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:

Kinetic models of metabolism can be used to study the dynamic characteristics of metabolic networks (Bordbar, ET AL. 2015; Andreozzi, Chakrabarti, ET AL. 2016). In these models, ordinary differential equations (ode) are used to express the rate of change of metabolite concentrations (x) as a function of the reaction fluxes (v) in the metabolic network (Equation 1). The matrix $\mathbf{S}$ 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 a kinetic model (Equation 1b) is dependent on the enzyme kinetic mechanism that is used to model the reaction (Link, Christodoulou, AND Sauer 2014; Srinivasan, Cluett, AND Mahadevan 2015). Accordingly, f is a nonlinear function of the metabolite concentrations, enzyme kinetic parameters (θ) and other input concentrations (u).

Unlike CBMs, where the ability to predict the steady state responses of metabolic networks is dependent on the stoichiometry of the network, the prediction of dynamic responses of metabolic networks to

different perturbations using kinetic models of metabolism is also dependent on the numerical values of the enzyme kinetic parameters (θ in Equation 1) (Andreozzi, Miskovic, AND Hatzimanikatis 2016). 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 parameter values might play a crucial role (Srinivasan, Cluett, AND Mahadevan 2015; Vital-Lopez, Maranas, AND Armaou 2006). The use of in vitro, or unreliable in vivo parameter estimates reduces confidence in the model predicted behaviour (Andreozzi, Miskovic, AND Hatzimanikatis 2016). Consequently, the reduction in confidence hampers the use of these models to gain insight into the functioning of metabolic networks (Chakrabarti, ET AL. 2013; Bordbar, ET AL. 2015). The corresponding increase in uncertainty in model predicted responses becomes an obstacle for using the predicted responses as a basis for designing the metabolic networks to achieve any of the aforementioned goals of metabolic engineering and design.

Parameter estimation methods based on optimization principles are typically used to determine true parameter values based on available experimental data. Under the assumption that all time dependent intracellular metabolite concentrations can be measured, a parameter estimation problem can be formulated as a nonlinear programming problem (Equation 1) to estimate the values of enzyme kinetic parameters, θ, based on the measured data. The minimization of least square error between the measured (x^*) and modeled (x) concentrations, weighted by the variance in the experimental data σ^*_{kl} for each concentration at each time point, is used as an objective function (Equation 2a) for the optimization problem (Equation 2). The parameter values are determined within fixed upper (θ_u) and lower (θ_l) bounds (Equation 2b).

$$\min_{\theta} \sum_{k=1}^{m} \sum_{l=1}^{d} \left(\frac{x^*_{kl} - x_{kl}}{\sigma^*_{kl}} \right)^2 \tag{2a}$$

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

However, not all metabolite concentrations used in the model (Equation 1) can be measured. Additionally, measurable fluxes in the metabolic network also need to be included as part of the parameter estimation problem. In such scenarios, the parameter estimation problem is modified to suit a new system of equations shown below (Equation 3). The new system of equations is obtained by augmenting the original system

(Equation 1) with Equation (3c) that models the relationship between the measurable metabolite concentrations and fluxes (y) and the unmeasured concentrations (x) that are used in the original model (Equation 1) above. The parameter vector (θ) is augmented with additional parameters that define this relationship. These additional parameters also need to be estimated.

$$\dot{x} = \mathbf{S}v \tag{3a}$$

$$v = f(x, y, \theta, u) \tag{3b}$$

$$\dot{y} = h(x, y, \theta, u) \tag{3c}$$

In systems identification, the measured concentrations and fluxes (y) are called output or observed variables, and the unmeasured concentrations (x) are called the state variables. For estimating θ, the metabolite concentrations x in the optimization problem (Equation 2) are substituted with the output variables y.

The ability to determine unique solutions to parameters θ is governed by the identifiability of these parameters in the model (McLean AND McAuley 2012; Raue, ET AL. 2009). The identifiability of parameters in nonlinear models can be classified into two categories: structural (or a priori) and practical (or posterior) identifiability. Any system (Equation 3) is 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 (4) below.

$$\Phi(\theta_1, u) = \Phi(\theta_2, u) \iff \theta_1 = \theta_2 \tag{4}$$

Accordingly, the system can have a unique solution, a finite number of non-unique solutions or an infinite number of solutions for all input functions, and is 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 unobservable state variables and the observable output variables. 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 (4) is referred to as practical identifiability. So, the effect of the available experimental data on the ability to estimate unique parameter values is determined by the practical identifiability of the parameter. Accordingly, practical identifiability of a parameter is contingent upon the nature, quality and quantity of data available to estimate the parameter as opposed to the structure and parameterization of the model.

Thus, 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 also are 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.

Methods and tools for structural identification of parameters based on differential algebra (Ljung AND Glad 1994; Audoly, ET AL. 2001; Bellu, ET AL. 2007) and profile likelihood (Raue, ET AL. 2009) are available. However, only the profile likelihood-based methods enable experimental design by facilitating practical identification of parameters. Nonetheless, this method still depend on solving a non-convex nonlinear least squares problem (Equation 2) to get likelihood estimates of parameters, and hence still suffers from all the inherent difficulties associated with obtaining global optimal solutions for non-convex optimization problems. This also makes it un-scalable for experimental design and practical identifiability of parameters in kinetic models of large metabolic networks.

In this paper, we propose a scalable methodology to establish practical identifiability for parameters in kinetic models of metabolism using steady state concentration and flux data. We present a computer algebra-based method that can facilitate experimental design through practical identification of parameters separately for each individual reaction within a metabolic network. For the purposes of this method we assume that all intracellular metabolite concentrations and fluxes can be measured. We identify and design experiments to estimate parameters in a small metabolic network model of glucoeneogenesis in *Escherichia coli* (Kotte, ET AL. 2014; Srinivasan, Cluett, AND Mahadevan 2017) to illustrate the utility of our method.

