its constituent reactions, they do not account for the various non-catalytic regulatory interactions that are also responsible for metabolic function. These shortcomings prevent CBMs from being used to fully understand the steady state as well as the dynamic characteristics of metabolic networks.

In contrast, the effects of regulatory interactions and changes in metabolite concentrations on different characteristics of metabolism can be studied using kinetic models of metabolism (Saa AND Nielsen 2017). These models account for changes in metabolite concentrations subject to thermodynamic and regulatory constraints that underly metabolic networks in addition to its stoichiometry (Link, Christodoulou, AND Sauer 2014). Kinetic models can not only help us better understand lesser known and understood characteristics of metabolism like bistability (Kotte, ET AL. 2014), and their role in human health, but can also improve predictions about the impact of engineering design perturbations on metabolism, and help propose alternative designs to achieve metabolite production goals (Khodayari, ET AL. 2016).

Kinetic models differ from CBMs in their use of mechanistic enzyme kinetic rate laws to model enzyme catalyzed fluxes within a metabolic network (Srinivasan, Cluett, AND Mahadevan 2015; Saa AND Nielsen 2017). The use of kinetic models requires information on the enzyme kinetic rate laws that will be used to model the fluxes within a metabolic network, as well as numerical values for the parameters used in these rate laws. Analyzing the ability of a metabolic network to exhibit dynamic characteristics like multiple steady states and oscillations, irrespective of the structure of the network, is one example where kinetic rate laws and parameter values play a crucial role (Srinivasan, Cluett, AND Mahadevan 2017).

Despite their importance, the parameterization of kinetic models is still a problem for which solutions are a subject of debate within the modeling community. Typically, enzyme kinetic rate laws are parameterized based on in vitro observations of enzyme activity, as opposed to observations made under in vivo conditions (Heijnen 2005; Smallbone, ET AL. 2007). However, some researchers have questioned their rele-

vance for gleaning information on the dynamics of metabolism under in vivo conditions as opposed to in vitro conditions (Heijnen 2005; Heijnen AND Verheijen 2013). On the other hand, some reports have shown that despite the large uncertainties associated with parameters estimated based on in vivo experimental data (Link, Christodoulou, AND Sauer 2014), in vitro parameter estimates are a reasonable approximation of values that would be applicable under in vivo conditions (Ron Milo paper comparing in vitro vs invivo enzyme turn over rates in PLoS Computational Biology).

Authors have sought to constrain and determine uncertainty in parameter estimates and consequent model predictions by using Monte Carlo approaches for kinetic modeling of metabolism (Andreozzi, Miskovic, AND Hatzimanikatis 2016). These approaches also allow for the integration of experimentally observed in vivo data. ORACLE (Wang and Haztzimanikatis, 2004) and Ensemble modeling (Tan and Liao, 2008), are two such examples. These and other Monte Carlo kinetic modeling methods have been previously reviewed (Srinivasan, Cluett, AND Mahadevan 2015). Bayesian approaches to improve parameter estimation and quantify estimation uncertainty have also been proposed (Saa AND Nielsen 2016). Vanlier, C. Tiemann, ET AL. 2013 provide a review of different Bayesian approaches to quantify parameter uncertainty.

Related to the development of these methods to quantify parameter estimation uncertainty, the importance of model parameter identifiability, i.e., the condition that needs to be satisfied to estimate unique kinetic parameter values from experimental data, is often overlooked (Ljung AND Glad 1994; Berthoumieux, ET AL. 2013). In a model, a parameter is said to be structurally a priori identifiable if its values can be uniquely estimated from available experimental data, independent of all other model parameters. However, if a parameter cannot be uniquely estimated due to redundant model parameterization, or due to the nonlinear relationships between the model parameters, then the parameter is said to be structurally non-identifiable. Conversely, if the ability to estimate unique parameter values is compromised due to the inability of the available data to capture the requisite information needed to estimate the parameters in the modeled system, and the uncertainty in parameter estimates is unquantifiable, the parameter is said to be practically non-identifiable (Ljung AND Glad 1994).

Authors have proposed ways to assess parameter identifiability by proposing approximate kinetic models of metabolism that utilize empirical enzyme kinetic rate laws with parameters that have physical significance,

and are identifiable (Heijnen 2005; Smallbone, ET AL. 2007). Significant work has also been done towards the development of methods for structural identification of parameters in kinetic models of metabolism (Ljung AND Glad 1994; Nikerel, ET AL. 2009; Berthoumieux, ET AL. 2013; Raue, ET AL. 2014)(paper from Rudiyanto Gunawan on model discrimination and sensitivity analysis).

Methods to improve practical identifiability through a priori experimental design have also been developed, with a focus on kinetic models of metabolism (Gadkar, Gunawan, AND Doyle 2005; Vanlier, C. a. Tiemann, ET AL. 2014; Raue, ET AL. 2014). Some of these methods are limited by their applicability to approximate kinetic models only (Nikerel, ET AL. 2009; Berthoumieux, ET AL. 2013), while some of them suffer from computational limitations when applied to kinetic models of large metabolic networks (Gadkar, Gunawan, AND Doyle 2005; Raue, ET AL. 2014)(Banga method using FIM for D-optimal design, ??).

In this paper, we propose a scalable methodology that uses available steady state fluxomics, metabolomics and proteomics data to test the practical identifiability of parameters for each individual reaction in kinetic models of metabolism. We demonstrate how the computer algebra-based method that we have developed to perform this test can also facilitate the design of experiments for generating the data required to estimate unique parameter values for all reaction fluxes in a metabolic network. In doing so, we are able to propose the number and types of perturbations that will provide the most useful data for parameter estimation, as well as test the identifiability of different enzyme kinetic rate laws that are typically used to model fluxes in metabolic networks. We illustrate our methodology using a small metabolic network model of glucoeneogenesis in *Escherichia coli* (Kotte, ET AL. 2014; Srinivasan, Cluett, AND Mahadevan 2017) under the assumption that all intracellular metabolite concentrations and fluxes can be measured.

2 Methods

In Section 2.1 we present the typical structure of a kinetic model of metabolism and a preliminary description of the least squares method by which the model parameters are estimated from experimental data. In Section 2.2 we provide formal mathematical definitions for structural and practical identifiability. The computer-algebra system based method to assess identifiability that we have developed is described in Section 2.3 ~~that follows~~. We also provide a description of a numerical equivalent for the computer algebra-based method to

estimate parameters numerically in Section 2.4. In Section 2.5, we define a quantitative metric to describe the identifiability of parameters, and in Section 2.6 a complete description of the small metabolic network used to demonstrate the methodology that we have developed is provided. Finally, a description of how the method can be used for experimental design for parameter estimation is given in Section 2.7.

2.1 Parameter estimation for kinetic models of metabolism

In kinetic models of metabolism, ordinary differential equations (ODE) are used to express the rate of change of metabolite concentrations ($x \in \mathbb{R}^{n_x}$) as a function of the reaction fluxes ($v \in \mathbb{R}^{n_r}$) in the metabolic network (Equation 1). The matrix $\mathbf{S} \in \mathbb{R}^{n_x \mathrm{x} n_r}$ in Equation (1a) defines the stoichiometric relationship between the fluxes and the concentrations of the metabolic network.

$$\dot{x} = \mathbf{S}v \tag{1a}$$

$$v = f(x, \theta, u) \tag{1b}$$

The expression for the nonlinear function (f) used to describe each reaction flux v_i in v, $i = 1, 2, ..., n_r$, in a particular kinetic model (Equation 1b) is dependent on the enzyme kinetic mechanism that is used to model the reaction (Srinivasan, Cluett, AND Mahadevan 2015). Accordingly, f is typically a nonlinear function of the vector of metabolite concentrations ($x \in \mathbb{R}^{n_x}$), the vector of enzyme kinetic parameters ($\theta \in \mathbb{R}^{n_p}$) and other input concentrations ($u \in \mathbb{R}^{n_u}$).

Parameter estimation methods based on optimization principles are used to determine ~~the~~ parameter ~~values~~ estimates from experimental data. Under the assumption that all intracellular metabolite concentrations and fluxes can be measured, a parameter estimation problem can be formulated as a nonlinear programming problem (Equation 2) to estimate the values of enzyme kinetic parameters, $\theta \in \mathbb{R}^{n_p}$, based on the measured data.

$$\min_{\theta} \sum_{k=1}^{n_x+n_r} \sum_{l=1}^{n_t} \left(\frac{y^*_{kl} - y_{kl}}{\sigma^*_{kl}} \right)^2 \tag{2a}$$

$$\theta_l \leq \theta \leq \theta_u \tag{2b}$$

Here $y \in \mathbb{R}^{n_x+n_r}$ is the combined vector of concentrations ($x \in \mathbb{R}^{n_x}$) and fluxes ($v \in \mathbb{R}^{n_r}$), at each time point $l = 1, 2, ..., n_t$. The minimization of the least squares error between the measured (y^*) and modeled (y) concentrations and fluxes, weighted by the variance in the experimental data σ^*_{kl} for each concentration and flux, at each time point $l = 1, 2, ..., n_t$, is used as an objective function (Equation 2a) for the optimization problem. The least squares parameter values are determined within fixed upper (θ_u) and lower (θ_l) bounds (Equation 2b).

