1 Introduction:

1.1 Identifiability analysis: Definitions and Formulations

Any nonlinear dynamical system can be represented by a set of states $\mathbf{x}$, observables $\mathbf{y}$ that are dependent on the states, parameters μ, and inputs u as in Equation (1).

$$\dot{\mathbf{x}} = g(\mathbf{x}, \mu, u) \tag{1a}$$

$$\mathbf{y} = h(\mathbf{x}, \mu, u) \tag{1b}$$

Identifiability concerns with the ability to determine a unique solution to the problem of estimating parameters μ from given data on the system observables $\mathbf{y}$ for inputs u (McLean AND McAuley 2012). The identifiability of parameters in nonlinear models of physical processes can be classified into two categories: structural and practical identifiability.

Any system (Equation 1) is said to be structurally or a priori identifiable if, for an input-output mapping defined by $\mathbf{y} = \Phi(\mu, u)$ for at least one input function u, any two values of parameters μ_1 and μ_2 satisfy the relationship in Equation (2) below.

$$\Phi(\mu_1, u) = \Phi(\mu_2, u) \iff \mu_1 = \mu_2 \tag{2}$$

Accordingly, any system that has an infinite number of solutions to the parameter estimation problem for all input functions is said to structurally non-identifiable. 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.

When parameters are estimated on the basis of noisy data, the ability to estimate unique parameter values to satisfy Equation (2) is referred to as practical identifiability. 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. 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 Methods:

We use a profile likelihood-based approach (Raue, ET AL. 2009) to establish structural and practical identifiability of parameters in nonlinear kinetic models of metabolism. Briefly, the approach seeks to establish the existence/non-existence of bounds in confidence intervals for the estimates of parameters in nonlinear models. The profile likelihood is calculated based on Equation (3) for each parameter θ_i where $\chi^2(\theta_i)$ is given by Equation (4).

$$\chi^2_{PL}(\theta_i) = \min_{\theta_{j \neq i}} \left[\chi^2(\theta)\right] \tag{3}$$

$$\chi^2(\theta) = \sum_{k=1}^{m} \sum_{l=1}^{d} \left(\frac{y^*_{kl} - y_{kl}}{\sigma^*_{kl}} \right)^2 \tag{4}$$

In the minimization objective shown in Equation (4) for parameter estimation, y^*_{kl} is the available experimental time course data for each observable state k at each l time point. The difference between the data and the model estimates at these time points, y_{kl} is weighted by the variance in the experimental data σ^*_{kl}. An algorithm to calculate the profile likelihood, $\chi^2_{PL}(\theta_i)$, based on Equation 3 is given below.

The identifiability of parameters is established through the confidence intervals of their estimates, $\left[\sigma_i^-, \sigma_i^+\right]$. The likelihood-based confidence interval for any parameter whose profile likelihood is estimated can be written on the basis of a threshold Δ_α in the likelihood as in Equation (5).

$$\{\theta | \chi^2(\theta) - \chi^2(\hat{\theta}) < \Delta_\alpha\} \tag{5}$$

The threshold Δ_α in the likelihood is the 1-α quantile of the χ^2 distribution, represented as $\chi^2(\alpha, df)$. The confidence intervals obtained hold for df degrees of freedom. For a choice of df=1 the confidence intervals will hold for each parameter individually, and confidence intervals that hold jointly for all parameters can be obtained by choosing the number of parameters as df.

The visualization of structurally and practically non-identifiable parameters using the profile likelihood approach is illustrated in Figure 1. The points of intersection between the profile likelihood curves (solid line) with the one parameter likelihood threshold ($\Delta_\alpha = \chi^2(\alpha, 1)$, dashed line) provide the confidence intervals of the parameter θ_i. The confidence intervals of a structurally non-identifiable parameter are unbounded, i.e., $[-\infty, +\infty]$ (Figure 1a), while the confidence intervals of a practically non-identifiable parameter are unbounded in at least one direction, i.e., $\left[\sigma_i^-, \sigma_i^+\right]$ where either $\sigma_i^- = -\infty$ or $\sigma_i^+ = +\infty$ (Figure 1b). If a parameter's estimates have a finite confidence interval then the parameter is said to be identifiable (Figure 1c). Note that the horizontal dotted lines in Figure 1 represent the confidence interval thresholds (Δ_α) that are used to establish identifiability.