2 Methods

2.1 A method to determine practical identifiability of kinetic models of metabolism

We provide the mathematical framework for practical 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. In kinetic models of metabolic networks (Equation 1), the fluxes are expressed as a function of the metabolite concentrations x and the kinetic parameters θ (Figure 1a). The value of every flux v_i is expressed using one of the many available enzyme kinetic formulations (Equation 3b). Without loss of generality, all of these kinetic formulations can be expressed as nonlinear algebraic equations (Figure 1a).

Let $\theta \in \mathbb{R}^p$ in Equation (3b) for each flux v_i in the network. As stated earlier, for each experiment $j = 1, 2, ..., n$, we assume that all metabolite concentrations (x) and reaction fluxes (v) are measurable. The pertinent information for each experiment 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, 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 θ for each flux v_i (Figure 1b). This is done by solving a system of nonlinear algebraic equations in $\mathbb{R}^p$ for each flux v_i, shown in Equation (5).

$$v_{i,k} = f_k(\mathbf{x}_k, \theta, u_k) \qquad \forall k = \{1, 2, ..., p\} \subset \{1, 2, ..., n\} \tag{5}$$

Each equation in (5), indicated by the index k, corresponds to the kinetic rate law expression $f(x, \theta, u)$ for v_i, described earlier in Equation (3b), written for concentrations and fluxes obtained from experiment k. Solving the system in Equation (5) results in $\mathbb{R}^p$ nonlinear expressions for parameters in θ (Equation 6), 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).

$$\theta_k = g_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u}) = \frac{N_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u})}{D_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u})} \tag{6}$$

The identifiability of parameter θ_k for flux v_i can be established by determining the value of $D_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u})$

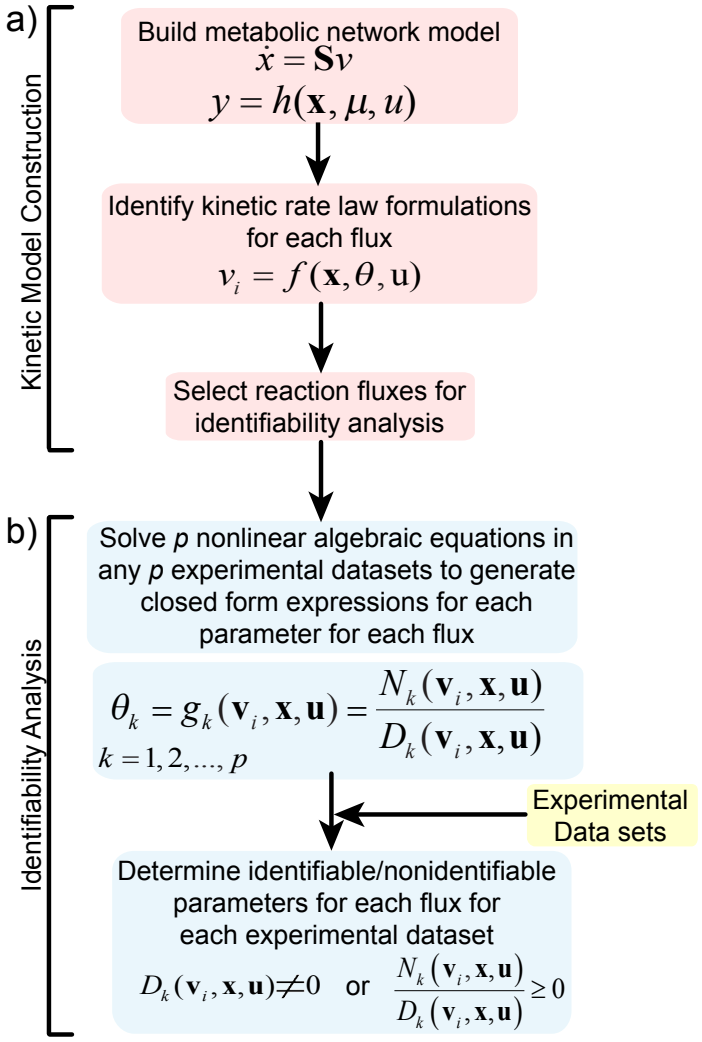

Figure 1. A flow diagram showing the methodology developed to establish practical identifiability of parameters in kinetic models of metabolism. a) The steps for the construction of a kinetic model of a metabolic network. The choice of rate law formulations to describe metabolic fluxes influences the 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.

(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 non-negative value of $g_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u})$ as identifiable (Figure 1b).

In the following sections we provide a previously published kinetic model of a small gluconeogenic network (section 2.2), followed by a demonstration of our methodology to establish practical identifiability for one of the fluxes in this network (section 2.3).

2.2 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 the utility of identifiability analysis for experimental design for parameter estimation in kinetic models of metabolism. The kinetic model is described below.

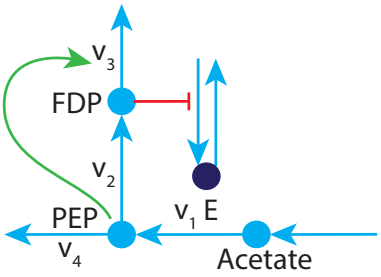

Figure 2. The small metabolic network for gluconeogenesis used to demonstrate our practical identifiability method for kinetic models of metabolism.

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

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

$$\frac{d}{dt}E = v_{e,max}\left(\frac{1}{1+\left(\frac{fdp}{K_e^{fdp}}\right)^{n_e}}\right) - dE \tag{9}$$

The kinetic expressions for fluxes v_1 through v_4 are given below. The consumption of acetate through v_1 and conversion of *pep* through v_2 are expressed in Equations (10) and (11) respectively using Michaelis-Menten

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

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

$$v_2 = V_2^{max} \frac{pep}{pep + K_2^{pep}} \tag{11}$$

$$v_3 = V_3^{max} \frac{f\tilde{dp}\left(1 + f\tilde{dp}\right)^3}{\left(1 + f\tilde{dp}\right)^4 + L_3\left(1 + \frac{pep}{K_3^{pep}}\right)^{-4}} \tag{12}$$

The allosterically regulated flux v_3 for the consumption of *fdp* is expressed in Equation (12) using the Monod-Wyman-Changeux (MWC) model for allosterically regulated enzymes, where $f\tilde{dp}$ refers to the ratio of *fdp* with respect to its allosteric binding constant K_3^{fdp}. The flux v_4 for the export of *pep* is expressed as a linear equation dependent on *pep* in Equation (13).

$$v_4 = k_4^{cat}.pep \tag{13}$$

2.3 Identifiability of parameters in a kinetic model of gluconeogenesis

Here, we demonstrate the use of our computer algebra-based methodology to establish practical identifiability of parameters for flux v_1 in the small model of gluconeogenesis (Figure 2) described in section 2.2.