2.2 Structural and practical identifiability of parameters in kinetic models

In the Introduction, we briefly mentioned that the ability to estimate unique parameter values from available experimental data is governed by the identifiability of these parameters in the model (Ljung AND Glad 1994; Vanlier, C. A. Tiemann, ET AL. 2012; Berthoumieux, ET AL. 2013; Raue, ET AL. 2014). Below, we provide a formal definition of structural and practical identifiability of parameters.

The parameters in θ in any nonlinear model (Equation 1) are said to be structurally identifiable if, for an input-output mapping defined by $y = \Phi(\theta, u)$ for at least one input function u, any two values of parameters θ_1 and θ_2 satisfy the relationship in Equation (3):

$$\Phi(\theta_1, u) = \Phi(\theta_2, u) \iff \theta_1 = \theta_2 \tag{3}$$

Accordingly, if parameters in θ have a unique value, a finite number of non-unique values or an infinite number of values for all input functions, they are said to be structurally globally identifiable, locally identifiable or non-identifiable, respectively. So, the structural identifiability of parameters in a dynamic model helps establish the presence or absence of a relationship between the unmeasured and measured concentrations/fluxes, as well as correlations between different model parameters (Rudiyanto Gunawan paper on model discrimination). Consequently, the effect of model structure and parameterization on the ability to infer true parameter values from experimental data is determined by the structural identifiability of the parameter.

Experimental data from many physical systems is usually noisy, and when parameters are estimated on the basis of noisy data, the ability to estimate unique parameter values to satisfy Equation (3) is referred

to as practical identifiability. If a single unique parameter satisfying Equation (3) can be found, then θ is said to be globally practically identifiable. Whereas, if parameter estimates with quantifiable uncertainties can be found, then the θ is said to be locally identifiable. The absence of unique parameter estimates for θ leads to practical non-identifiability. The practical identifiability of a parameter is hence contingent upon the nature, quality and quantity of data available to estimate the parameter as opposed to the structure and parameterization of the model.

So, on the one hand, establishing the structural identifiability of parameters enables one to propose models that are not only appropriate representations of physical processes, but are also parameterized in such a way that the value of these parameters can be estimated from measurable data. On the other hand, establishing practical identifiability of parameters in any model helps design experiments that are minimal, informative and useful for parameter estimation.

2.3 A method to determine identifiability of kinetic models of metabolism

We provide the mathematical framework for identification of parameters in kinetic models of metabolism in this section. A summary of the methodology in the form of a flow diagram is shown in Figure 1. As indicated in Figure 1a, the first step involves the construction of the kinetic model (Equation 1) of the metabolic network with n_r reaction fluxes.

For each flux v_i, $i = 1, 2, ..., n_r$, in the kinetic model, let $\theta \in \mathbb{R}^p$ in Equation (1b). If data from n_E experiments is available for the chosen metabolic network, as stated earlier, for each experiment $j = 1, 2, ..., n_E$, we assume that all metabolite concentrations ($x \in \mathbb{R}^{n_x}$) and reaction fluxes ($v \in \mathbb{R}^{n_r}$) are measurable. We discuss the implications of relaxing this assumption in the results section. The pertinent information for each experiment j is available as a vector of concentrations and fluxes, $\mathbf{x}_j$ and $\mathbf{v}_j$, respectively (Figure 1b).

In order to establish the practical identifiability of kinetic parameters for each flux v_i, $i = 1, 2, ..., n_r$, we describe a computer algebra-based method. The primary use of the computer algebra system is to obtain closed-form expressions for each parameter in $\theta \in \mathbb{R}^p$ for each flux v_i (Figure 1b). This is done by first selecting a combination of $p \leq n_E$ experimental data. The fluxes and concentrations from p different

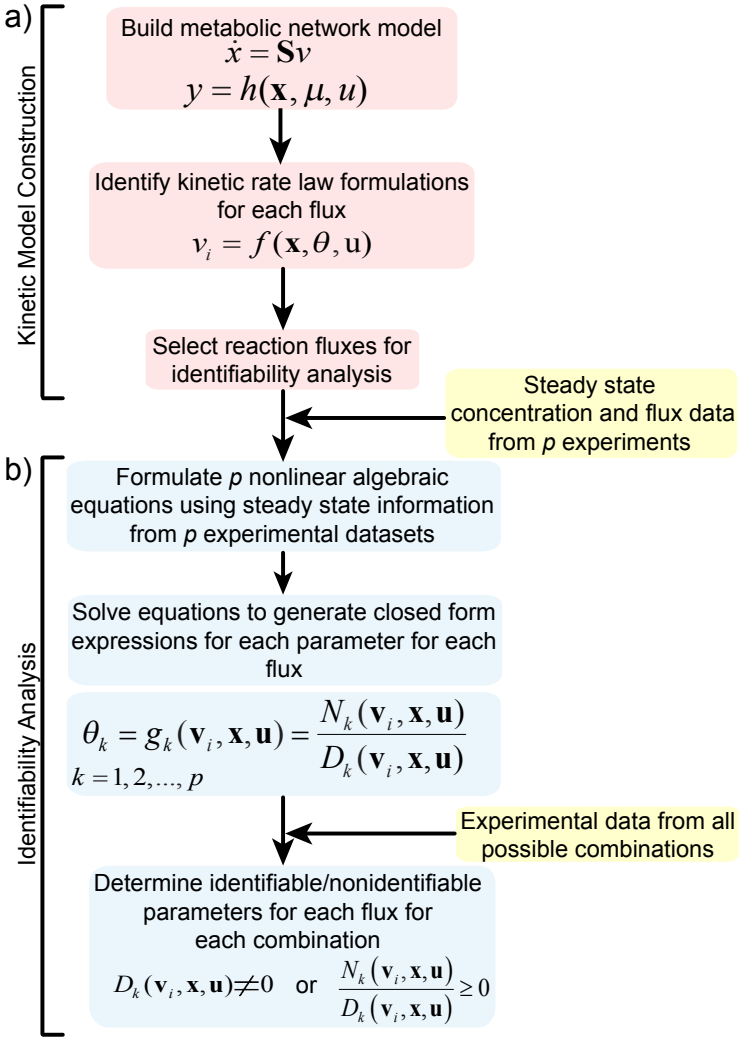

Figure 1. A flow diagram showing the methodology developed to establish practical identifiability of parameters in kinetic models of metabolism. a) The steps for the construction of a kinetic model of a metabolic network. The choice of rate law formulations to describe metabolic fluxes influences the identification methodology. The identifiability of parameters for each flux can be established independently. b) The steps for practical identifiability analysis for parameters of a single flux.

experiments are then used to formulate a system of nonlinear algebraic equations in $\mathbb{R}^p$ for each flux v_i, shown in Equation (4).

$$v_{i,j} = f_j(\mathbf{x}_j, \theta, \mathbf{u}_j) \qquad \forall j = \{1, 2, ..., p\} \subset \{1, 2, ..., n_E\} \tag{4}$$

Here, $v_{i,j}$ refers to the flux v_i obtained from experiment j. $\mathbf{x}_j$ and $\mathbf{u}_j$ are the vector of metabolite and other input concentrations from each experiment j, and θ is a vector in $\mathbb{R}^p$, whose elements are denoted by θ_k.

Each equation in (4), indicated by the index j, corresponds to the kinetic rate law expression $f(x, \theta, u)$ for each v_i, $i = 1, 2, ..., n_r$, described in Equation (1b), written for concentrations ($\mathbf{x}_j$, $\mathbf{u}_j$) and fluxes ($v_{i,j}$) obtained from experiment j. Solving the system in Equation (4) results in $\mathbb{R}^p$ nonlinear expressions for each parameter θ_k in $\theta \in \mathbb{R}^p$ (Equation 5), where $N(\mathbf{v}_i, \mathbf{x}, \mathbf{u})$ is the numerator of g, and $D(\mathbf{v}_i, \mathbf{x}, \mathbf{u})$ is the denominator of g (Figure 1b). Note that $\mathbf{v}_i$, $\mathbf{x}$ and $\mathbf{u}$ are used to denote vector of vectors of fluxes for reaction i ($\mathbf{v}_i$), metabolite ($\mathbf{x}$) and input ($\mathbf{u}$) concentrations, respectively, obtained from p experiments denoted by the index $j = 1, 2, ..., p$.

$$\theta_k = g_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u}) = \frac{N_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u})}{D_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u})} \tag{5}$$

The identifiability of parameter θ_k, $k = 1, 2, ..., p$, for flux v_i can be established by determining the value of $D_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u})$ (Figure 1b): any parameter θ_k is said to practically identifiable if $D_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u}) \neq 0$, and practically non-identifiable if $D_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u}) = 0$. Furthermore, the physical properties of the kinetic parameters can be used to distinguish between identifiable and non-identifiable parameter values by designating only parameters with a positive value of $g_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u})$ as identifiable (Figure 1b). The solution $g_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u})$ in Equation (5) is unique for an identifiable θ_k, and an infinite number of solutions are possible for a non-identifiable θ_k. However, if there are multiple but finite solutions $g_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u})$, then the corresponding parameter θ_k is locally identifiable.

