

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

Shyam Srinivasan<sup>a</sup>, William R. Cluett<sup>a</sup> and Radhakrishnan Mahadevan<sup>\*,a,b</sup>

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 interaction, 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 assess 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 of 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 implications 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 propose alternative designs to achieve metabolite production goals (Khodayari, ET AL. 2016).

Kinetic models differ from CBMs in their use of heavily parameterized mechanistic enzyme kinetic rate laws to model enzyme catalyzed fluxes within a metabolic network. These parameters represent various aspects of the enzyme kinetic rate laws (Srinivasan, Cluett, AND Mahadevan 2015; Saa AND Nielsen 2017). Hence, the use of kinetic models requires information on all the enzyme kinetic rate laws that will be used to model all the fluxes within a metabolic network, as well as numerical values for all 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 might play a crucial role (Srinivasan, Cluett, AND Mahadevan 2017).

The development of these enzyme kinetic rate laws is based on in vitro observations of enzyme activity. Accordingly, authors have questioned their relevance for gleaned information on the dynamics of metabolism under in vivo conditions, as opposed to in vitro conditions (Heijnen 2005; Heijnen AND Verheijen 2013). This

problem is further compounded by the fact that typical parameter values used in kinetic models of metabolism are estimated in vitro, and may not be applicable under in vivo conditions (Heijnen 2005; Smallbone, ET AL. 2007). Also, parameters estimated based on in vivo experimental data are associated with large uncertainties (Link, Christodoulou, AND Sauer 2014). Using in vitro, or unreliable in vivo parameter estimates to study the steady state and dynamic responses of metabolic networks to different perturbations, under in vivo conditions, reduces confidence in the model predicted behaviour. Consequently, this hampers the use of these models to gain insight into metabolic network function (Andreozzi, Miskovic, AND Hatzimanikatis 2016; Vasilakou, ET AL. 2016), and the increase in prediction uncertainty 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 (Saa AND Nielsen 2017).

Although some authors have sought to quantify this uncertainty using different techniques (Vanlier, C. Tiemann, ET AL. 2013; Andreozzi, Miskovic, AND Hatzimanikatis 2016), others have proposed to alleviate as well as constrain the uncertainty in parameter estimates and consequent model predictions by using a Monte Carlo approach to kinetic modeling of metabolism that allows for integration of experimentally observed in vivo data (Srinivasan, Cluett, AND Mahadevan 2015). Bayesian approaches to improve parameter estimation and quantify estimation uncertainty have also been proposed (Saa AND Nielsen 2016).

In spite of the development of these methods to quantify parameter estimation uncertainty, model parameter identifiability, a necessary, and sometimes sufficient condition to estimate unique kinetic parameter values from experimental data, is often overlooked (Ljung AND Glad 1994; Berthoumieux, ET AL. 2013). Briefly, it concerns with the ability to estimate unique values for all model parameters from observed experimental data. In a model, any parameter is said to be structurally or a priori identifiable if its values can be uniquely estimated independent of all other model parameters from available experimental data. However, if parameters cannot be uniquely estimated independent of each other due to redundant model parameterization, or due to the nonlinear relationship between the model parameters, then the parameters are 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 to overcome concerns with parameter identifiability by proposing approximate kinetic models of metabolism that utilize empirical enzyme kinetic rate laws whose parameters 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 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 can also facilitate the design of experiments that are minimal and informative to generate data required to estimate unique parameter values for all reaction fluxes in a metabolic network. In doing so, we not only propose the number and types of perturbations that will provide the most useful data for parameter estimation, but also test the identifiability of different enzyme kinetic rate laws that are typically used to model fluxes in metabolic networks.

For the purposes of this method we assume that all intracellular metabolite concentrations and fluxes can be measured. We illustrate our methodology to identify parameters and design experiments to identify parameters in a small metabolic network model of gluconeogenesis in *Escherichia coli* (Kotte, ET AL. 2014; Srinivasan, Cluett, AND Mahadevan 2017).

## 2 Methods

### 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 ( $x$ ) 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.

$$\dot{x} = \mathbf{S}v \quad (1a)$$

$$v = f(x, \theta, u) \quad (1b)$$

The expression for the nonlinear function ( $f$ ) used to describe each reaction flux  $v_i$  in  $v$ ,  $i = 1, 2, \dots, n_r$ , in a 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 a nonlinear function of the vector of metabolite concentrations ( $x$ ), the vector of enzyme kinetic parameters ( $\theta$ ) and other input concentrations ( $u$ ).

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 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,  $\theta$ , based on the measured data.

$$\min_{\theta} \sum_{k=1}^m \sum_{l=1}^d \left( \frac{y_{kl}^* - y_{kl}}{\sigma_{kl}^*} \right)^2 \quad (2a)$$

$$\theta_l \leq \theta \leq \theta_u \quad (2b)$$

Here  $y = [x, v]^T$  is the vector of both concentrations ( $x$ ) and fluxes ( $v$ ). The minimization of least square error between the measured ( $y^*$ ) and modeled ( $y$ ) concentrations and fluxes, weighted by the variance in the

experimental data  $\sigma_{kl}^*$  for each concentration and flux, 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 ( $\theta_u$ ) and lower ( $\theta_l$ ) 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  $\theta$  in any nonlinear model (Equation 1) are said to be structurally identifiable if, for an input-output mapping defined by  $y = [x, v]^T = \Phi(\theta, u)$  for at least one input function  $u$ , any two values of parameters  $\theta_1$  and  $\theta_2$  satisfy the relationship in Equation (3):

$$\Phi(\theta_1, u) = \Phi(\theta_2, u) \iff \theta_1 = \theta_2 \quad (3)$$

Accordingly, if parameters in  $\theta$  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  $\theta$  is said to be globally practically identifiable. Whereas, if parameter estimates with quantifiable uncertainties can be found, then the  $\theta$  is said to be locally identifiable. The absence of unique parameter estimates for  $\theta$

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 practical 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  $v_i$ ,  $i = 1, 2, \dots, n_r$ , 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  $j = 1, 2, \dots, n_E$ , we assume that all metabolite concentrations ( $x$ ) and reaction fluxes ( $v$ ) are measurable. We discuss the implications of relaxing this assumption later. 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  $v_i$ ,  $i = 1, 2, \dots, n_r$ , 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  $\theta$  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 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).