In flux v_1, the concentration of the enzyme E is used as a variable. Since we assume that steady state experimental information is only available for metabolite concentrations and fluxes, the expression previously given for v_1 (Equation 10) 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 (14). 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.

$$v_1 = V_1^{max} \frac{ac}{ac + K_1^{ac}(ne)} \tag{14}$$

We choose the expression for flux v_1 given in Equation (14) to demonstrate our method for practical identifiability.

As mentioned in section **??**, practical identifiability is a necessary condition to estimate true values of both V_1^{max} and K_1^{ac} from experimental data. Here, we assume that the concentrations and fluxes from at least two different experiments are available i.e., in Equation (5) $k = \{1, 2\}$. We label the available concentrations and fluxes as $ac^{(k)}$ and ${v_1}^{(k)}$, respectively. Then, the nonlinear algebraic equations shown in Equation (5) can be formulated for v_1 as:

$$ {v_1}^{(k)} = V_1^{max} \frac{ac^{(k)}}{ac^{(k)} + K_1^{ac}(ne)} \qquad\qquad k = \{1, 2\} \tag{15} $$

Solving this simultaneous system of k equations using Mathematica (Wolfram Research, USA), a computer algebra system, we get $p = 2$ nonlinear algebraic expressions for parameters V_1^{max} (Equation 16a) and $K_1^{ac}(ne)$ (Equation 16b). These expressions have the form previously shown in Equation (6).

$$ V_1^{max} = \frac{v_1^{(1)} v_1^{(2)} (ac^{(1)} - ac^{(2)})}{v_1^{(2)} ac^{(1)} - v_1^{(1)} ac^{(2)}} \tag{16a} $$

$$ K_1^{ac}(ne) = \frac{ac^{(1)} ac^{(2)} (v_1^{(1)} - v_1^{(2)})}{v_1^{(2)} ac^{(1)} - v_1^{(1)} ac^{(2)}} \tag{16b} $$

In Equation (16), the denominator of the right hand side expression is used to test the identifiability of parameters V_1^{max} (Equation 16a) and $K_1^{ac}(ne)$ (Equation 16b) for different available experimental data combinations. 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 (16) should be non-negative (Figure 1b).

2.4 Data for establishing parameter identifiability in kinetic model of gluconeogenesis

Steady state information on the metabolome and the fluxome can be gathered under different physiological conditions. One way to alter the physiological conditions is to change the substrate concentration under

which the cells grow. In the small metabolic network (Figure 2), the acetate concentration plays the role of a substrate, and determines the acetate uptake flux v_1. Thus, steady state metabolite concentrations and fluxes can be calculated under various acetate concentrations to form different experimental data combinations that measure cellular response to changes in the substrate concentration.

Physiological changes can also be brought about by perturbing the expression levels for different enzymes within a metabolic network. The model of gluconeogenesis (Figure 2) described in section 2.2 has three different fluxes (v_1, v_2 and v_3) whose enzyme expression parameters (k_1^{cat}, V_2^{max} and V_3^{max}) can be perturbed to simulate the repression and over expression of the corresponding enzymes. Accordingly, in addition to measuring network response to substrate perturbations, changes in steady state concentrations and fluxes can also be observed for enzyme expression perturbations.

In total, based on the above discussion, we can perturb four different model parameters ($acetate$, k_1^{cat}, V_2^{max} and V_3^{max}) to obtain experimental data in silico. The 18 different parameter values used to generate experimental data are given in Table 1.

Table 1. Table showing the perturbed values of all parameters used to generate experimental data in silico for testing practical identifiability of all fluxes in the small metabolic network.

Experiment Type	Perturbed Parameter (wild type value a.u.)	Perturbed Values				
wild type						
acetate perturbation	$acetate$ (0.1)	0.05	0.09	0.11	0.5	1.0
v_1 perturbation	k_1^{cat} (1)	0.5	0.9	1.1	1.5	2.0
v_2 perturbation	V_2^{max} (1)	0.5	0.9	1.1	1.5	2.0
v_3 perturbation	V_3^{max} (1)	0.5	0.9	1.1	1.5	2.0

The minimum number of experiments from which data is required for identifying all the parameters of a given flux is determined by dimension $\mathbb{R}^p$ of the parameter space of a chosen flux v_i. For instance, as demonstrated above, data from two distinct experiments is required for identifying v_1, as v_1 has two parameters ($\mathbb{R}^2$). v_2 is also modeled by a parameter space with dimensions of $\mathbb{R}^2$. Hence, multiple combinations of data generated from any two different experiments are used to test the identifiability of v_1 and v_2. The total number of such possible combinations is 306 (18 x 17) from the 18 different experiments shown in Table 1. Similarly, for identifying v_3, that is described by three parameters ($\mathbb{R}^3$), we need data from a combination

of three different experiments. Based on the 18 experiments in Table 1, we have 4896 distinct combinations (18 x 17 x 16) of three experiments.