2.4 Numerical parameter identification and estimation

The identification methodology described in the previous section can be extended to be used in conjunction with an optimization-based procedure for numerical estimation of parameters from experimental data. The $\mathbb{R}^p$ nonlinear algebraic equations formulated for each flux v_i (Equation 4) can be written as shown in Equation (6) below to be solved as an optimization problem. Minimization of the sum of the absolute error values is used as the objective, and the $\mathbb{R}^p$ nonlinear algebraic equations are provided as inequality constraints.

$$\min_{\theta,\varepsilon} \sum_{j=1}^{p} |\varepsilon_j| \tag{6a}$$

$$v_{i,j} - f_j(\mathbf{x}_j, \theta, \mathbf{u}_j) \leq \varepsilon_j \qquad j = \{1, 2, ..., p\} \tag{6b}$$

$$\theta_l \leq \theta_j \leq \theta_u \qquad j = \{1, 2, ..., p\} \tag{6c}$$

The formulation of the above optimization problem is similar to the least squares problem shown earlier in Equation (2). Notice that the parameter values are bounded above and below by θ_u and θ_l, respectively. These bounds can be specified to be arbitrarily large to include all possible parameter values.

2.5 Degree of identifiability: A quantitative measure of practical identifiability

We express the practical identifiability of kinetic parameters using a simple quantitative term called the degree of identifiability. We describe the degree of identifiability of any single parameter as the percentage of all data combinations (used to test for practical identifiability) that can identify that parameter.

As an example, if 90% of all the experimental data combinations used for testing can identify a parameter θ_i, then the degree of identifiability of θ_i is said to be 0.9 or 90%. On the other hand, if only 10% of the combinations can identify another parameter θ_j, then θ_j has a degree of identifiability of 0.1 or 10%. Furthermore, we can create a hierarchy of practically identifiable parameters using their degrees of identifiability. In the above instance of the two parameters θ_i and θ_j that have degrees of identifiability of 90% and 10% respectively, θ_i is classified to be more identifiable than θ_j due to its relatively higher degree of identifiability.

Determining this hierarchy of identifiable parameters can help in distinguishing parameters that can be

identified by any type and any combination of experiments from parameters that can be identified by only a select type and combination of experiments. Such a classification can subsequently be used to design minimal sets of experiments that can practically identify all kinetic parameters used to model a metabolic network, going from the least identifiable parameter to the most identifiable parameter.

2.6 Kinetic model of gluconeogenesis in *E. coli*

A previously proposed kinetic model (Kotte, ET AL. 2014; Srinivasan, Cluett, AND Mahadevan 2017) for acetate consumption through gluconeogenesis (Figure 2) is used as a case study to illustrate identifiability analysis for experimental design for parameter estimation in kinetic models of metabolism. The kinetic model is described below.

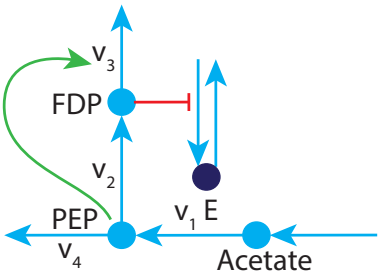

Figure 2. The previously published small metabolic network for gluconeogenesis used to demonstrate our practical identifiability method for kinetic models of metabolism.

$$\frac{d}{dt}pep = v_1 - v_2 - v_4 \tag{7}$$

$$\frac{d}{dt}fdp = v_2 - v_3 \tag{8}$$

$$\frac{d}{dt}E = v_5 - dE \tag{9}$$

The kinetic expressions for fluxes v_1 through v_5 are given below. The consumption of acetate through v_1 and conversion of *pep* through v_2 are expressed in Equations (10) and (12) respectively using Michaelis-Menten

kinetics. The acetate flux through v_1 is also governed by the quantity of available enzyme E.

$$v_1 = k_1^{cat} E \frac{ac}{ac + K_1^{ac}} \tag{10}$$

The model for flux v_1 of the small network (Figure 2), uses the concentration of the enzyme E as a variable (Equation 10). Since we assume that steady state experimental information is only available for metabolite concentrations and fluxes, and not for enzymes (again the details on relaxing this assumption are discussed later), the expression in Equation (10) for v_1 cannot be used for identifying parameters k_1^{cat} and K_1^{ac}. So, we modify the Michaelis-Menten kinetic rate law expression to eliminate the enzyme concentration E as a variable in Equation (11). Consequently k_1^{cat} is replaced by V_1^{max} as a parameter to describe v_1. The corresponding enzyme binding constant is denoted as $K_1^{ac}(ne)$ to distinguish it from the enzyme binding constant calculated in the presence of measured enzyme concentration data.

$$v_1 = V_1^{max} \frac{ac}{ac + K_1^{ac}(ne)} \tag{11}$$

We choose the expression for flux v_1 given in Equation (11) to demonstrate our method for practical identifiability.

$$v_2 = V_2^{max} \frac{pep}{pep + K_2^{pep}} \tag{12}$$

$$v_3 = V_3^{max} \frac{f\tilde{d}p \left(1 + f\tilde{d}p\right)^3}{\left(1 + f\tilde{d}p\right)^4 + L_3 \left(1 + \frac{pep}{K_3^{pep}}\right)^{-4}} \tag{13}$$

The allosterically regulated flux v_3 for the consumption of *fdp* is expressed in Equation (13) using the Monod-Wyman-Changeux (MWC) model for allosterically regulated enzymes, where $f\tilde{d}p$ refers to the ratio of *fdp* with respect to its allosteric binding constant K_3^{fdp}.

The practically identifiability of parameters of a given flux are determined by solving a system of nonlinear algebraic equations using a computer algebra system (Section 2.3). We find that the nonlinearity of the MWC kinetic rate law used to model the allosteric regulation of v_3 makes it computationally intractable for determining the closed form expressions of the three parameters V_3^{max}, K_3^{fdp} and K_3^{pep} using a computer algebra system (Mathematica or SymPy in Python). In order to overcome this computational obstacle, we

model the reaction rate for v_3 using the convenience kinetic rate law formulation (Liebermeister AND Klipp 2006). The corresponding expression obtained for v_3 is given below (Equation 14).

$$v_3 = V_3^{max} \left(\frac{1}{1 + \frac{K_3^{pep}}{pep}} \right) \left(\frac{\frac{fdp}{K_3^{fdp}}}{1 + \frac{fdp}{K_3^{fdp}}} \right) \tag{14}$$

The flux v_4 for the export of *pep* is expressed as a linear equation dependent on *pep* in Equation (15).

$$v_4 = V_4^{max}.pep \tag{15}$$

The production of enzyme E is represented by flux v_5. The inhibition of this flux by *fdp* is modeled using Hill kinetics, where K_e^{fdp} represents the Hill binding constant for the inhibiting metabolite *fdp*, n_e is the Hill exponent, and V_e^{max} is the maximum reaction rate for v_5.

$$v_5 = V_e^{max} \left(\frac{1}{1 + \left(\frac{fdp}{K_e^{fdp}} \right)^{n_e}} \right) \tag{16}$$

2.7 Experimental design through practical parameter identification

Not all metabolite concentrations and fluxes in the model (Equation 1) change for any random experiment. This makes unambiguous estimation of parameters impossible, either due to the inherent correlation between changes in different concentrations or fluxes, or due to the homeostasis of the concentrations and fluxes under the chosen experimental conditions (Heijnen AND Verheijen 2013). In such scenarios, the need to design experiments to effect a change in, and discriminate between changes in different concentrations/fluxes becomes necessary.

Following the methodology described in Section 2.3, and demonstrated in Section 3.1 for a single flux using data from a combination of two different experiments, all distinct combinations of data sets obtained from experiments described in Section S3.1 of the Supplementary Information can be tested for their ability to practically identify any of the fluxes in the small metabolic network. This step would determine the degree of identifiability (defined in Section 2.5) of each parameter in each flux in the model, and help distinguish experiment combinations that contribute to identifiability from combinations that do not practically identify

any parameter in the model (Figure 1b). In doing so, it is possible to obtain a minimal and informative collection of experiments that can be performed to identify as many model parameters as possible (Figure S5). Consequently, the set of experiments can be used to estimate all the identifiable parameters in the model. This is formally explained below.

The identifiability of each parameter based on each experiment with index $j = 1, 2, ..., n_E$ is established based on the methodology described in Section 2.3 (Figure 1b), and demonstrated in Section 3.1. Subsequently, for any flux v_i, and for any combination of p experimental data sets, if the experimental concentrations and fluxes ($\mathbf{x}_j$ and $\mathbf{v}_j$, respectively, where $j = 1, 2, ..., p$) do not satisfy the condition for identifiability for any parameter θ_k in $\theta \in \mathbb{R}^p$ (Figure 1b), then at least one of the p experiments needs to be changed to make parameter θ_k identifiable. Consequently, the corresponding experiment cannot be used for estimating parameter θ_k, and needs to be discarded from the set of all necessary experiments. Furthermore, another experiment from $j = 1, ..., n_E$ needs to be selected to replace the discarded experiment such that parameter θ_k is identifiable. This process has to be repeated until all parameters in $\theta \in \mathbb{R}^p$ are identifiable for flux v_i. In doing so, we can arrive at a set of p experiments that will always result in practically identifiable parameters for flux v_i. Note that if none of the n_E pre-selected experiments satisfy the identifiability condition, then we can design an $(n_E + 1)^{th}$ experiment that can replace one of the experiments that causes practical non-identifiability. This analysis can be performed for each flux in a metabolic network independent of all the other fluxes, making it theoretically scalable even to genome-scale models of metabolism.