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

Here,  $v_{i,j}$  refers to the value of 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  $\theta$  is a vector in  $\mathbb{R}^p$ , whose elements are denoted by  $\theta_k$ .

Each equation in (4), indicated by the index  $j$ , corresponds to the kinetic rate law expression  $f(x, \theta, u)$  for each  $v_i$ ,  $i = 1, 2, \dots, n_r$ , 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  $\theta_k$  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  $j = 1, 2, \dots, p$ .

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

The identifiability of parameter  $\theta_k$ ,  $k = 1, 2, \dots, p$ , 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  $\theta_k$  is said to be 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). The solution  $g_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u})$  in Equation (5) is unique for an identifiable  $\theta_k$ , and an infinite number of solutions are possible for a non-identifiable  $\theta_k$ . However, if there are multiple but finite solutions  $g_k(\mathbf{v}_i, \mathbf{x}, \mathbf{u})$ , then the corresponding parameter  $\theta_k$  is locally identifiable.

## 2.4 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  $\theta_i$ , then the degree of identifiability of  $\theta_i$  is said to be 0.9 or 90%. On the other hand, if only 50% of the



combinations can identify another parameter  $\theta_j$ , then  $\theta_j$  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  $\theta_i$  and  $\theta_j$  that have degrees of identifiability of 90% and 50% respectively,  $\theta_i$  is classified to be more identifiable than  $\theta_j$  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.

## 2.5 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 found based on experiments described in section S3.2 of the Supplementary Information 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 2.4) 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 2). 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  $j = 1, 2, \dots, n_E$  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  $j = 1, 2, \dots, p$ ) do not satisfy the condition for identifiability for any parameter  $\theta_k$  in  $\theta \in \mathbb{R}^p$  (Figure 1b), then at least one of the  $p$  experiments needs to be changed to make parameter  $\theta_k$  identifiable. Consequently, the corresponding experiment cannot be used for estimating parameter  $\theta_k$ , and needs to be discarded from the set of all necessary experiments. Furthermore, another experiment from  $j = 1, \dots, n_E$  needs to be selected to replace the discarded experiment such that parameter  $\theta_k$  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 2.3 to practically identify parameters in flux  $v_1$  of the small gluconeogenic network (Figure S1, model given in section S3.1 of the Supplementary Information). Identifiability results for fluxes  $v_1$ ,  $v_2$ ,  $v_3$  and  $v_5$  are also discussed in this section. We show that the ability to establish practical identifiability of kinetic parameters in metabolic network models relies upon the nonlinearity of the kinetic rate law formulations used to describe the fluxes. We also 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.

Then, 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. Here we show how the informativeness of a given type of experiment to identify a specific flux can be deduced from its contribution

222 towards the practical identification of the parameters for a given flux.

223 Then, in section 3.3 we provide a motivation for the need for experimental design, especially to identify  
 224 kinetic models of metabolic networks. We do this by looking at the number of combinations that can identify  
 225 the maximum number of parameters. In section ??

226 Finally, we discuss some results arising out of the use of our methodology to determine the identifiability  
 227 of parameters when data with additive noise is used in section 3.5 

### 228 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 S16 and Section S3.1 of the Supplementary Information). We choose the expression for flux  $v_1$  given in Equation (S5) for this demonstration.

$$v_1 = V_1^{max} \frac{ac}{ac + K_1^{ac}(ne)} \quad (S5)$$

Since  $\theta = \{V_1^{max}, K_1^{ac}(ne)\} \in \mathbb{R}^2$  for  $v_1$ , as mentioned in Section S3.2, 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 S3.2 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:


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


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 6a)

and  $K_1^{ac}(ne)$  (Equation 6b). These expressions have the form shown in Equation (5).

$$\theta_1 = V_1^{max} = \frac{v_1^{(1)}v_1^{(2)}(ac^{(1)} - ac^{(2)})}{v_1^{(2)}ac^{(1)} - v_1^{(1)}ac^{(2)}} \quad (6a)$$

$$\theta_2 = K_1^{ac}(ne) = \frac{ac^{(1)}ac^{(2)}(v_1^{(1)} - v_1^{(2)})}{v_1^{(2)}ac^{(1)} - v_1^{(1)}ac^{(2)}} \quad (6b)$$

In Equation (6), the denominator of the right hand side expression is used to test the identifiability of parameters  $V_1^{max}$  (Equation 6a) and  $K_1^{ac}(ne)$  (Equation 6b) 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 (6) should be non-negative (Figure 1b). 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 available, in addition to the measured concentration of the metabolites and fluxes. In doing so, we get two expressions similar to the one shown in Equation (6) for  $k_1^{cat}$  and  $K_1^{ac}$ . Here, the value of  $V_1^{max}$  in Equation (S5) above is substituted with  $V_1^{max} = k_1^{cat}E$  instead. The globally identifiable parameter values for all fluxes modeled as Michaelis-Menten equations ( $v_1$  and  $v_2$ ) in provided in  Supplementary Figure S2a and S2b.

We apply the method described in Section 2.3 and Figure 1b, and illustrated above for  $v_1$ , to other fluxes of the network (Figure S16) not described using Michaelis-Menten equations ( $v_3$  and  $v_5$ ). On applying this method to  $v_3$ , which is modeled using MWC kinetics, we find that we could not apply our methodology to determine closed form solution for all the parameters in  $v_3$ . However, when  $v_3$  is instead modeled using Convenience kinetics, we find that the parameters are locally identifiable based on the definition we provide in Section 2.3. All three parameters  $V_3^{max}$ ,  $K_3^{fdp}$  and  $K_3^{pep}$  used to model  $v_3$  have two possible roots. We still could not determine identifiable values for these parameters (Supplementary Figure ) using the constraints on the physical relevance of their values (see Section 2.3 and Figure 1). In the box plot in Supplementary Figure S3a and b we show all the possible estimated parameter values for both roots of all three parameters modeling  $v_3$ , when experimental data is generated using the MWC model for  $v_3$ . Due to the numerous estimated parameter values, we can designate these parameters are practically non-identifiable. Although

using the Convenience kinetics model for  $v_3$  to generate the in silico experimental data leads to a reduction in the uncertainty of the estimated parameters (Supplementary Figure S3c and d), we still end up with non-unique parameter estimates. This leads us to conclude that under the current conditions, flux  $v_3$  is structurally locally identifiable and practically non-identifiable.