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 50% of the combinations can identify another parameter θ_j, then θ_j has a degree of identifiability of 0.5 or 50%. 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 50% 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 Experimental design through practical parameter identification

Following the methodology described in section 2.1, and demonstrated in section 2.3 for a single flux using data from a combination of two different experiments, all distinct combinations found based on experiments described in section 2.4 can be tested for their ability to practically identify any of the three 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 distinguishing 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

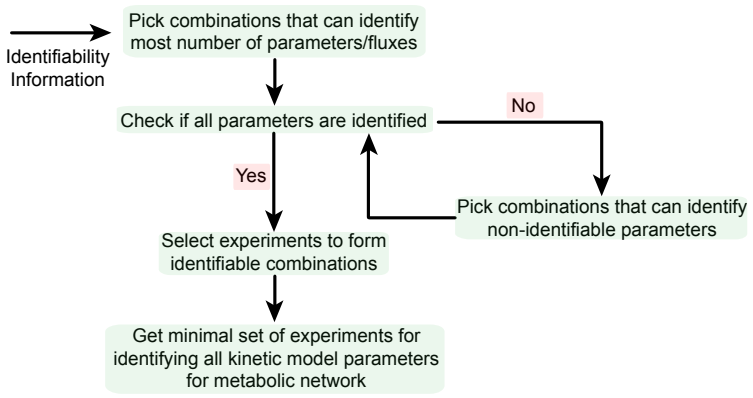

Figure 3. Flow diagram showing a method for experimental design that uses our methodology for practical identification of parameters to determine the number and type of experiments required to identify all fluxes within a given metabolic network.

performed to identify as many model parameters as possible (Figure 3). Consequently, the set of experiments can be used to estimate the identifiable parameters in the model. This is formally explained below.

The identifiability of each parameter based on each experiment indexed as $j = \{1, ..., n\}$ is established based on the methodology described previously in section 2.1 and demonstrated in Section 2.3 (Figure 1b). Subsequently, for any flux v_i and for any p combinations of indices j, if the experimental concentrations $(\mathbf{x}_j)$ and fluxes $(\mathbf{v}_j)$ do not satisfy the condition for identifiability for any parameter θ_k in $\theta \in \mathbb{R}^p$, i.e., $D_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u}) = 0$ for any k, 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 parameter estimation and needs to be discarded from the set of all necessary experiments. Furthermore, another experiment from $j = \{1, ..., n\}$ needs to be selected 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 pre-selected experiments satisfy the identifiability condition, then we can design an $(n+1)^{th}$ 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 show that the ability to establish practical identifiability of kinetic parameters in metabolic network models using our methodology, described in section 2.1, relies upon the nonlinearity of the kinetic rate law formulations used to describe the fluxes. We use the gluconeogenic model (Figure 2) described in section 2.3 as an example. Then, in section 3.2 we provide a motivation for the need for experimental design, especially to identify kinetic models of metabolic networks. We do this by looking at the number of combinations that can identify the maximum number of parameters. In section 3.3 we show that the degree of identifiability of the maximum reaction rate parameter is always higher than the degree of identifiability of the corresponding enzyme binding parameter irrespective of enzyme kinetic rate law used to describe the flux. In section 3.4 we discuss the ability to determine the type of experiments for parameter identification for a given flux based on the informativeness of a given type of experiment. In this section we show how the informativeness of a given type of experiment to identify a specific flux can be deduced from its contribution towards the practical identification of the parameters for a given flux. Finally, we discuss some results arising out of the use of our methodology to determine the identifiability of parameters when data with additive noise is used in section 3.5.

3.1 Nonlinearity of enzyme kinetic rate law expression affects identifiability analysis

Practical identifiability of parameters in kinetic models of metabolism using the methodology described in section 2 is governed by the nonlinear complexity of the enzyme kinetic rate law used to model a specific flux. We demonstrate one example of how the methodology works in section 2.3 for parameters of flux v_1. The expression in Equation (16) is obtained by using a computer algebra system. To recall, in this specific case, a computer algebra system is used to solve for the parameters of a flux described using the Michaelis-Menten kinetic rate law when data from two different experiments is available.

However, 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 and 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 17).

$$v_3 = V_3^{max} \left(\frac{1}{1 + \frac{K_3^{pep}}{pep}} \right) \left(\frac{\frac{fdp}{K_3^{fdp}}}{1 + \frac{fdp}{K_3^{fdp}}} \right) \tag{17}$$

Using this expression for analysis, we find that each of the parameters V_3^{max}, K_3^{fdp} and K_3^{pep} have two different closed-form expressions owing to the presence of a square root term in their solutions. These distinct expressions are denoted by (1) and (2) following the respective parameter names throughout the rest of the document: $V_3^{max}(1)$, $K_3^{fdp}(1)$, $K_3^{pep}(1)$, and $V_3^{max}(2)$, $K_3^{fdp}(2)$, $K_3^{pep}(2)$.

The impact of 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 (Chassagnole, ET AL. 2002; Peskov, Mogilevskaya, AND Demin 2012; Heerden, ET AL. 2014). However, the identifiability of the parameters used in these models has never been truly examined. Unlike the case of the MWC model, if determining closed-form expressions for the parameters of these models are tractable, we believe our methodology can help in elucidating the identifiability of these parameters. If not, metabolic network fluxes can be expressed using alternative kinetic rate law models whose parameters can be tested for identifiability, and subsequently experiments can be designed for their estimation and model validation.

Next, we look at some of the results we obtained for the practical identification of different parameters for all three fluxes v_1, v_2 and v_3 in the small gluconeogenic network model described in section 2.2, with the flux for v_3 described by Equation (17). We start by looking at one of motivations for the need for designing experiments to collect data for parameter estimation.