3 Results

First, in Section 3.1, we demonstrate the use of the methodology that we described in Section 2.1 to practically identify parameters in flux v_1 of the small gluconeogenic network (Figure 2) model given in Section 2.6. We discuss the ability of the proposed methodology to determine the structural identifiability of parameters modeling v_1, v_3 and v_5 in Section 3.2. In Section 3.3 that follows, we show how the demonstrated methodology is capable of practically identifying and estimating parameters for fluxes v_1, v_2, v_3 and v_5 using steady state flux values and metabolite concentrations. The various ways in which this information can be used for designing experiments to generate data that can facilitate estimation of identifiable parameters are discussed

in Section 3.4. The contribution of the uncertainty in the data arising from either the differences between in vivo and in vitro kinetics, or the noise present in experimentally measured quantities towards identifying parameters in enzyme kinetic models is discussed finally in Section 3.5.

3.1 Identifying parameters in kinetic models of metabolism: an example

In this section, we illustrate the proposed methodology step by step to identify parameters of flux v_1 in the small metabolic network (Figure 2 and Section 2.6). We choose the expression for flux v_1 given in Equation (11) for this demonstration.

Since $\theta = \{V_1^{max}, K_1^{ac}(ne)\} \in \mathbb{R}^2$ for v_1, as mentioned in Section S3.1, we need steady state concentration and flux measurements from at least two different experiments. So, from the $n_E = 21$ different experiments described in Section S3.1 and Table S1, we can choose multiple combinations of $p = 2$ experiments to satisfy the data requirements for identifying v_1 i.e., in Equation (4) $j = \{1, 2\}$. We label the available concentrations and fluxes as $ac^{(j)}$ and ${v_1}^{(j)}$, respectively. Then, the nonlinear algebraic equations shown in Equation (4) can be formulated for v_1 as:

$$ {v_1}^{(j)} = V_1^{max} \frac{ac^{(j)}}{ac^{(j)} + K_1^{ac}(ne)} \qquad\qquad j = \{1, 2\} $$

Solving this simultaneous system of equations in $\mathbb{R}^2$ using Mathematica (Wolfram Research, USA), a computer algebra system, we get $p = 2$ nonlinear algebraic expressions for parameters V_1^{max} (Equation 17a) and $K_1^{ac}(ne)$ (Equation 17b). These expressions have the form shown in Equation (5).

$$ \theta_1 = V_1^{max} = \frac{v_1^{(1)} v_1^{(2)} (ac^{(1)} - ac^{(2)})}{v_1^{(2)} ac^{(1)} - v_1^{(1)} ac^{(2)}} \tag{17a} $$

$$ \theta_2 = K_1^{ac}(ne) = \frac{ac^{(1)} ac^{(2)} (v_1^{(1)} - v_1^{(2)})}{v_1^{(2)} ac^{(1)} - v_1^{(1)} ac^{(2)}} \tag{17b} $$

To test the practical identifiability of the parameters in Equation 17, we substitute any suitable in silico experimental data and determine the value of the denominator of the right hand side expression. Since the enzyme binding constant ($K_1^{ac}(ne)$) and the maximum reaction rate (V_1^{max}) cannot be negative, we

can further constrain the criteria for identifiability for both these parameters by saying that the evaluated expressions in Equation (17) should be positive (Figure 1b). The parameter values that are obtained for V_1^{max} and $K_1^{ac}(ne)$ by substituting in silico steady state experimental data are shown in Supplementary Figure S1. Due to the numerous possible parameter values seen in Supplementary Figure S1, we can conclude that both V_1^{max} and K_1^{ac} are practically non-identifiable.

We can also apply the proposed methodology to practically identify parameters in v_1 under the assumption that the protein concentration for the enzyme E is also available, in addition to the measured metabolite concentrations and fluxes. In doing so, we get two expressions similar to the one shown in Equation (17) for k_1^{cat} and K_1^{ac}. Here, the value of V_1^{max} in Equation (11) is substituted with $V_1^{max} = k_1^{cat}E$ instead. The corresponding identifiability expressions for k_1^{cat} and K_1^{ac} are given in Equation (18).

$$k_1^{cat} = \frac{v_1^{(1)} v_1^{(2)} \left(ac^{(1)} - ac^{(2)}\right)}{v_1^{(2)} ac^{(1)} E^{(1)} - v_1^{(1)} ac^{(2)} E^{(2)}} \tag{18a}$$

$$K_1^{ac} = \frac{ac^{(1)} ac^{(2)} \left(v_1^{(1)} E^{(2)} - v_1^{(2)} E^{(1)}\right)}{v_1^{(2)} ac^{(1)} E^{(1)} - v_1^{(1)} ac^{(2)} E^{(2)}} \tag{18b}$$

We show the parameter value for k_1^{cat} and K_1^{ac} that are obtained through the practical identifiability analysis in Figure 3a when in silico experimental data is substituted in Equation (18). Through Equation (18) and Figure 3a we are able to show that the uncertainty in the parameter estimates (Supplementary Figure S1) can be resolved through the incorporation of the available enzyme concentrations. Thus, having more experimental information can help resolve practical identifiability.

In the following section we present results from the identifiability analysis of fluxes v_2, v_3 and v_5 in the small metabolic network (Figure 2), using the methodology (Figure 1) that we have demonstrated above for v_1.

3.2 Establishing Structural identifiability of parameters based on closed-form solutions

For the proposed methodology (Figure 1) to work, it should be possible to obtain closed form solutions for each parameter in the enzyme kinetic model for each flux as shown in Equation (5). Since the ability to

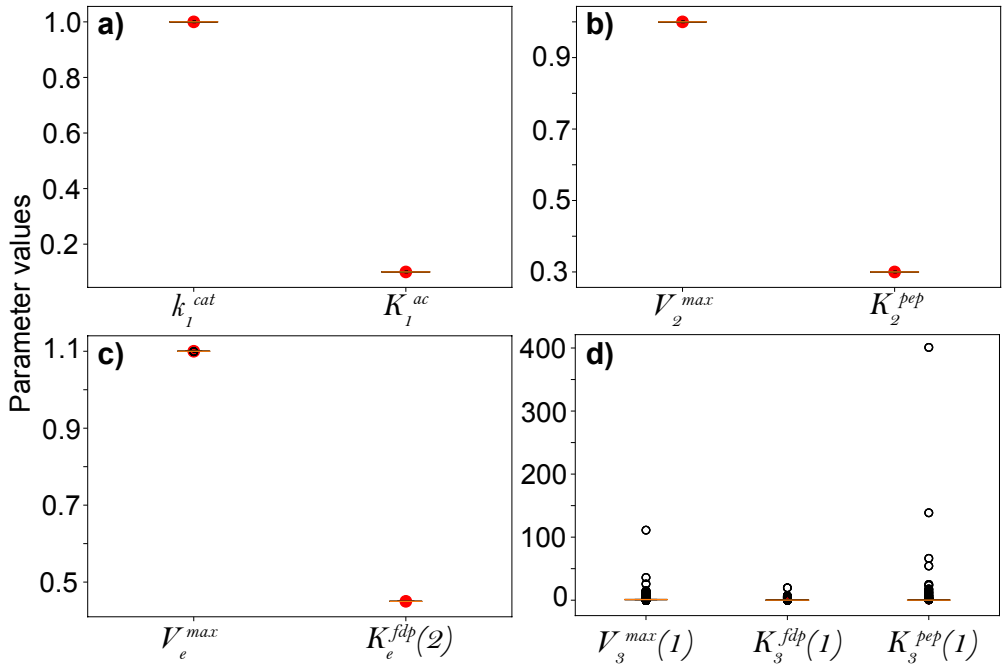

Figure 3. Distribution of predicted parameter values when performing practical identifiability analysis using closed-form solutions for each parameter in flux a) v_1, b) v_2, c) v_5, and d)v_3. For v_1, we have assumed that enzyme concentration is available and have accordingly identified and estimated k_1^{cat}, as opposed to V_1^{max}. The parameter values for only the second root of K_e^{fdp} in v_5 ($K_e^{fdp}(2)$) is shown, since $K_e^{fdp}(1)$ is not estimated by any combination of two experiments, and V_e^{max} is estimated by all combinations. Only one of the two roots for v_3 is shown in panel d. The estimated data for the second root has a similar distribution to that of the first root and is shown in the Supplementary Information. Data is generated using the Convenience Kinetic model for allosteric regulation for v_3.

obtain closed-form solutions for each parameter is dependent on the model structure, any parameter that has non-unique closed-form solutions can be called a structurally non-identifiable parameter. However, if the number of solutions that the parameter has are finite, then the parameter is only locally structurally identifiable.