However,  $K_3^{fdp}$  and  $V_3^{max}$  become globally identifiable when  $K_3^{pep}$  is assumed to be known and fixed at a constant value, and  $K_3^{pep}$  and  $V_3^{max}$  become globally identifiable when  $K_3^{fdp}$  is assumed to be known and fixed at a constant value (Supplementary Figure S4).

Through Supplementary Figure S4, we also want to bring attention to the fact that parameter non-identifiability remains when there is a mismatch between the kinetic rate law used to generate the in silico data, and the kinetic rate law used to identify parameters. For example, in the specific case of  $v_3$ , we see that globally structurally identifiable parameters  $V_3^{max}$  and  $K_3^{fdp}$  (when  $K_3^{pep}$  is fixed) are practically non-identifiable when data generated from experiments using the MWC kinetics is used to model  $v_3$  (Supplementary Figure S4a). The same is true of globally structurally identifiable parameters  $V_3^{max}$  and  $K_3^{pep}$  when  $K_3^{fdp}$  is fixed (Supplementary Figure S4b). Instead, when Convenience kinetics is used to generate in silico data, under the aforementioned conditions all the parameters are globally structurally and practically identifiable (Supplementary Figure S4c and d).

Similarly, we applied the demonstrated methodology to identify flux  $v_5$  that describes a transcription/translation reaction using Hill kinetics, based on only the available metabolomics and the fluxomics information. When we tried identifying the three parameters  $V_e^{max}$ ,  $K_e^{fdp}$  and  $n_e$  in  $v_5$ , we could not obtain closed form solutions for these parameters using the computer algebra system. However, when we fixed the Hill coefficient  $n_e$  as a known constant parameter, we were able to obtain closed form solutions for  $V_e^{max}$  and  $K_e^{fdp}$ . Thus, similar to  $v_3$  earlier, the dimension of the parameter space for  $v_5$  had to be reduced from  $\mathbb{R}^3$  to  $\mathbb{R}^2$  to enable the application of the procedure that we have described in Section 2.3. In this instance, the identification required data from two different experiments. Again, just as we see for  $v_3$ , there are two different roots for  $K_e^{fdp}$ , making it locally structurally identifiable. On the other hand,  $V_e^{max}$  has only one possible root, thereby it is a globally structurally identifiable parameter. However, only one of the two roots ( $K_e^{fdp}(2)$ ) is practically identifiable based on the physical relevance criteria stated in Section 2.3 and Figure

1b. Thus, enforcing this criteria results in global practical identifiability of both  $V_{e\max}$  and  $K_e^{f_{dp}}$  in  $v_5$  (Supplementary Figure S2c).

In the following section, we describe and summarize the results on the identifiability of parameters in these fluxes ( $v_1$ ,  $v_2$ ,  $v_3$  and  $v_5$ ) based on the expressions and the identifiable roots we discussed in this section.

### 3.2 Degree of identifiability of different enzyme kinetic rate laws

In Figure 3 and Supplementary Figure S5, we show the number and percentage of experimental data combinations that are capable of identifying each parameter in each of the four fluxes  $v_1$  (Figure 3a and e),  $v_2$  (Figure 3b and f),  $v_3$  (Figure 3c and g) and  $v_5$  (Figure 3d and h) belonging to the network shown in Figure S16. Based on the definition given in Section 2.4, the percentages refer to the degree of identifiability of each parameter. We see from Figure 3 and Supplementary Figure S5 that the degree of identifiability of the maximum reaction rates ( $V_i^{\max}$ ), whenever they are identifiable for each of the four fluxes, is greater than or equal to the degree of identifiability of the corresponding enzyme binding ( $K_i$ ) constants, or the allosteric activation constant ( $K_3^{pep}$ ), in the respective reaction rate law models. 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. The fact that in silico experimental data is generated using different kinetic rate law models for  $v_3$  (MWC in Figure 3a-d vs Convenience kinetics in Figure 3e-h) has no bearing on this observation. We accordingly surmise 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.

When data based on the MWC model for  $v_3$  is used for identifying parameters, less than 50% of all data combinations can identify  $v_1$  (47.5% and 40.0% in Figure 3a). Even though the degree of identifiability of  $K_1^{ac(ne)}$  is significantly different when data derived from the Convenience kinetics model for  $v_3$  is used for identification (25% in Figures 3e), the degree of identifiability is still less than 50% for both parameters. In contrast, all available data combinations (from the MWC model and the Convenience kinetics model) satisfy the requirements for identifying  $v_2$  (100% in Figures 3b and f).

The difference in the degrees of identifiability between  $v_1$  (Figure 3a) and  $v_2$  (Figure 3b) can be attributed to the ability of data from different combinations of experiments to satisfy the conditions determined for practical identifiability of that parameter. To elaborate, we showed in Equation (6) and Section 3.1 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. So, in this instance, recall that the ability of a combination of experiments to be useful is determined by the ability of the input change (changes in the input acetate concentration) to effect a change in the measured value of  $v_1$ . Similarly, the identification of parameters for  $v_2$  (Figure 3b) requires the experiments to distinguish between values of both  $v_2$  as well as  $pep$ . We elaborate on the type and nature of experiments that are useful and informative for parameter identification in  $v_1$  and  $v_2$  later in Section 3.4.

The difference in the degree of identifiability of  $v_1$  that comes about from the use of experimental data from different rate laws used to model  $v_3$  (Figures 3a and e) can be again attributed to the usefulness of experimental data. However, the difference in data usefulness between similar experiments arising due to differences in the kinetic rate law used for  $v_3$  could be due to the difference in the dynamics of the network represented by either of the two models. The interaction between different components of the network (Supplementary Figure S16) as exemplified by the different models plays an important role in determining the dynamic and the steady state response of the network. The difference in dynamics in turn contributes to different experimentally observed values of metabolite concentrations and fluxes (Supplementary Figure S1), which in turn changes the usefulness of the given combination of datasets, and its subsequent effect of parameter identifiability.