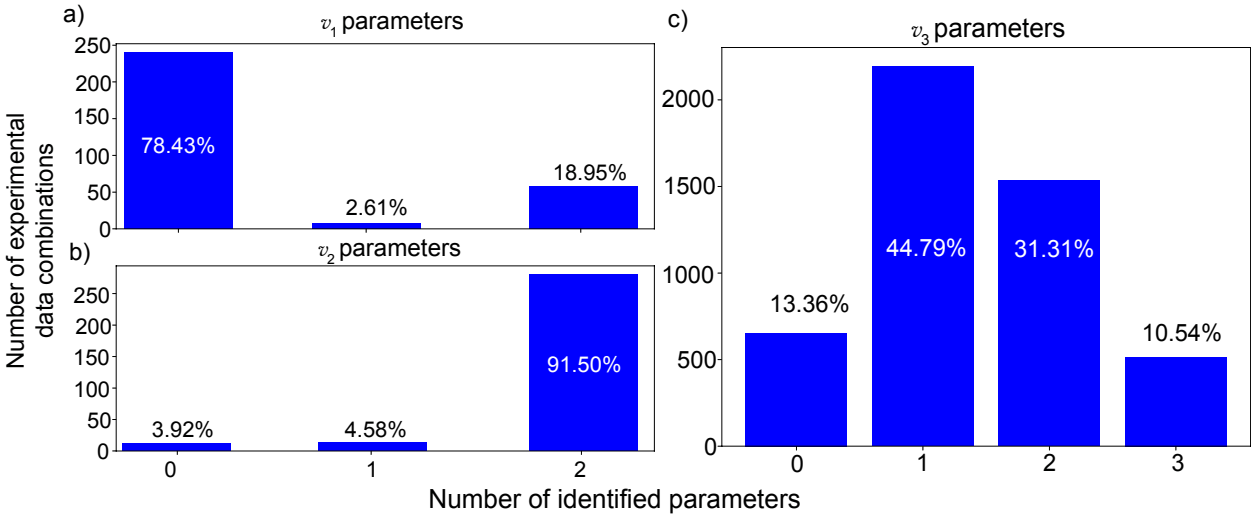

Figure 4. Utility of experimental data combinations on the basis of their ability to identify the most number of parameters. Information is shown for parameters modeling fluxes a) v_1, b) v_2 and c) v_3. The total possible number of parameters that can be identified by data from combinations of a) two, b) two, and c) three experiments is shown on the horizontal axis. The vertical axis represents the total number of combinations that can identify the corresponding number of parameters. The percentages shown in the plots represent the fraction of the total combinations used to test identifiability of parameters for a given flux. A total of 306 data combinations are used for identifiability analysis for a) v_1 and b) v_2, and 4896 combinations are used to practical identifiability of c) v_3. Section 2.4 provides more details on how the combinations of experimental data are generated.

3.2 Utility of experimental data for practical parameter identification

Although the need for better parameter estimation methods is well researched, within the kinetic modeling community, the need for experimental design methods to satisfy the data needs for parameter estimation are not considered seriously. Hence, we find the need to stress the necessity for experimental design methodologies, in this case, tailored specifically for metabolic network models.

We do this by showing how useful the multiple combinations of experimental data are towards identifying parameters for each given flux (Figure 4). We express the usefulness of any given data combination on the basis of the number of parameters that the combination can identify. The higher the number of parameters a data combination can identify, the greater is the usefulness of that combination of experiments.

The number of data combinations that can identify a specific number of parameters within a given flux is shown in Figure 4. The percentage of all data combinations that can estimate a given number of parameters within a given flux are also shown in Figure 4. For flux v_1 (Figure 4a) about 78% of the 306 data combinations cannot identify any parameter used to model the flux (V_1^{max} and $K_1^{ac}(ne)$). Only a single parameter can be identified by about 2% of the combinations, while only about 19% of the data combinations can identify both the parameters used in the model. In contrast, most data combinations (>90%) are not wasted, and can identify both parameters in v_2 (Figure 4b). Less than 5% of the experimental data combinations are wasted from not being able to identify any parameter. In the case of v_3 (Figure 4c), about 45% of the 4896 data combinations can identify only one parameter, while 13% of the combinations cannot identify any parameter. Only 10% of combinations can identify v_3 completely.

Thus, Figure 4 gives credence to the idea that careful experimental design is necessary to minimize usage of resources devoted to performing experiments for parameter estimation. For instance, choosing any of the experiment combinations from the 78% percent that does not identify any parameter in v_1, or the 13% that does not identify any parameter in v_3 can lead to experiments that are not informative. This can lead to potential waste of resources used to perform these experiments.

In addition to providing information on the fraction of experiments that are informative and those that are not informative, Figure 4 also informs on the ease of identifying parameters for a single flux. For the small network (Figure 2), we find that more than 90% of all the data combinations tested for their ability to

identify v_2 are actually capable of identifying V_2^{max} and K_2^{pep} (Figure 4b). In contrast, less than 20% of the combinations can identify both parameters in v_1. Since minimization of resource utilization requires us to use the least number of experiments to identify both v_1 as well as v_2, this indicates the need to carefully choose combinations of experiments that can identify v_1. A careful consideration of experiments for identifying v_2 is however not a major concern since there is a very high probability that experiments chosen to identify v_1 will also identify v_2.

Although the calculation of the utility of experiment combinations in terms of their ability to identify the maximum number of parameters gives an idea on the informativeness of experiments used to gather the data, further analysis from the perspective of identifiability of the parameters of different fluxes is required to be able to design experiments based on this information. Hence, we look at the degree of identifiability (see section 2.5) of each parameter of each flux in the following section to be able to classify parameters, and consequently fluxes, from most identifiable to least identifiable.

3.3 Maximum reaction rates are more identifiable than enzyme binding constants

In Figure 5 we show the number and percentage of combinations that are capable of identifying each parameter in each flux. Based on the definition given in section 2.5, the percentages refer to the degree of identifiability of each parameter. The three panels in Figure 5 represent the degree of identifiability for parameters modeling the three different fluxes of the small network individually.

In regards to the identifiability of v_1 (Figure 5a), v_2 (Figure 5b) and v_3 (Figure 5c), the degree of identifiability of the maximum reaction rates (V_i^{max}) in each of the three fluxes is higher than the degree of identifiability of the corresponding enzyme binding (K_i) constants and the allosteric activation constant (K_3^{pep}). We observe this trend irrespective of the fact that only v_1 and v_2 are represented by the same enzyme kinetic rate law (Michaelis-Menten), while the convenience kinetic rate law is used to model v_3. These observations lead us to conclude that the maximum reaction rate parameters are always more identifiable (as indicated by their higher degree of identifiability) than their enzyme binding constant counterparts, irrespective of the enzyme kinetic rate law used to model the corresponding flux.