We demonstrated the structural identifiability of parameters modeling v_1 in Section 3.1. We have shown that the parameters have only one unique closed-form solution, and accordingly are structurally identifiable. Since v_2 is also expressed using the Michaelis-Menten model, just like v_1, we find that the parameters (V_2^{max} and K_2^{pep}) are also structurally identifiable. The closed-form expressions for these parameters are similar to the ones shown in Equation 17, with ac replaced by pep, and v_1 replaced by v_2.

However, we find that the parameters used to model v_3 using the Convenience kinetics rate law, and v_5 using the Hill kinetic rate law are not structurally identifiable as they are given in Section 2.6.

First, for v_3, we find that the parameters V_3^{max}, K_3^{fdp} and K_3^{pep} have two different closed-form solutions. Thus, based on the presence of non-unique but finite number of possible solutions for these parameters we can classify v_3 as a locally structurally identifiable flux. In order to alleviate local structural identifiability, we reduced the dimension of the parameter space for v_3. Originally, $\theta \in \mathbb{R}^3$ for v_3. By reducing the dimension of θ to $\mathbb{R}^2$, we were able to obtain a structurally identifiable model for v_3. To reduce the dimension of the parameter space for v_3, we fix either K_3^{fdp} or K_3^{pep} as a known quantity, and identify the other unfixed parameter along with V_3^{max}. This results in unique closed-form expressions for both V_3^{max} and the other unfixed parameter (K_3^{pep} or K_3^{fdp}).

While v_3 is an allosterically regulated metabolic flux, v_5 describes a transcription/translation reaction using Hill kinetics. We apply our proposed methodology to identify parameters modeling v_5 using only the available experimental data on the metabolite concentrations and the fluxes within the metabolic network. We could not obtain closed form solutions for parameters V_e^{max}, K_e^{fdp} and n_e in v_5 using the computer algebra system. So, instead of changing the model as we did for v_3 (see Section 2.6), we resorted to reducing the dimension of the parameter space by fixing one of the three parameters, the Hill coefficient n_e. We illustrated the consequence of reducing the dimension of the parameter space of the Convenience kinetic model for v_3 earlier. With a fixed and known n_e, K_e^{fdp} has two possible closed-form solutions, which make

it a locally structurally identifiable parameter. On the other hand, V_e^{max} has only one unique closed-form solution, and therefore is structurally identifiable.

We have now established conditions for structural identifiability of all the major fluxes in the small metabolic network (Figure 2). We have shown how our proposed methodology can be used to establish conditions for structural identifiability using steady state information on the model variables. We next discuss the practical identifiability of the parameters in v_2, v_3 and v_5 whose parameters are structurally identifiable only under certain conditions.

3.3 Relationship between structural and practical parameter identifiability

We mention in Section 2.2 that, by definition, unique parameter values based on the model structure are possible for any structurally identifiable parameter. Together with this definition for structural identifiability, we also introduced the concept of practical parameter identifiability. To recall, we mentioned that it should be possible to estimate unique parameter values based on all available experimental data for any practically identifiable parameter.

As shown in Figure 1 and illustrated for v_1 in Section 3.1, to determine the practical identifiability of parameters we test for the existence of a non-zero denominator of the closed-form expressions of the parameters. We also reduce the possible space within which a parameter could be practically identifiable by checking for the physiological feasibility of the parameter values that are obtained through this analysis (Figure 1b). If the resulting parameter values obtained from various combinations of experimental data for each closed-form expression are unique, then the parameter is practically identifiable. However, if a non-unique number of parameter values are possible from multiple combinations of experimental steady state data, then the parameter is said to be practically non-identifiable. In conjunction with the conditions for structural identifiability demonstrated earlier in Section 3.2, if the parameter has only one unique closed-form expression, and its value is also unique, then the parameter is both structurally and practically identifiable. If either of these conditions are not satisfied, the parameters can be either locally structurally or practically identifiable or non-identifiable.

Accordingly, both v_1 and v_2 are not only structurally identifiable due to the presence of unique closed-

form expressions for their parameters, they are also practically identifiable because the parameters in the respective models possess unique values based on distinct combinations of experimental data (Figure 3a and b).

Regarding v_5, we showed earlier in Section 3.2 that the identifiability of v_5 can be analyzed only when the Hill coefficient n_e is held constant. So, in subsequent discussions, the dimension of the v_5 parameter space is kept at $\mathbb{R}^2$ by fixing the value of n_e. Under these conditions, we find that the structurally identifiable parameter V_e^{max} is also practically identifiable, i.e., it has only one unique value based on all available in silico experimental data (Figure 3c). However, recall that unlike V_e^{max}, K_e^{fdp} is only locally structurally identifiable as it has two possible closed-form expressions. Nonetheless, despite its local structural identifiability, we find that the K_e^{fdp} is also practically identifiable, like V_e^{max}, with only one unique parameter value (Figure 3c).

We find that the practical identifiability of v_5, despite the local structural identifiability of one of its parameters, is due to the enforcement of the physiological relevance criteria on the parameters i.e., only one of the two closed-form expressions for K_e^{fdp} is physiologically relevant. The other solution always acquires a negative value that has no physiological meaning. Thus, by reducing the practically identifiable space of parameters, we have shown that our methodology can establish global practical identifiability even when the parameters are only locally structurally identifiable.

Similar to K_e^{fdp} in v_5, we also explained the local structural identifiability of V_3^{max}, K_3^{fdp} and K_3^{pep} modeling v_3 in Section 3.2. These parameters have two possible closed-form expressions. In Figure 3d we show the numerical values for one of the two possible closed-form expressions for V_3^{max}, K_3^{fdp} and K_3^{pep}. The numerical values for the second closed-form expressions of the three parameters is presented in Supplementary Figure, and they also have a similar distribution. Based on the prior definition for practically identifiable parameters, the numerous possible values that the three parameters can acquire (Figure 3d) leads us to conclude that the parameters in v_3 are practically non-identifiable when they are only structurally locally identifiable. Also, unlike v_5, which is practically identifiable in the presence of local structural identifiability, parameters for v_3 are practically non-identifiable even after the reduction in the practically identifiable parameter space realized using the physiological relevance condition (Figure 1b).

However, we find that v_3 is practically identifiable when its parameters are also structurally identifiable

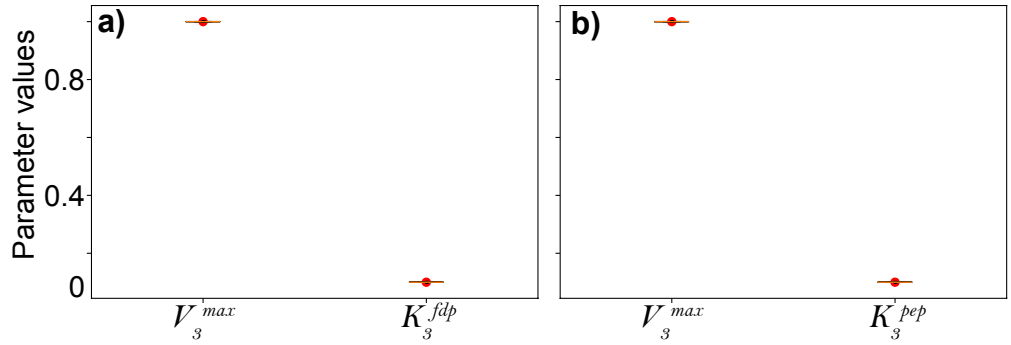

Figure 4. Distribution of predicted parameter values when performing practical identifiability analysis using closed-form solutions for each parameter in flux v_3. The globally identifiable parameter values of a) V_3^{max} and K_3^{fdp} when K_3^{pep} is held constant, and b) V_3^{max} and K_3^{pep} when K_3^{fdp} is held constant.

(Figure 4). Earlier in Section 3.2 we had mentioned that V_3^{max} and K_3^{fdp} are structurally identifiable only when K_3^{pep} is fixed, and V_3^{max} and K_3^{pep} are structurally identifiable when K_3^{fdp} is fixed. Under these scenarios we find the structurally identifiable parameters to also be practically identifiable (Figure 4).

In conjunction with the practical identifiability of v_5 established earlier, we see that it is possible to delineate between structural and practical identifiability of parameters in kinetic models of metabolism only under certain conditions, and not in others.

3.4 A priori experimental design through practical parameter identification

The analysis of parameter practical identifiability can be used to gather information on the type of experiments that can provide useful data for parameter estimation. For instance, during practical identifiability analysis, if either the denominator of the closed-form expression is zero, or if the parameter values that are obtained are not physiologically feasible (Figure 1b), then the experimental data set concerned is said to be incapable of practically identifying that said parameter. Consequently, the data from that combination of experiments is considered non-informative. When this analysis is repeated for multiple combinations of steady state data from the 21 different in silico experiments, we can determine the number of different experimental data combinations that can practically identify each parameter in each flux (Figure 5).

As described in Section 2.5, the information on the number of experimental data sets that can practically identify each parameter can be used to determine the degree of identifiability of the corresponding parameters.