Nonetheless, we would like to note that these variations in the dynamics contributed by the difference between the rate laws can be resolved with more information, i.e., experimental proteomic data that provide protein concentrations for different experiments. We earlier saw that relaxing the assumption on the available experiment data, i.e., assuming the availability of protein concentration ( $E$ ), in addition to the metabolite concentrations ( $pep$  and  $fdp$ ) and fluxes, results in removal of uncertainty and global identifiability of  $v_1$  (Supplementary Figure S2). Recall that any predicted parameter value for both  $k_1^{cat}$  and  $K_1^{ac}$  matches with the exact known value irrespective of the model (MWC vs Convenience kinetics) from which experimental

data is obtained when enzyme concentrations are included for identifiability (Supplementary Figure S2a). Additionally, we also see an improvement in the degree of identifiability of the corresponding parameters  $k_1^{cat}$  and  $K_1^{ac}$  (Supplementary Figure S5a and b) when enzyme concentrations are available.

We would also like to point out that a difference in either the ability to estimate true parameter values, or the the identifiability of parameter modeling  $v_2$  does not arise due to difference in the kinetic models used to obtain in silico experimental data. This could be because of the fact that any and all available information that is required for the identification of parameters in  $v_2$  is encompassed in the steady state data used for identifiability analysis.

However, unlike  $v_2$ , we see that the degree of identifiability of both locally identifiable roots of parameters in  $v_3$  changes with the change in the data (MWC vs Convenience kinetics) used for the analysis (Figures 3c and g and Supplementary Figures S5b and e). Despite the difference in the dynamics represented by the different data sets, and its impact on the usefulness of data sets for identifiability purposes that was explained earlier, we could also allude to the structural non-identifiability between the parameters in  $v_3$  for the impact that it would have on the practical identifiability of the parameters using in silico data from different sources. In order to ascertain this, we plot the degrees of identifiability of the parameters  $V_3^{max}$ ,  $K_3^{fdp}$  and  $K_3^{pep}$  under conditions when they are globally identifiable in Supplementary Figure S6. To recall from Section 3.1,  $V_3^{max}$  and  $K_3^{fdp}$  are globally identifiable when  $K_3^{pep}$  is constant, and  $V_3^{max}$  and  $K_3^{pep}$  are globally identifiable when  $K_3^{fdp}$  is known and fixed. Now that structurally non-identifiability between parameters has been alleviated, we see that the uncertainty introduced by the data obtained from the MWC model still results in a degree of identifiability of only about 5% for  $K_3^{fdp}$  (Supplementary Figure S6a). However, when  $K_3^{fdp}$  is held constant all MWC model-based experimental datasets contribute to local practical identifiability of parameters  $V_3^{max}$  and  $K_3^{pep}$  (Supplementary Figure S6c), albeit with a bot of uncertainty (Supplementary Figures S6b). This ties back with our earlier observation that these parameters are not globally practically identifiable from data sets obtained using the MWC model (Supplementary Figure S4a and b). However, we see from Supplementary Figures S6c and d that the globally structurally identifiable parameters are also practically identifiable based on the data obtained from the Convenience kinetics model. This can be corroborated by the parameter values shown in Supplementary Figure S4c and d, where data from the



Convenience kinetics model is used for analysis.

We did not observe any differences in the degree of identifiability of parameters in  $v_5$  between the analyses that used data from either the MWC model or the Convenience kinetics model for  $v_3$  (Figure 3d and h). The identifiability of parameters based on the first root obtained from the computer algebra system are shown in Supplementary Figure S5c and f. We use a similar reasoning to that we suggested for identifiability of  $v_2$  above to justify our results.

### 3.3 Rationale for a priori experimental design through practical parameter identification



(should this section moved to SI?) Although the need for better parameter estimation methods is well researched, within the kinetic modeling community (include references related to Banga papers on global optimization methods for parameter estimation, Rudiyanto Gunawan, Lennart Ljung and Heijnen papers here), the need for experimental design methods to satisfy the data needs for parameter estimation has not gained enough traction (include Rudiyanto Gunawan papers, D-optimal design using FIM here). 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 flux  $v_1$  (Figure 4a),  $v_2$  (Figure 4b),  $v_3$  (Figure 4c) and  $v_5$  (Figure 4d). We express the usefulness of data from any given combination of experiments on the basis of the number of parameters that the data from the combination can identify. The higher the number of parameters a combination of experiments can identify, the greater is the usefulness of that combination of experiments. The percentage of all data combinations that can estimate a given number of parameters within a given flux is also shown in Figure 4.

For  $v_1$  (Figure 4a), data from about 50% of the 240 combinations of experiments cannot identify any parameter used to model the flux ( $V_1^{max}$  and  $K_1^{ac}(ne)$ ). In contrast, data from only about 5% of the experiments is useless towards identifying any parameter in  $v_3$ , while there are no useless combinations of experiments from which data cannot be used to identify either  $v_2$ , or  $v_5$ . These data are also in agreement with

the degree of identifiability observed for parameters in all these fluxes (Figure 3 and Supplementary Figure S5). Thus, when it is possible to obtain steady state data from numerous experiments to estimate parameters for a small metabolic network, there is a very high possibility of performing experiments that could provide non-informative data that can result in parameter non-identifiability without a priori experimental design.

We see that in instances where the identifiability of the fluxes differ based on the data set used (data from MWC model vs data from Convenience kinetics model), the utility of the data sets towards identifying parameters are changes. For instance, when data from the MWC model is used to identify parameters in  $v_3$  (Figure 4c), less than 10% of the data sets can identify all three parameters. Whereas, when data from the Convenience kinetics model is used for the same purpose, slightly more than 50% of the combinations can identify all three parameters. Note that this observations should be considered in light of the structural local identifiability and the practical non-identifiability of these parameters that we mentioned earlier in Sections 3.1 and 3.2.