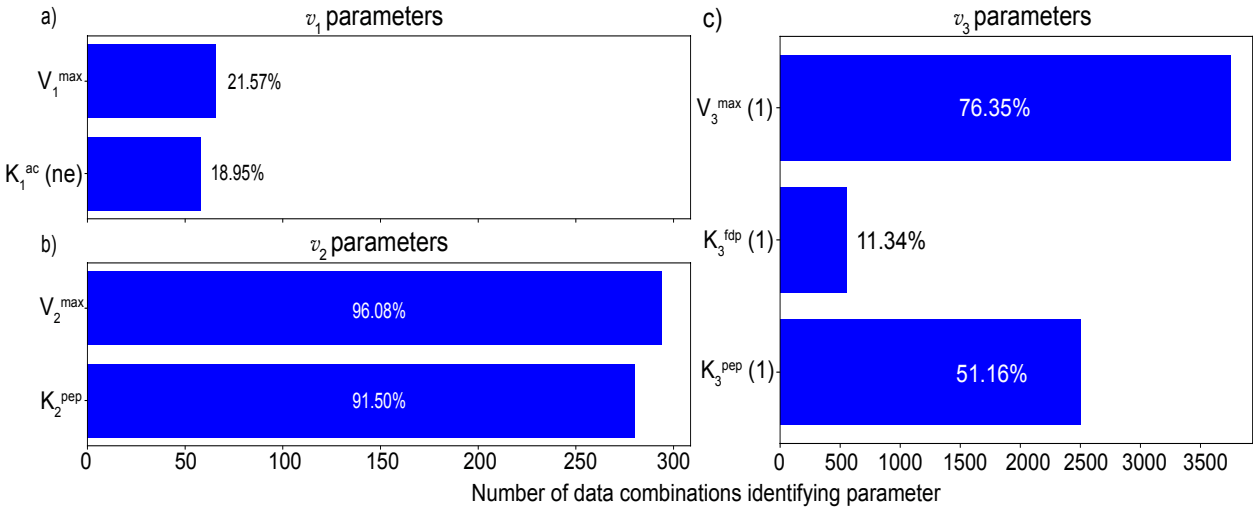

Figure 5. The number of data combination from 18 different in silico experiments that can practically identify each parameter in fluxes a) v_1, b) v_2 and c) v_3 when there is no noise in the input experimental data. The percentage of total experimental data combinations that can identify each parameter is specified either in, or right next to the bar showing the total number of combinations identifying a given parameter. Since the total number of experiments used to identify parameters in v_1 and v_2 is two, the total number of data combinations used to identify parameters in fluxes v_1 and v_2 is lower (at 306) that the total number of combinations used to test identifiability of parameters for v_3 (at 4896), which is modeled with three parameters and requires data from at least three different experiments.

From Figure 5 we see that the variation in the degree of identifiability is not only dependent on the type of parameter being identified, but is also associated with the flux that these parameters model. In systems identification terminology, this can be tied to selecting experiments that are persistently excitable for the flux being identified. In systems identification, 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. Input signals that guarantee to fully excite the dynamics of the system are termed as persistently excitable signals. In linear systems, persistence of excitation of any input signal can be theoretically guaranteed. However, for nonlinear systems, like the metabolic network we deal with in this paper, persistence of excitation of input signals cannot be theoretically guaranteed, and should be assessed on a case by case basis. The lack of persistence of excitation requires the design of experiments to satisfy the data needs for complete system identification.

In the case of identifying fluxes within a metabolic network individually, each flux is not independent from the other fluxes that form the metabolic network. Each flux is connected to one or more fluxes through the metabolites that not only act as substrates and products, but also as allosteric activators and inhibitors (e.g., effect of *pep* on v_3 in Figure 2). Hence, careful consideration of experiments is a necessity so that the data acquired can satisfy conditions for practical identifiability for all parameters modeling a flux, and subsequently, all fluxes within a network. In the next section we discuss the dependence of the variation in the degree of identifiability for parameters of different fluxes from the point of view of the informativeness of experiments from which data is collected for identification.

3.4 Experiment requirements for identifying parameters depend on the position of the flux in the metabolic network

In this section, we demonstrate a specific case of lack of persistence of excitation and the subsequent experimental design required to overcome this hurdle for parameter identifiability for a small nonlinear metabolic network model. Earlier, we mentioned that the informativeness of experiments for identifying parameters in a given flux depends on its ability to satisfy the conditions determined for practical identifiability of that

parameter. For example, we demonstrated in section 2.3 that for a combination of any two experiments to be capable of identifying V_1^{max} and K_1^{ac} in v_1, the experiments must have distinct acetate concentrations as well as a different uptake flux v_1 between them (Equation 16). Thus, in this instance, the informativeness of the experiments (changes in the input acetate concentration and v_1) for identifying parameters of v_1 is determined by the ability of the input change to effect a change in the measured value of v_1. Similarly, the identification of parameters for v_2 requires the experiments to distinguish between values of both v_2 as well as *pep*.

As such, as per Figure 5a only 20% of all data combinations can satisfy these requirements, and can consequently identify v_1. Whereas, more than 90% of the available data combinations satisfy the requirements for identifying v_2 (Figure 5b). Thus, identifiability analysis is crucial to determine the minimum number of experiments, as well as the nature of experiments that can help identify parameters for both v_1 and v_2. Recall that we use experimental data from five different types of experiments to test the practical identifiability of parameters in the model (Section 2.4). In order to determine the type of experiments that help satisfy identifiability conditions for both parameters of v_1, we look at the distribution of the five different types in all data combinations that can identify these 2 parameters (Figure 6a and b).

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 in comparison to the other types of experiments. This matches with requirements for identifiability on the basis of the informativeness of experiments that was laid out earlier i.e., experiments must be able distinguish between different values of v_1 and acetate. Accordingly, only about 20% of the 306 different experiment combinations used to test identifiability are informative enough to discriminate both acetate as well as v_1, and can identify V_1^{max} and $K_1^{ac}(ne)$ (Figure 5a).