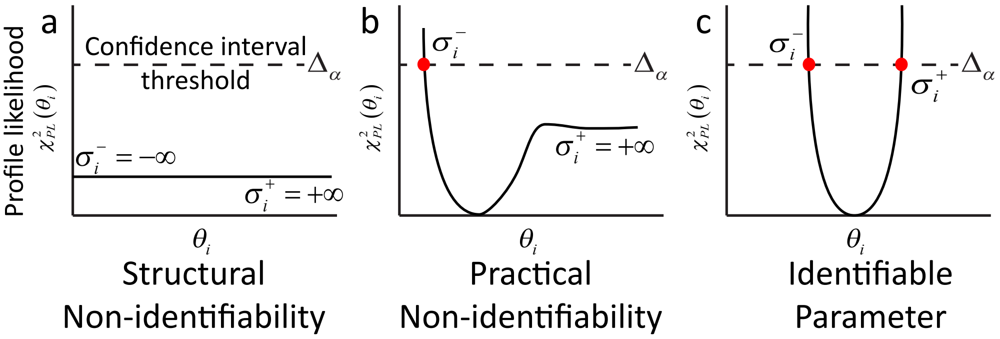

Figure 1: The profile likelihood estimates of a) a structurally non-identifiable, b) a practically non-identifiable and c) an identifiable parameter. The horizontal axis represents the changes in the value of the parameter (θ_i) whose identifiability is being determined and the profile likelihood ($\chi^2_{PL}(\theta_i)$) is shown in the vertical axis. The confidence interval threshold (Δ_α) used to determine the identifiability of the parameter is denoted by the horizontal dotted line. Identifiable parameters are distinguished from non-identifiable parameters by the presence of both upper and lower bounds on their confidence interval estimates $\left[\sigma_i^-, \sigma_i^+\right]$.

2.1 Algorithm to calculate profile likelihood:

1. Start by solving the numerical optimization problem in Equation 3 for a fixed value of θ_i to determine the initial point on the profile likelihood curve.

2. Take an increasing/decreasing step θ_{step} in the direction of θ_i.

The increasing/decreasing steps in θ_i can be adapted to the shape of the profile likelihood curve that is encountered while running the algorithm. Accordingly, the step θ_{step} should satisfy Equation (6) below, where θ^{k-1} refers to parameter estimates obtained from iteration $k-1$ of the above algorithm and $q \in [0, 1]$.

$$\chi^2\left(\theta^{k-1} + \theta_{step}\right) - \chi^2(\theta^{k-1}) \approx q.\Delta_\alpha \tag{6}$$

3. Re-optimize $\theta_{j \neq i}$ using Equation 3.
4. Repeat the last two steps until a desired threshold Δ_α is exceeded or a maximal amount of steps is reached.

Due to the dependence of practical parameter identifiability on the experimental data, the profile likelihood approach can be used to design experiments in such a way that the observables that are derived from these experiments can improve the practical identifiability of the parameters. We show how experimental design can have a meaningful impact on parameter identification and estimation in Figure 2. Assuming a parameter θ_i is practically non-identifiable (Figure 2a), performing a profile-likelihood based identifiability analysis using simulated data can help determine the nature of experiments needed to make the parameter identifiable (Figure 2b). In contrast, performing non-informative experiments without prior knowledge on their ability to change the identifiability of the parameter may provide data that cannot be used to estimate parameter θ_i (Figure 2c).

2.2 A method to establish posterior identifiability of metabolic network models:

This section details a method to establish the practical (posterior) identifiability of metabolic network models using the algebraic relationship between fluxes. Every flux, v, in a kinetic model of a metabolic network can be expressed as a nonlinear algebraic equation (Equation 7). The fluxes are expressed as a function of the metabolite concentrations x and the kinetic parameters θ in Equation (7).