Although one could argue based on earlier observations that uncertainty in the model from which data is obtained, and the structural non-identifiability of some of the parameters in the models of the aforementioned fluxes that there may not be a need for experimental design if in vivo model uncertainty and structural non-identifiability can be resolved, one  is to only look at Supplementary Figure S7a and d. In these figures we show the utility of multiple combinations of metabolite and protein concentration, and flux data towards identifying  $k_1^{cat}$  and  $K_1^{ac}$  in  $v_1$  when data from the MWC model and Convenience kinetics model are used, respectively. We see that about 45% of the 240 combinations of data from 21 different experiments cannot identify either one of the two parameters, when these parameters are globally structurally and practically  identifiable.

In addition to providing information on the fraction of experiments that are informative and those that are not informative, Figure 4 and Supplementary Figure S7 also informs the ease of identifying parameters for a single flux. In conjunction with the degrees of identifiability shown in Figure 3 and Supplementary Figure S5, we see that identifying  $v_2$  is much easier than identifying  $v_1$ . Since minimization of resource utilization requires us to use the least number of experiments to identify both  $v_1$  and  $v_2$ , 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  $v_2$ . A similar argument can also be for determining experiments that can identify  $V_e^{max}$  and  $K_e^{fdp}$  in  $v_5$ : since  $V_e^{max}$  has an identifiability of 100%, the choice of experiments to completely identify parameters in  $v_5$  rests with experiments that can identify  $K_e^{fdp}$ .

Accordingly, if we can establish a hierarchy of identifiable parameters and the corresponding fluxes they model, this section gives credence to the idea that a priori experimental design is necessary, even when all possible in vivo measurements can be obtained, to get the most informative data for parameter identification and estimation at minimum cost. Additionally, having the ability to see the contribution of each type of experiment towards identifiability would prove sufficient to decide on the total number and type of experiments that need to be performed to identify all fluxes in a kinetic model of metabolism. So, next we look at the contribution of different types of experiments towards identifying parameters in each flux in the small network.

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

In the previous sections, we showed how ability to structurally identify parameters in a kinetic model of metabolism is dependent on the kinetic rate law used to model a specific flux (Section 3.1). We also discussed how the the informativeness of experiments used to practically identify parameters in kinetic models is not only dependent on the in silico rate law, but is also a function of the uncertainty in the experimental data brought about by the differences between the in silico rate law and the in vivo rate law from which data is obtained (Section 3.2). Subsequently, we also showed the need for a priori experimental design to generate informative data to identify parameters in kinetic models.

In systems identification terminology, all the aforementioned observations regarding data requirements for parameters identification 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.

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 or transcriptional activators and inhibitors (e.g., effect of *pep* on  $v_3$ , or the effect of *fdp* on  $v_5$  and consequently on  $v_1$  in Figure S16). 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 (Heijnen AND Verheijen 2013). So, in this 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.

We demonstrated in section 3.1, using  $v_1$  and  $v_2$  as examples, that the ability of data from a given combination of experiments to identify parameters in a given flux depends on its ability to satisfy the conditions determined for practical identifiability of that parameter. Although we were able to determine the requirements for identifying  $v_1$  and  $v_2$  based on the closed form expressions of the parameters, this may not be possible when a complex nonlinear rate law is used to model a flux. Using the Convenience Kinetic rate law to model  $v_3$  is a case in point. So, in order to determine the type of experiments that help satisfy identifiability conditions for parameters of any flux modeled by any usable rate law, we look at the distribution of the different types of experiments in all data combinations that can identify the parameters within a flux. Recall that we use experimental data from five different types of experiments to test the practical identifiability of parameters in the model (Section S3.2 and Table S1).

The contribution from different types of experiments towards identification of parameters in different fluxes is given in Figure 5. In Figure 5a and 5b, the contribution from different experiment types for identifying  $V_1^{max}$  and  $K_1^{ac}(ne)$  are given, respectively. The contribution of different experiments towards identifiability for  $V_2^{max}$  (Figure 5c) and  $K_2^{pep}$  (Figure 5d) are also shown. Note that the contributions are only shown for data sets obtained using the MWC model in Figure 5. Experiment type contributions towards identifiability based on data sets obtained from the Convenience kinetics model for  $v_3$  are shown separately in Supplementary Figure .

In agreement with what we discussed in Section 3.1, i.e., experiments must be able distinguish between

different values of  $v_1$  and acetate, 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 (Figure 5a and b). Since less than 50% of all data combinations can satisfy these requirements, and can consequently identify  $v_1$  (Figure 3a), we say that identifiability analysis is crucial to determine the minimum number of experiments, as well as the nature of experiments that can help identify parameters for  $v_1$ .

In contrast to  $v_1$ , we see that both the enzyme perturbation and the acetate perturbation experiments have the same contribution towards combinations that can identify parameters for  $v_2$  (Figure 5c and d). Hence, in conjunction with the high degree of identifiability seen in Figure 3b 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, this confirms our earlier hypothesis that in comparison to selecting experiments to identify  $v_1$ , there is very little restriction on the types of experiments that are informative enough to identify  $v_2$ .

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 (6) 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 S16) 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  $v_2$  (Figure 3b), i.e., 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 with respect to the inputs and the outputs of the metabolic 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 (Supplementary Figure S11). In this case, the top three panels (Supplementary Figure S11a-c) show the distribution of different experiment types for identifying each parameter with respect to their first root identified as  $V_3^{max}(1)$ ,  $K_3^{fdp}(1)$  and  $K_3^{pep}(1)$ . The bottom panels (Supplementary Figure S11d-f) show the corresponding distribution of experiments for the second root of each of three parameters in  $v_3$ :  $V_3^{max}(2)$ ,  $K_3^{fdp}(2)$  and  $K_3^{pep}(2)$ .

In all of the above scenarios for  $v_1$  (Figure 5a and b),  $v_2$  (Figure 5c and d) and  $v_3$  (Supplementary Figure S11), 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 and blue surfaces are superimposed on top of each other. This is also seen for experiments identifying  $V_e^{max}$  in  $v_5$  (Figure 5e). However, this is not the case with the distribution of experiments required for identifying  $K_e^{fdp}$  in  $v_5$  (Figure 5f). 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 (Figure 3d), the choice and number of experiments required to identify  $v_5$  completely hinges upon identifiability of  $K_e^{fdp}$  from the chosen experiments.