In contrast to v_1, we see that the different enzyme perturbation experiments have a higher contribution towards data combinations that can identify parameters for v_2 (Figure 6c and d). Hence, in conjunction with the high degree of identifiability seen in Figure 5b for v_2, we can deduce that almost all experiment types are persistently excitable for v_2 i.e., most experiments can bring about noticeable changes to both *pep* and v_2. Consequently, in comparison to selecting experiments to identify v_1, there is very little restriction

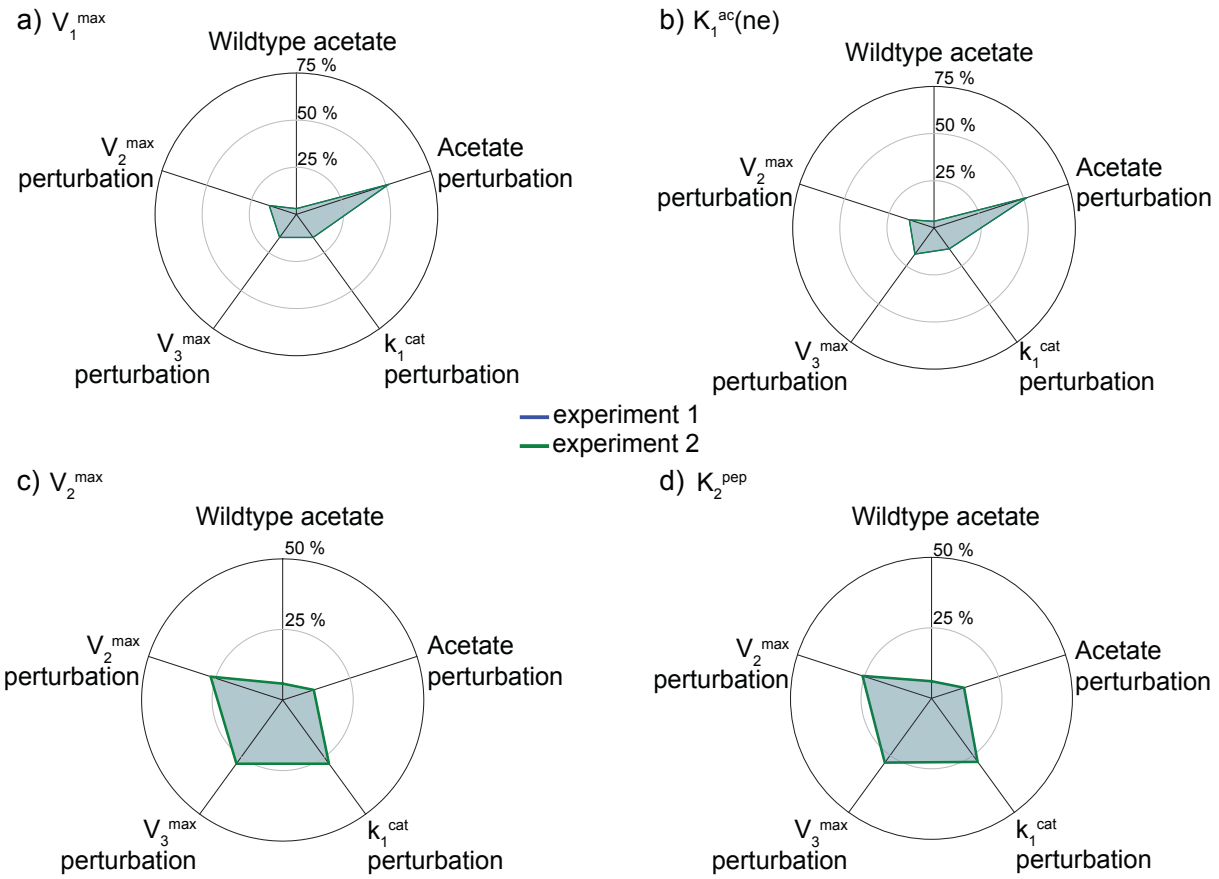

Figure 6. The contribution of different experiments types used in a combination of two experiments ($k = \{1, 2\}$) that can practically identify parameter a) V_1^{max}, b) K_1^{ac}, c) V_2^{max} and d) K_2^{pep}. The percentages reflect the fractional contribution of each experiment type towards all identifiable data combinations.

on the types of experiments that are informative enough to identify v_2. This observations is in line with the one made in section 3.2 based on the total number of combinations that can identify both parameters in v_2.

Both 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 of experiment informativeness on the position of the flux can be further elucidated using v_1 and v_2 as examples. We know from Equation (16) that identifiability of v_1 requires changes in both acetate 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. In this specific metabolic network, since the substrate is an input variable to the model, and v_1 is the corresponding uptake flux, the substrate can be easily perturbed to create persistently excitable experiments to identify parameters in v_1. 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.

On the other hand, the Michaelis-Menten model for v_2 also requires changes in *pep* and v_2 for persistently excitable experiments to identify v_2. 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 in the Appendix) 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). Hence, it is theoretically possible for any of the five different experiment types to be persistently excitable to identify v_2. This is confirmed by the high degree of identifiability for both parameters of v_2, where in more than 90% of all data combinations can identify the parameters. Thus, 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 in the metabolic network with respect to the inputs and the outputs of the network,

as well as the various regulatory interactions present within the network. We can extend these observations to justify the observed contribution of experiments towards identifying parameters for v_3 as well (Figure **??**).

3.5 Effect of noise on degree of identifiability is dependent on nonlinearity of enzyme kinetic rate law models

In reality, experimental data from biological systems is usually noisy. Noise in data gathered from experiments is associated with biological noise attributable to the stochasticity of cellular function, as well as measurement noise associated with the techniques used to obtain concentration and flux measurements. However, the previous sections describe practical identifiability for parameters obtained using noise-free in silico experimental data. So, in order to illustrate the ability of our method to test the practical identifiability of parameters under realistic circumstances, we apply our methodology on in silico data (concentration and flux) with 5% additive noise.