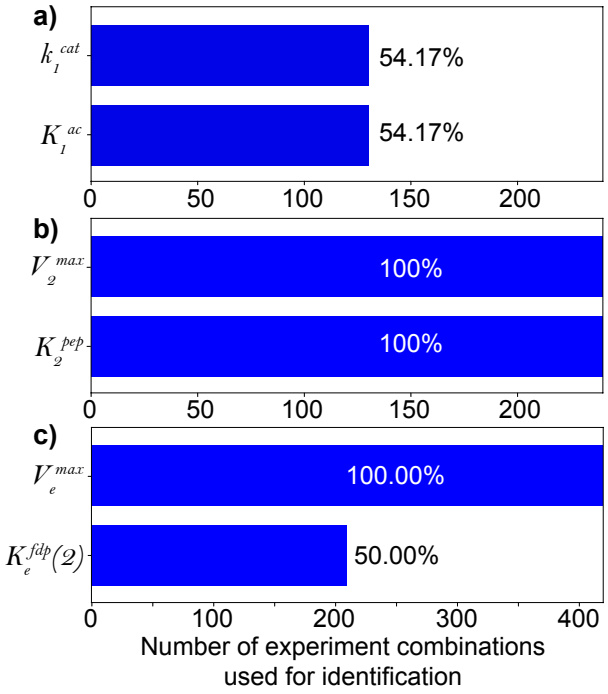

Figure 5. The number of data combination from 21 different in silico experiments that can practically identify each parameter in fluxes a) v_1, b) v_2, and c) v_5 when there is no noise in the input experimental data. The percentage of total combinations of experimental data used for analysis (240 for v_1 and v_2, and 421 for v_5) that can identify each parameter is also specified. v_1, v_2 and v_5 require data from two experiments for analysis. The contribution of different experiment type towards identifying each parameter is shown in the spider plots.

Subsequently, this information can be used to classify parameters based on their ease of identifiability.

In Figure 5 we show the number of experimental data combinations that are capable of identifying each parameter, and consequently, the degree of identifiability of each parameter (percentage experimental data combinations that are capable of identifying each parameter) in flux v_1 (Figure 5a), v_2 (Figure 5b) and v_5 (Figure 5c). The degree of identifiability for V_1^{max} and $K_1^{ac}(ne)$ in v_1 is shown in Supplementary Figures S2, and the degree of identifiability for both closed-form expressions of the three parameters of v_3 are shown in Supplementary Figure S3. The degree of identifiability of each parameter is also given in these figures as percentages.

It is important to recall that both v_1 (Supplementary Figure S1) and v_3 (Figure 3d) are not practically identifiable (Section 3.3). While v_1 becomes practically identifiable (Figure 3a) if enzyme concentrations are utilized to alleviate uncertainties in the enzyme turn over rates (k_1^{cat}), v_3 becomes practically identifiable only when the dimension of the parameter space is reduced (Figure 4). In these cases the degree of identifiability (Supplementary Figure S2 for v_1 and Supplementary Figure S3 for v_3) refers to the number of experimental data sets that can determine physiologically relevant values for the corresponding parameters.

Based on their degrees of identifiability, we see that the maximum reaction rates (V_i^{max}) are more or similarly identifiable in comparison to the corresponding enzyme binding (K_i) constants or the activation/inhibition constants, in the respective reaction rate law models (Figure 5, Supplementary Figures S2 and S3). We make this observation despite the fact that the four fluxes are modeled using three different enzyme kinetic rate laws: v_1 and v_2 are modeled using the Michaelis-Menten rate law, v_3 is modeled using the Convenience kinetic rate law and v_5 is modeled as a Hill equation with inhibition. As mentioned earlier, for v_1 (Supplementary Figure S2) and v_3 (Supplementary Figure S3) we have shown that a greater number of experimental data sets can predict physiologically relevant or non-zero positive values for V_1^{max} and V_3^{max} than for $K_1^{ac}(ne)$ and K_3^{fdp} or K_3^{pep}, respectively.

The degree of identifiability of v_3, when its parameters are structurally and practically identifiable (Sections 3.2 and 3.3) are shown in Supplementary Figure S4. Accordingly, we find that with the exception of V_1^{max} (Supplementary Figure S2) and k_1^{cat} in v_1 (Figure 5a), all data sets used to test practical identifiability can determine unique values for parameters when the corresponding parameter is structurally identifiable.

We can attribute the difference in the degree of identifiability between v_1 (Figure 5a and Supplementary Figure S2) and the other fluxes (v_2, v_3 and v_5) to the ability of data from different combinations of experiments to satisfy the conditions for practical identifiability of that parameter, that can be determined a priori. In systems identification terminology, data requirements for parameter identification can be tied to selecting experiments that are persistently excitable for the flux being identified. Any input signal should be rich or informative enough to guarantee full excitement of the dynamics of the system (Ljung AND Glad 1994). Only information obtained from such changes in the input can be used to completely identify the system over its entire dynamic range. So, the ability of data from a combination of different experiments to practically identify parameters of a given flux is governed by the ability of the experiment to generate distinct measured concentrations and fluxes that will satisfy the identifiability conditions.

In turn, the degree of identifiability of parameters and the informativeness of the corresponding experiments used to identify them can be explained by the position of the flux in the metabolic network. The position of any given flux in the metabolic network determines the specific experiment that is persistently excitable enough to identify the parameters of that flux. This dependency can be further elucidated using v_1 and v_2 as examples.

We know from Equation (18) and Section 3.1 that for a combination of any two experiments to be capable of identifying v_1, the experiments must generate data that have distinct acetate concentrations, E and v_1. We also know, based on our knowledge of the Michaelis-Menten kinetic rate law that changes in the substrate concentration of a reaction can bring about a nonlinear change in the value of the corresponding reaction rate. So, in this instance, since the substrate is an input variable to the model, and v_1 is the corresponding uptake flux and E is a system variable, the substrate can be easily perturbed to create persistently excitable experiments to identify parameters in v_1. We see the consequence of this requirement in the degree of identifiability of k_1^{cat} and K_1^{ac} (Figure 5a). We can generalize this observation for the identification of all uptake fluxes in all metabolic networks, i.e., at a minimum, a change in the input substrate concentration may be necessary for an informative experiment to identify the uptake flux parameters.

Similarly, the identification of parameters for v_2 (Figure 5b) requires that persistently excitable experiments distinguish between values of both v_2 as well as *pep*. However, since both of these are system outputs,

satisfaction of this condition cannot be guaranteed without an analysis of the dynamics of the metabolic network, and how changes in the input (acetate) bring about changes in the two requisite output quantities. Previous dynamical analysis of the network (Figure 2) has already established the existence of a functional relationship between *pep* and v_2, and the input acetate concentration and the levels of expression of the different enzymes within the network (Srinivasan, Cluett, AND Mahadevan 2017). The 100% degree of identifiability seen for v_2 (Figure 5b) confirms the theoretical possibility for any type of perturbation experiment to be persistently excitable to identify v_2. Overall, this analysis informs us that the degree of identifiability and consequently, the type of experiments needed to identify different parameters varies widely depending on the position of the flux with respect to the inputs and the outputs of the metabolic network, as well as the various regulatory interactions present within the network (e.g., effect of *pep* on v_3, or the effect of *fdp* on v_5 and consequently on v_1 in Figure 2).

From the above example we can summarize that identification of individual fluxes within a metabolic network necessitates a careful consideration of experiments such that the data acquired can satisfy conditions for practical identifiability for all parameters modeling a flux, and subsequently, all fluxes within a network (Heijnen AND Verheijen 2013). To facilitate the design of experiments based on their ability to satisfy requirements for practical identifiability of parameters, we determine the occurrence of each type of steady state perturbation experiment within combinations that can practically identify each parameter (Figure 6, Supplementary Figures S2, S3 and S4). So, with our proposed methodology it is possible to identify the types of perturbation experiments that would be informative for identifying each parameter in each flux with steady state concentration and flux data.

As mentioned in Supplementary Section S3.1, we use experimental data from five different types of experiments to test the practical identifiability of parameters in the model (Supplementary Table S1). In Figures 6a, 6b and 6c, the contribution from different experiment types for identifying parameters in v_1, v_2 and v_5 are respectively shown as spider plots.