$$v = f(\mathbf{x}, \theta) \tag{7}$$

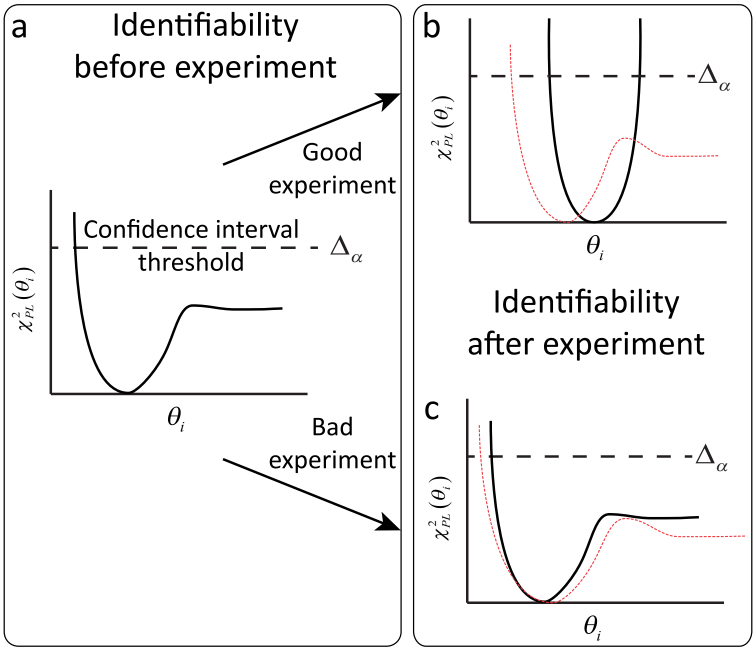
a
Identifiability before experiment
Good experiment
Bad experiment
Confidence interval threshold
Δ_α
$\chi^2_{PL}(\theta_i)$
θ_i
b
c
Identifiability after experiment

Figure 2: Cartoon illustrating the utility of identifiability analysis for experimental design. a) The profile likelihood of a practically non-identifiable parameter that needs to be estimated based on both good and bad experimental data. The changes in the profile likelihood of the parameter when estimated with b) good experimental data and c) bad experimental. The practical identifiability of a parameter is dependent on the quality and quantity of experimental data used to estimate the parameter. The identifiability of the parameter, as determined using its profile likelihood, does not change due to the poor quality/quantity of experiments.

Given the nonlinear nature of this model, the function f in Equation (7) can expressed, without loss of generality as,

$$v = \frac{N(\mathbf{x}, \theta)}{D(\mathbf{x}, \theta)} \tag{8}$$

where $N(\mathbf{x}, \theta)$ is the numerator of f, and $D(\mathbf{x}, \theta)$ is the denominator of f.

If $\theta \in \mathbb{R}^p$, given a set of experimental measurements for the metabolite concentrations $\mathbf{x}$ and the reaction fluxes $\mathbf{v}$, theoretically, it is possible to choose p sets of data from these measurements to solve for the p parameters in θ. However, if any these datasets do not satisfy the condition that $D(\mathbf{x}, \theta) \neq 0$, then the number of experiments required to estimate the p parameters in θ can be established to be greater than p. An example is shown below.

This analysis can be performed for each flux in a metabolic network independent of all the other fluxes. This enables this method to be scalable to even genome-scale models. The following section demonstrates this methodology for one of the fluxes in the gluconeogenic model of Kotte et al., (**Kotte2014**).

2.3 Identifiability analysis of parameters in a kinetic model of gluconeogenesis:

The proposed model for acetate consumption through gluconeogenesis and its corresponding kinetic model is used as a case study to illustrate the utility of identifiability analysis for the design of experiments for estimating parameters in kinetic models of metabolism. The kinetic model is described below.

$$\frac{d}{dt} pep = v_1 - v_2 - v_4 \tag{9}$$

$$\frac{d}{dt} fdp = v_2 - v_3 \tag{10}$$

$$\frac{d}{dt} E = v_{e,max} \left(\frac{1}{1 + \left(\frac{fdp}{K_e^{fdp}} \right)^{n_e}} \right) - dE \tag{11}$$

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 (12) and (13) respectively using Michaelis-Menten kinetics. The acetate flux through v_1 is also governed by the quantity of available enzyme E.

$$v_1 = k_1^{cat} E \frac{acetate}{acetate + K_1^{acetate}} \tag{12}$$

$$v_2 = V_2^{max} \frac{pep}{pep + K_2^{pep}} \tag{13}$$

$$v_3 = V_3^{max} \frac{\tilde{fdp}\left(1+\tilde{fdp}\right)^3}{\left(1+\tilde{fdp}\right)^4 + L_3\left(1+\frac{pep}{K_3^{pep}}\right)^{-4}} \tag{14}$$