50 different samples of noise with 0 mean and 5% standard deviation are drawn from a normal distribution and added to the steady state experimental data generated on the basis of experiments described in Section 2.4 to generate 50 different samples of noisy data. Subsequently, all 50 samples are tested for their ability to practically identify all three fluxes of the small network (Figure 2). The degree of identifiability of the different parameters using noisy experimental data are shown in the Appendix (Figure **??**).

We hypothesize that the nonlinearity of the enzyme kinetic rate law used to model a flux, and consequently the complexity of the closed form expression obtained for the various parameters used in the model has an impact on the ability of noisy data to practically identify the parameters. The difference in the identifiability attributable to the aforementioned factors can be seen in the difference in the degree of identifiability for the parameters of v_1, v_2 and v_3 in the small metabolic network that we use to demonstrate our methodology.

Despite the presence of noise, the degree of identifiability of v_1 and v_2 (Figure **??**a and b in the Appendix) does not change from the case where no noise is present in the experimental data (Figure 5). However, the degree of identifiability for parameters of v_3 is affected by the presence of noise in the data (Figure **??**c in the Appendix). This effect is represented by the presence of a non-zero standard deviation and corresponding error bars on the graphs in the degrees of identifiability of parameters modeling v_3 (Figure **??**c in the

Appendix). The standard deviation and the error bars represent the presence of data combinations that can identify a given parameter under certain values of noise, but not do so under different noise values. In the presence of 5% additive noise in the data, the conditions under which parameters for v_1 and v_2 can be identified are strictly satisfied by the same data combinations that can identify these parameters in the absence of noise, i.e., the changes in the acetate concentration (for v_1) and *pep* (for v_2) are significant enough to enable identification. The contribution of different experiment types towards identifiability also remains the same (figure not shown separately).

However, as stated in the previous section, the nonlinearity of the original kinetic rate law expression for v_3, and the complexity of the closed-form expression obtained for the parameters in the rate law model preclude us from testing the presence, and consequently the satisfaction of these conditions for parameters in v_3. Furthermore, the conditions for identifiability of v_3 parameters are not as simple as either requiring just a simple difference in *fdp* or *pep* or that of both concentrations. Thus, while under some noise values these identifiability conditions may be satisfied, it may not be possible to satisfy these conditions under other noise values. Accordingly, we end up with a distribution of experimental data combinations that can identify v_3 parameters depending on noise levels in the data. It is significant to note that the variability in the degree of identifiability of parameters is small ($< 2\%$) when noise is present.

These observations, in conjunction with the estimation of the degree of identifiability of the different parameters (Figure 5) help in determining not only the nature of experiments required for parameter identifiability (Figure 6), but also the order in which the type of experiments necessary for identifying the entire metabolic network needs to be evaluated.

For instance, in the case of the example network that we use in this paper, given the high degree of identifiability for v_2 and large spread in the frequency of different experiment types that can contribute to identifiability for v_2, the focus should be on first choosing experiments for identifying v_1 followed by selecting experiments for identifying v_3. It is possible that the three or at most four experiments chosen to identify parameters of both v_1 and v_3 would suffice to identify v_2 as well without the need to perform additional experiments.

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, the primary reason for non-identifiability of parameters from experimental data 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.

Although methods have been developed for practical parameter identification and experimental design, existing methods for experimental design based on practical identification of parameters are based on solving nonlinear least squares problems using optimization approaches that cannot guarantee global optimal solutions (Raue, ET AL. 2009). 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. While some authors have eschewed deterministic parameter estimation techniques in favour of Bayesian methods based on probabilistic estimation of parameters (Saa AND Nielsen 2016b; Saa AND Nielsen 2016a), The need to establish practical identifiability of parameters, even for these methods, still stands. Additionally, the ability to design experiments to get good priors for Bayesian approaches to parameter estimation can also be fulfilled through practical identifiability analysis.

Due to their fundamental aspect of requiring solutions to non-convex optimization problems, existing methods for practical identifications may not be scalable for identification and experimental design for parameters of large metabolic networks. 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.

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.

Audoly, S., ET AL. (2001) Global identifiability of nonlinear models of biological systems, *IEEE Trans. Biomed. Eng.* 48.1, 55–65.

Bellu, G., ET AL. (2007) DAISY: a new software tool to test global identifiability of biological and physiological systems. *Comput. Methods Programs Biomed.* 88.1, 52–61.

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

Chakrabarti, A., ET AL. (2013) Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints, *Biotechnol. J.* 8.9, 1043–1057.

Chassagnole, C., ET AL. (2002) Dynamic modeling of the central carbon metabolism of Escherichia coli, *Biotechnol. Bioeng.* 79.1, 53–73.

Heerden, J. H. van, ET AL. (2014) Lost in transition: start-up of glycolysis yields subpopulations of non-growing cells. *Science* 343.6174, 1245114.

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.

McLean, K. A. P. AND K. B. McAuley (2012) Mathematical modelling of chemical processes-obtaining the best model predictions and parameter estimates using identifiability and estimability procedures, *Can. J. Chem. Eng.* 90.2, 351–366.

Peskov, K., E. Mogilevskaya, AND O. Demin (2012) Kinetic modelling of central carbon metabolism in Escherichia coli, *FEBS J.* 279.18, 3374–3385.

Raue, A., ET AL. (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood, *Bioinformatics* 25.15, 1923–1929.

Saa, P. A. AND L. K. Nielsen (2016a) A probabilistic framework for the exploration of enzymatic capabilities based on feasible kinetics and control analysis, *Biochim. Biophys. Acta - Gen. Subj.* 1860.3.

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

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.

Srinivasan, S., W. R. Cluett, AND R. Mahadevan (2017) Model-based design of bistable cell factories for metabolic engineering, *Bioinformatics*.

Vital-Lopez, F., C. Maranas, AND A. A. Armaou (2006) Bifurcation analysis of metabolism of E. coli at optimal enzyme levels, in: *Proc. 2006 Am. Control Conf.* IEEE, 3439–3444.