The contribution of experiments that involve changes in the acetate concentrations, which consequently bring about changes in the value of v_1, contribute to a significant part ($> 50\%$) of the identifiable experimental data combinations for v_1 in comparison to the other types of experiments (Figure 6a and Supplementary

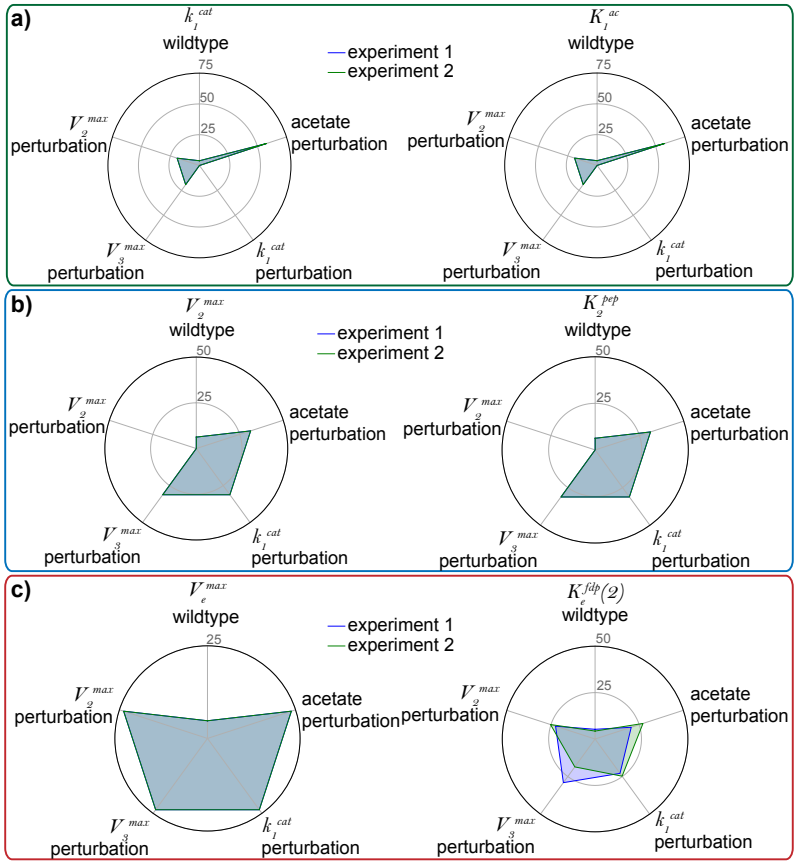

Figure 6. Frequency of each type of perturbation experiment in data sets that can identify parameters in fluxes a) v_1, b) v_2 and c)v_5. The frequencies are represented as a percentage of the total number of experiments present within all the data sets that can practically identify the parameters. The frequency of experiments that occur as the first experiment in a combination of two experiments are shown as a blue surface, and the frequency of experiments that occur as the second experiment in the combination are shown as a green surface.

Figure S2). This is in agreement with the condition for identifiability that we discussed earlier (Equations 18 and 17). Since less than 50% of all data combinations can satisfy these requirements, and can consequently identify v_1 (Figure 5a), we also say that identifiability analysis is crucial to determine the minimum number of experiments, along with the nature of experiments that can help identify parameters for v_1.

With v_2, we see that the enzyme perturbations as well as the acetate perturbation experiments have similar contributions towards datasets that can identify v_2 (Figure 6b). This also supports our arguments made earlier with regards to the identifiability conditions for v_2, and the reasons for the difference in the type of experiments that are informative between v_1 and v_2. Accordingly, we find that in comparison to selecting experiments to identify v_1, there is very little restriction on the types of experiments that are informative to identify v_2.

We can also extend these observations to justify the observed contribution of experiments towards identifying parameters for v_5 (Figure 6c), or determining physiologically relevant parameter values for structurally locally identifiable parameters of v_3 (Supplementary Figures S3 and S4).

In all of the above scenarios for v_1 (Figure 6a), v_2 (Figure 6b) and v_3 (Supplementary Figures S3 and S4), the distribution of experiment types between the two (v_1 and v_2) or three (v_3) required experiments is quite similar. Hence, the green/blue/yellow surfaces in the spider plots are superimposed upon each other. This is also seen for experiments identifying V_e^{max} in v_5 (Figure 6c). However, this is not the case for K_e^{fdp} in v_5 (Figure 6c). We see that when two experiments are required to identify K_e^{fdp}, the choice of the first experiment has a bearing on the choice of the second experiment, and vice-versa, so that data with enough information is available for the identification of K_e^{fdp}. Also, since V_e^{max} is globally identifiable using any experiment type (Figures 5c and 6c), the choice and number of experiments required to identify v_5 completely hinges upon the identifiability of K_e^{fdp} from the chosen experiments.

So, now we have established a hierarchy of parameters based on their identifiability. Parameters for v_2 and V_e^{max} in v_5 that are most identifiable are at the top of the hierarchy, while v_1 and K_e^{fdp} in v_5 fall at the bottom of the list as the experiments needed for their identification require careful consideration. When v_3 is structurally identifiable, their parameters also do not require any experimental design considerations.

3.5 Parameter non-identifiability due to uncertainty in Experimental Data

In all the aforementioned scenarios, the kinetic rate law from which data is derived is known and same as the model for which parameters are estimated. However, in reality, the kinetic rate law based on which metabolic networks function and from which in vivo experimental data is extracted is mostly unknown. The rate laws are primarily inferred through the parameter estimation procedure. This is one of the motivations for the development of approximate kinetic rate law models (Heijnen AND Verheijen 2013; Smallbone, ET AL. 2007; Berthoumieux, ET AL. 2013). So, there is a need to see if the methodology that we have developed here is capable of handling the uncertainty that arises due to the mismatch between the model and the data used to identify and estimate the parameters in the model.

The scope within which we have defined the model (Section 2.6) makes such an analysis possible by changing the enzyme kinetic rate law used to describe v_3. Note that the original description (Kotte, ET AL. 2014; Srinivasan, Cluett, AND Mahadevan 2017) of the network (Figure 2) uses the Monod-Wyman-Chageaux (MWC) model to describe the flux through v_3. Whereas, so far we have used a Convenience kinetic rate law description for both data generation as well as identifiability analysis. To determine the ability of our methodology to handle the in vivo-in vitro model uncertainty, we use the MWC model description to generate the in silico experimental data. This data will then be used to identify parameters in all the fluxes, including v_3 that is described by the Convenience kinetic model.

First, we find that the spread in the estimated Convenience kinetics parameter values, when v_3 is only locally structurally identifiable, is much larger than when there is mismatch between the model generating the data and the model that is being identified (Supplementary Figure). A more important observation is that even when the parameters are structurally identifiable in v_3 (achieved by assuming either K_3^{fdp} or K_3^{pep} as a known constant), they can at most only be locally practically identifiable. This is shown by the spread in the estimate values of the structurally identifiable parameters when steady state data based on the MWC model is used in Supplementary Figure.

Second, note that the dynamics of the network as represented by an MWC model for v_3 are different from the dynamic characteristics expressed when a Convenience kinetics model is used instead to describe v_3. Thus, this can bring about a change in the steady state concentrations and fluxes observed for the various

in silico experiments listed in Supplementary Table S1. For instance, since the enzyme concentration E is dependent on the dynamics of the network, the uptake flux v_1 can be different between the two models for the same acetate concentration (Equation 10). Consequently, as the enzyme concentration E is not part of the closed-form expression for V_1^{max} and $K_1^{ac}(ne)$ in Equation 17), the difference in the steady state data used for identification can result in a change in the spread (uncertainty) observed for estimated values of V_1^{max} and $K_1^{ac}(ne)$ (Supplementary Figure). Thus, while quantifiable, the uncertainty due to mismatch in the in vivo and in vitro information will carry over to the estimated parameters.

However, this issue can be resolved if more in vivo information is used for parameter identification. We first observe this scenario when Equation (18), which includes E, is used to identify k_1^{cat} and K_1^{ac}: these parameters are practically identifiable even when in silico steady state data from a mismatched model is used for identification (Supplementary Figure). We also observe this with the identification of v_2 and v_5 (Supplementary Figure). For these two fluxes all available and necessary steady state information are part of their identifiability expressions, thereby leaving no room for any uncertainties to propagate from the data through the practical identification process.

Apart from the mismatch between the in vivo and the in vitro enzyme kinetic rate laws, uncertainty in experimental data also arises due to the presence of noise in the measured experimental data. This noise could be attributed to the measurement error commonly encountered in process analytics. In order to test the robustness of our methodology to practically identify parameters using steady state data with measurement errors, we used in silico experimental data with 5% additive noise for practical identification, instead of the noise-free data that we have used so far.

We found that in every case where parameters are structurally identifiable, the noise did not have any effect on the identifiability of the parameters or their estimated values. We also found that inclusion of all necessary data (e.g., the presence or lack thereof of enzyme concentration E for v_1) can alleviate issues related to using experimental data with errors for identification and estimation: the degrees of identifiability of V_1^{max} and $K_1^{ac}(ne)$ had non-zero standard deviations associated with them, but the degrees of identifiability of k_1^{cat} and K_1^{ac} did not. Using a similar reasoning to the earlier scenario in the presence of mismatches between in vitro and in vivo model, we can say that there is no room for any uncertainties to propagate from the noisy

data when all necessary steady state information for identification is available. Thus, both v_2 and v_5 also did not show any differences in either their degree of identifiability or their estimated parameter values.

However, for v_3, whose parameters are only locally structurally identifiable, we found small non-zero standard deviations in the degrees of identifiabilities (Supplementary Figure) when noisy data is used. Although this was seen due to the differences in the number of data combinations that can estimate positive values for each of the three parameters between different noisy experimental data sets, we observe that the standard deviation in the estimated parameter values for each data, between different samples of noisy experimental data, is small (Supplementary Figure).