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

(Should this section be moved to the SI as well?) 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 S3.2 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 S16). 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 ??). 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 ??) help in determining not only the nature of experiments required for parameter identifiability (Figure ??), 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, there are 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

- Andreozi, 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.
- Andreozi, 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.
- 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.
- 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.

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

653 Heijnen, J. J. (2005) Approximative kinetic formats used in metabolic network modeling, *Biotechnol. Bioeng.*  
654 91.5, 534–545.

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

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

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

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

663 Ljung, L. AND T. Glad (1994) On global identifiability for arbitrary model parametrizations, *Automatica*  
664 30.2, 265–276.

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

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

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

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

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

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

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

679 — (2017) Model-based design of bistable cell factories for metabolic engineering, *Bioinformatics*.

680 Vanlier, J., C. A. Tiemann, ET AL. (2012) A Bayesian approach to targeted experiment design, *Bioinform-*  
681 *atics* 28.8, 1136–1142.

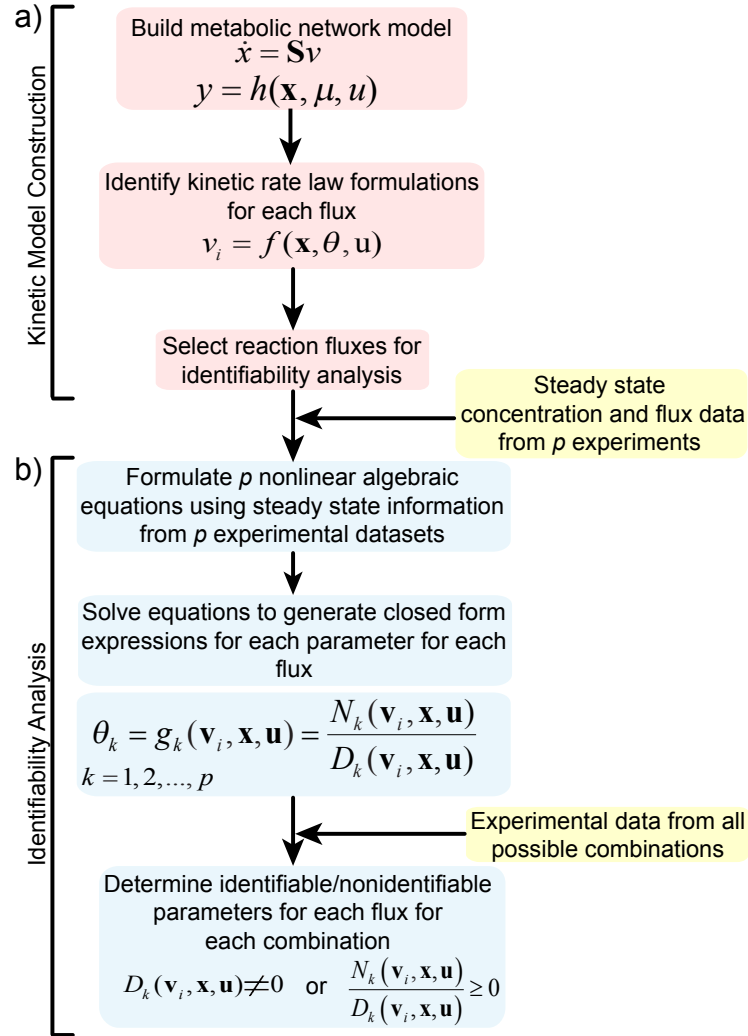
682 Vanlier, J., C. Tiemann, ET AL. (2013) Parameter uncertainty in biochemical models described by ordinary  
683 differential equations, *Math. Biosci.* 246.2, 305–314.

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

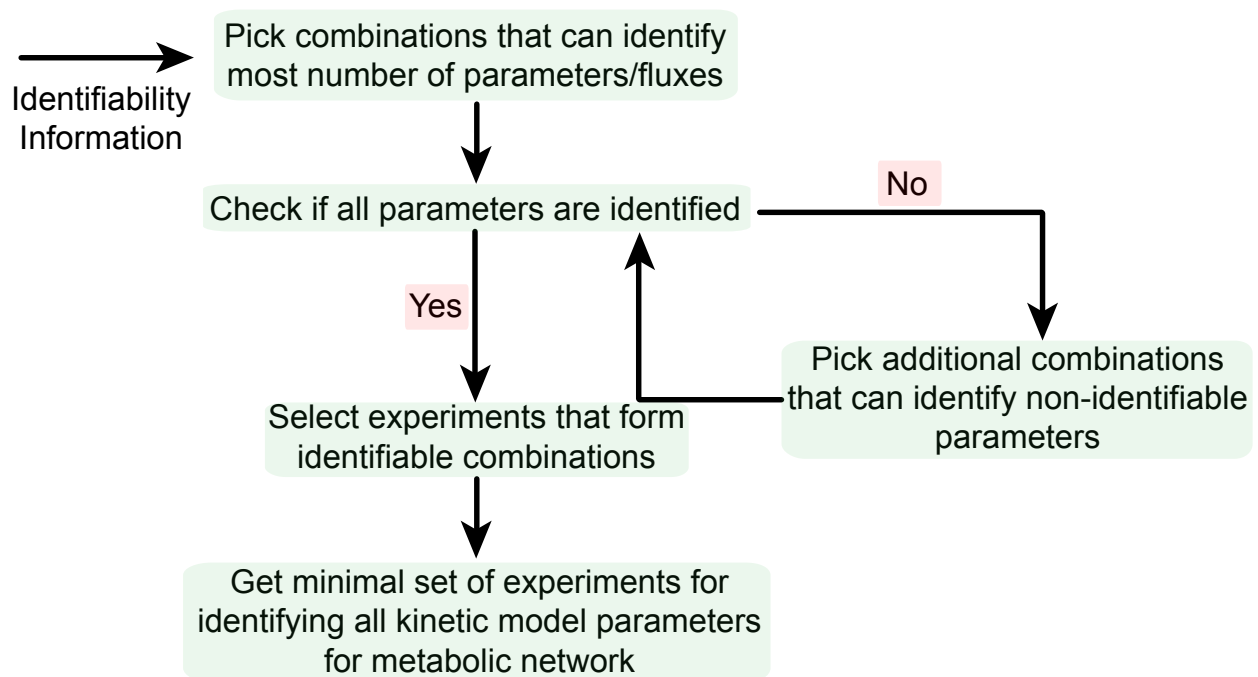
686 Vasilakou, E., ET AL. (2016) *Current state and challenges for dynamic metabolic modeling*.