4 Discussions

Parameter estimation for kinetic models has always focused on the ability to estimate parameters from existing data without the need for additional experiments, which might not be always possible if parameters are not identifiable from existing experimental data. The presence of noise is typically said to be a significant factor that results in non-identifiability. However, there different reasons for non-identifiability of parameters that we show with our work. First, non-identifiability could be structural to the model used to represent the flux, and cannot be alleviated without reduction in the parameter space. Otherwise, non-identifiability of parameters can be attributed to the lack of information about the dynamics of the system whose parameters are being estimated within the chosen experimental data. The informativeness of experiments can be tied back to their ability to discriminate the dynamics of the system under two or more different input conditions. Thus, the presence of noise only serves to exacerbate the inability of experiments to discriminate the dynamics of the systems.

Previously, methods have been developed for practical parameter identification and experimental design for kinetic models of metabolism. These methods for experimental design based on practical identification of parameters rely on solving nonlinear least squares problems using optimization approaches that cannot guarantee global optimal solutions (**Raue2009a**), or calculating the Fischer Information Matrix (FIM) to obtain information on the structural and practical identifiability of parameters in kinetic models. Either of these types of methods become computationally cumbersome for models of large genome-scale, or even

central carbon scale metabolic networks. Some authors have eschewed deterministic parameter estimation techniques in favour of Bayesian methods based on probabilistic estimation of parameters and experimental design (**Saa2016a**; Saa AND Nielsen 2016) that has the possibility of overcoming some of the issues with the deterministic techniques.

In this document, we have presented a scalable method to practically identify parameters in kinetic models of metabolism, and use it to design experiments that are minimal and informative for estimating the parameters that does not require solutions to non-convex optimization problems. By establishing identifiability for each flux within a metabolic network individually, we hope to overcome the scalability obstacle. Furthermore, we believe our method offers an algorithmic alternative to determine persistently excitable experiments that can enable identification of all fluxes within a metabolic network. Using a small metabolic network for gluconeogenesis, we have demonstrated that the identifiability of parameters for a given flux is dependent on the position of the flux within the metabolic network. We have also shown the ability to use our analysis to design the minimal number of experiments that are most informative for identifying all fluxes within a metabolic network.

We find that the identifiability of parameters in kinetic models of metabolism using steady state information is dependent on the kinetic rate law used to model the fluxes within metabolism. The impact of the formulation and nonlinearity of a kinetic rate law expression affecting the practical identifiability of parameters in the expression may not be an unique problem isolated to the system that we are investigating. Complicated expressions for describing fluxes have been extensively used to model observed experimental data for different fluxes in a variety of organisms (**Chassagnole2002a**; **Peskov2012**; **VanHeerden2014**). However, authors have favored working with approximate kinetic models of metabolism whose parameters are easily identifiable and estimable instead of trying to establish the identifiability of the parameters used in these models (mention Heijnen papers on resolving identifiability using approximate models here).

We have shown that in some instances (e.g., v_5) local practical identifiability could be resolved to obtain global practical identifiability using constraints on the values of the parameters such that they are physically relevant. We have also shown that the structural identifiability of the parameters in any given kinetic rate law model has a bearing on the ability to determine the practical identifiability of parameters using steady

state metabolomic, fluxomic and proteomic information. We find that these can sometimes be resolved by reducing the dimension of the parameter space that is being identified: $\theta \in \mathbb{R}^3$ to $\theta \in \mathbb{R}^2$ for both v_3 and v_2. Additionally, we would also like to point out that discrepancies between in vivo kinetic rate law from which typical experimental data is obtained, and the in vitro rate law used in kinetic models can itself lead to practical parameter non-identifiability or local identifiability. This can lead to uncertainty in parameter estimates made from in vivo experimental data.

Our work adds to this existing body of work wherein we develop a method for practical identifiability tailored for use with nonlinear enzyme kinetic rate laws that are typically used to model fluxes in metabolic networks. With our work we hope to change the status quo in the application of systems identification techniques for kinetic models of metabolic networks. Our methodology fills the niche gap of experimental design for parameter estimation by providing a way to design informative experiments to obtain data required for parameter estimation by spending the least amount of resources. In the future, we believe our work can be extended and formulated as a mixed integer linear programming problem that can be solved to determine the type and total minimum number of experiments necessary to estimate all parameters in kinetic models of genome-scale metabolic networks.

References

Andreozzi, S., A. Chakrabarti, ET AL. (2016) Identification of metabolic engineering targets for the enhancement of 1,4-butanediol production in recombinant E. coli using large-scale kinetic models, *Metab. Eng.* 35, 148–159.

Andreozzi, S., L. Miskovic, AND V. Hatzimanikatis (2016) iSCHRUNK – In Silico Approach to Characterization and Reduction of Uncertainty in the Kinetic Models of Genome-scale Metabolic Networks, *Metab. Eng.* 33, 158–168.

Apaolaza, I., ET AL. (2017) An in-silico approach to predict and exploit synthetic lethality in cancer metabolism, *Nat. Commun.* 8.1, 459.

Berthoumieux, S., ET AL. (2013) On the identifiability of metabolic network models, *J. Math. Biol.* 67.6-7, 1795–1832.

Bordbar, A., D. McCloskey, ET AL. (2015) Personalized Whole-Cell Kinetic Models of Metabolism for Discovery in Genomics and Pharmacodynamics, *Cell Syst.* 1.4, 283–292.

Bordbar, A., J. M. Monk, ET AL. (2014) Constraint-based models predict metabolic and associated cellular functions, *Nat. Rev. Genet.* 15.2, 107–120.

Chandrasekaran, S., ET AL. (2017) Comprehensive Mapping of Pluripotent Stem Cell Metabolism Using Dynamic Genome-Scale Network Modeling, *Cell Rep.* 21.10, 2965–2977.

Di Filippo, M., ET AL. (2016) Zooming-in on cancer metabolic rewiring with tissue specific constraint-based models, *Comput. Biol. Chem.* 62, 60–69.

Gadkar, K. G., R. Gunawan, AND F. J. Doyle (2005) Iterative approach to model identification of biological networks, *BMC Bioinformatics* 6.1, 155.

Heijnen, J. J. (2005) Approximative kinetic formats used in metabolic network modeling, *Biotechnol. Bioeng.* 91.5, 534–545.

Heijnen, J. J. AND P. J. T. Verheijen (2013) Parameter identification of in vivo kinetic models: Limitations and challenges, *Biotechnol. J.* 8.7, 768–775.

Khodayari, A., ET AL. (2016) A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains, *Nat. Commun.* 7, 13806.

Kotte, O., ET AL. (2014) Phenotypic bistability in Escherichia coli's central carbon metabolism. en, *Mol. Syst. Biol.* 10.7, 736.

Liebermeister, W. AND E. Klipp (2006) Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theor. Biol. Med. Model.* 3, 41.

Link, H., D. Christodoulou, AND U. Sauer (2014) Advancing metabolic models with kinetic information, *Curr. Opin. Biotechnol.* 29.1, 8–14.

Ljung, L. AND T. Glad (1994) On global identifiability for arbitrary model parametrizations, *Automatica* 30.2, 265–276.

Maia, P., M. Rocha, AND I. Rocha (2016) In Silico Constraint-Based Strain Optimization Methods: the Quest for Optimal Cell Factories. *Microbiol. Mol. Biol. Rev.* 80.1, 45–67.

Nikerel, I. E., ET AL. (2009) Model reduction and a priori kinetic parameter identifiability analysis using metabolome time series for metabolic reaction networks with linlog kinetics, *Metab. Eng.* 11.1, 20–30.

Raue, A., ET AL. (2014) Comparison of approaches for parameter identifiability analysis of biological systems, *Bioinformatics* 30.10, 1440–1448.

Saa, P. A. AND L. K. Nielsen (2016) Construction of feasible and accurate kinetic models of metabolism: A Bayesian approach. *Sci. Rep.* 6, 29635.

Saa, P. A. AND L. K. Nielsen (2017) Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks, *Biotechnol. Adv.* 35.8, 981–1003.

Smallbone, K., ET AL. (2007) Something from nothing - Bridging the gap between constraint-based and kinetic modelling, *FEBS J.* 274.21, 5576–5585.

Srinivasan, S., W. R. Cluett, AND R. Mahadevan (2015) Constructing kinetic models of metabolism at genome-scales: A review. *Biotechnol. J.* 10.9, 1345–59.

— (2017) Model-based design of bistable cell factories for metabolic engineering, *Bioinformatics*.

Vanlier, J., C. A. Tiemann, ET AL. (2012) A Bayesian approach to targeted experiment design, *Bioinformatics* 28.8, 1136–1142.

Vanlier, J., C. Tiemann, ET AL. (2013) Parameter uncertainty in biochemical models described by ordinary differential equations, *Math. Biosci.* 246.2, 305–314.

Vanlier, J., C. a. Tiemann, ET AL. (2014) Optimal experiment design for model selection in biochemical networks Optimal experiment design for model selection in biochemical networks, 1–22.

Zerfaß, C., J. Chen, AND O. S. Soyer (2018) Engineering microbial communities using thermodynamic principles and electrical interfaces, *Curr. Opin. Biotechnol.* 50, 121–127.