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

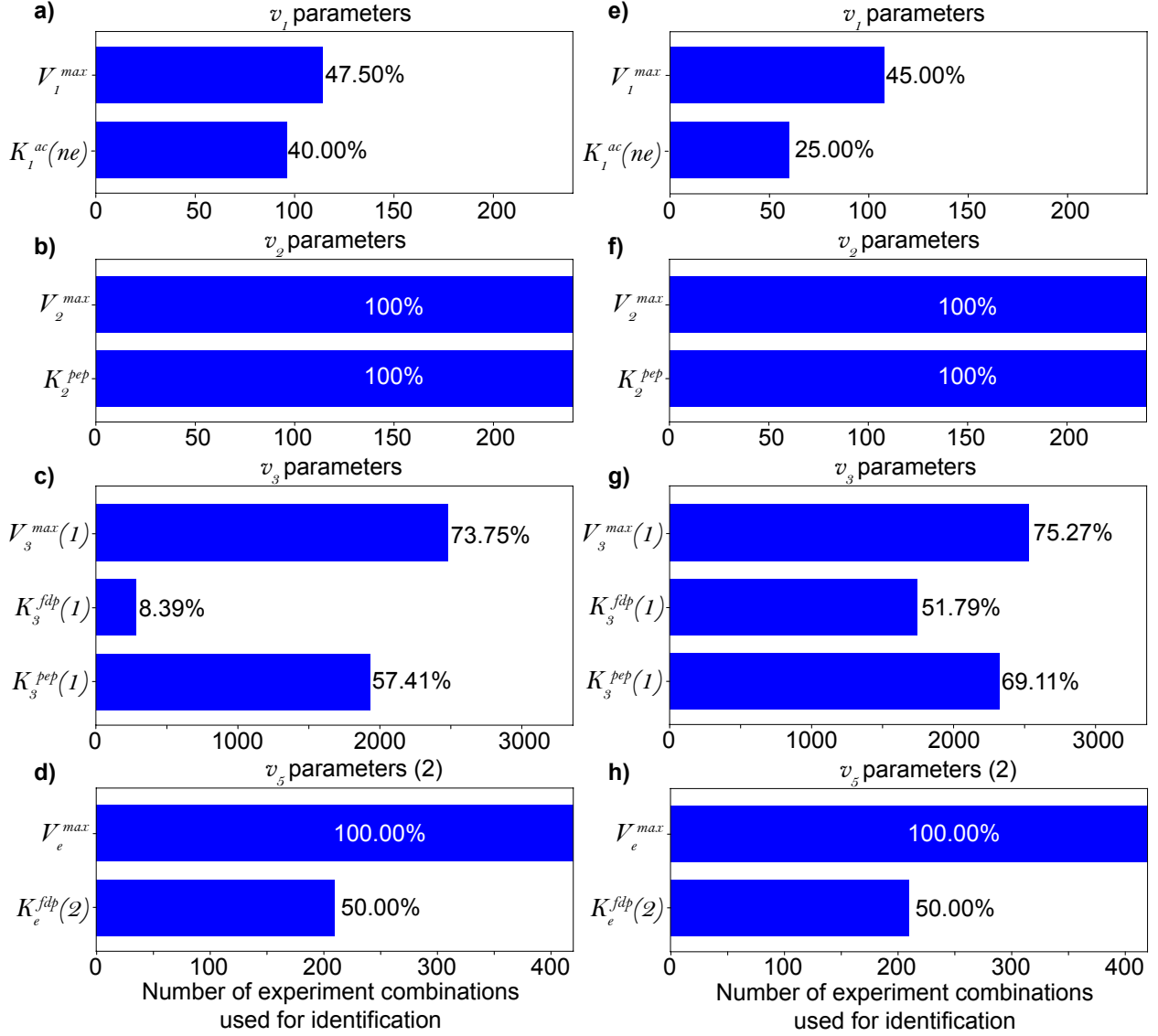
## 5 Tables, Figures and Figure Captions:



**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.

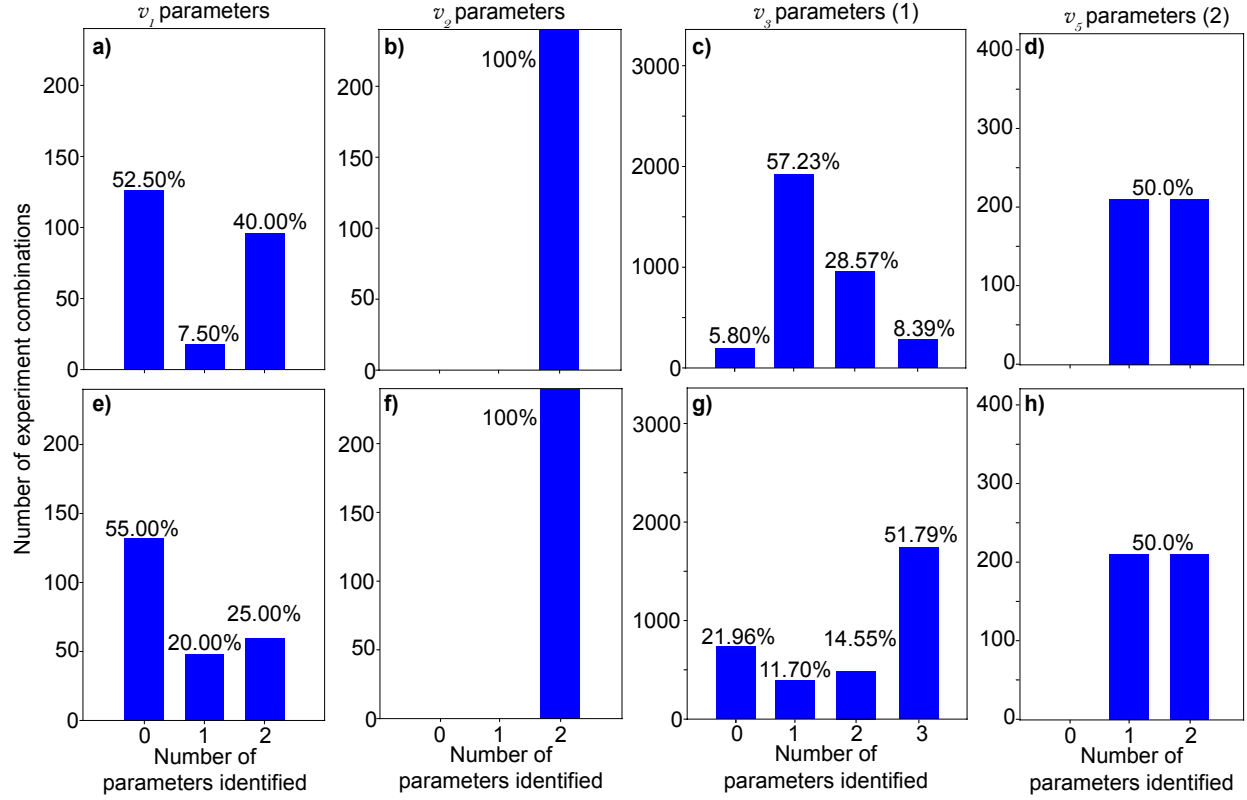


**Figure 2.** Flow diagram showing a method for experimental design that uses our methodology for practical identification of parameters to determine the number and type of experiments required to identify all fluxes within a given metabolic network.

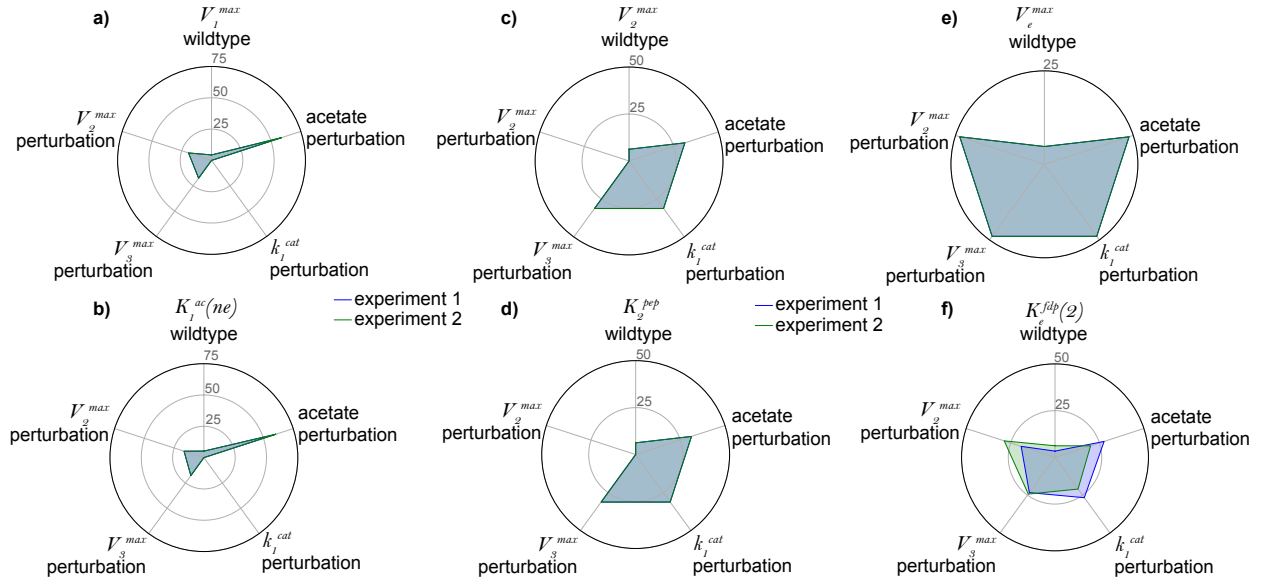


**Figure 3.** The number of data combination from 21 different in silico experiments that can practically identify each parameter in fluxes a) and e)  $v_1$ , b) and f)  $v_2$ , c) and g)  $v_3$ , and d) and h)  $v_5$  when there is no noise in the input experimental data. The percentage of total combinations of experimental data used for analysis (240 for  $v_1$  and  $v_2$ , 3360 for  $v_3$  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, and  $v_3$  requires data from three experiments. Results for only one of the two possible roots are shown for  $v_3$  and  $K_e^{fdp}$  in  $v_5$ . The results obtained using data from the MWC model for  $v_3$  are presented in the panels on the left hand side column and the panels on the right hand column show results obtained using data derived from the Convenience kinetics model for  $v_3$ .





**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)  $v_2$ , c)  $v_3$  and d)  $v_5$ . The total number of combinations of experimental data is shown in the vertical axis and the horizontal axis shows the total possible number of parameters that can be identified by data from combinations of a) two, b) two, c) three and d) two experiments. 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 240 data combinations are used for identifiability analysis for a)  $v_1$  and b)  $v_2$ , 3360 combinations are used to practically identify c)  $v_3$ , and 420 combinations are used to analyze the identifiability of d)  $v_5$ . Section S3.2 provides more details on how the combinations of experimental data are generated.



**Figure 5.** The contribution of different experiments types used in a combination of two experiments ( $j = \{1, 2\}$ ) that can practically identify parameter a)  $V_1^{max}$ , b)  $K_1^{ac}$ , c)  $V_2^{max}$ , d)  $K_2^{pep}$ , e)  $V_e^{max}$  and f)  $K_e^{fdp}(2)$ . The percentages reflect the fractional contribution of each experiment type towards all identifiable data combinations.

## 690 **6 Old Figures, Figure Captions and other miscellaneous text**

691 Methods and tools for structural identification of parameters based on differential algebra (Ljung AND Glad  
692 1994; Audoly, ET AL. 2001; Bellu, ET AL. 2007) and profile likelihood (Raue, ET AL. 2014) are available.