The allosterically regulated flux v_3 for the consumption of *fdp* is expressed in Equation (14) using the Monod-Wyman-Changeux (MWC) model for allosterically regulated enzymes, where $\tilde{fdp}$ refers to the ratio of *fdp* with respect to its allosteric binding constant K_3^{fdp}. The added flux v_4 for the export of *pep* is expressed as a linear equation dependent on *pep* in Equation (15).

$$v_4 = k_4^{cat}.pep \tag{15}$$

We use flux v_2 to demonstrate the identifiability analysis method described in the previous section. Flux v_2 has two parameters, V_2^{max} and K_2^{pep} that need to be estimated from experimental data. Here, we assume that at least two different sets of experimental data for the concentrations and fluxes are available. Accordingly, we label these dataset as pep^1, ${v_2}^1$ and pep^2, ${v_2}^2$ respectively. Subsequently, these experimental datasets can be included in the model to form two simultaneous nonlinear algebraic equations in the parameters V_2^{max} and K_2^{pep} (Equation 16).

$$V_2^{max} = \frac{v_2^1 v_2^2 (pep^1 - pep^2)}{v_2^2 pep^1 - v_2^1 pep^2} \tag{16a}$$

$$K_2^{pep} = \frac{pep^1 (v_2^1 pep^2 - v_2^2 pep^2)}{v_2^2 pep^1 - v_2^1 pep^2} \tag{16b}$$

2.4 Parameter estimation in kinetic models of metabolism:

The formulation for parameter estimation assumes that the concentrations and the fluxes corresponding to each perturbation are variables in the optimization problem formulated to estimate kinetic parameters.

Let $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{p} \in \mathbb{R}^l$ and $\mathbf{v} \in \mathbb{R}$ be the vector of concentrations, parameters and fluxes respectively. If the number of perturbations used for parameter estimation is in $\mathbb{R}^p$, then the vector of concentrations, fluxes and parameters that need to be estimated in the optimization problem changes to $\mathbf{x} \in \mathbb{R}^{mp}, \mathbf{v} \in \mathbb{R}^p$ and $\mathbf{p} \in \mathbb{R}^l$ respectively.

Let $i \in \mathbb{R}^{mp}$, $j \in \mathbb{R}^p$ and $k \in \mathbb{R}^l$ represent the indices for concentrations, fluxes and parameters respectively in Equations (18)-(17).

Formulation c): In order to estimate kinetic parameters for fluxes in models of metabolism, we propose a formulation accounting for the uncertainty in the concentration and flux estimates. The uncertainty terms (ϵ) are used as bounds for the flux and the concentrations and are included in the optimization objective as a variable to be penalized is given in Equation (17) below. Note that the concentration and the fluxes have different uncertainties, ϵ_x and ϵ_v respectively, associated with them. w_1, w_2, w_3 and w_4 are the weights associated with the four different terms in the objective.

$$\min_{\mathbf{x},\mathbf{p},\mathbf{v},\epsilon_x,\epsilon_v} \; w_1 \left\|\mathbf{v} - \mathbf{v}^*\right\| + w_2 \left\|\mathbf{x} - \mathbf{x}^*\right\| + w_3\epsilon_x + w_4\epsilon_v \tag{17a}$$

$$\text{st } N(\mathbf{x}, \mathbf{p}) - v_j D(\mathbf{x}, \mathbf{p}) = 0 \qquad \forall\; j \in \mathbb{R}^p \tag{17b}$$

$$x_i^*(1 - \epsilon_x) \leq x_i \leq x_i^*(1 + \epsilon_x) \qquad \forall\; i \in \mathbb{R}^{mp} \tag{17c}$$

$$v_j^*(1 - \epsilon_v) \leq v_j \leq v_j^*(1 + \epsilon_v) \qquad \forall\; j \in \mathbb{R}^p \tag{17d}$$

$$\mathbf{p}_{min} \leq \mathbf{p} \leq \mathbf{p}_{max} \tag{17e}$$

Table 1: Table showing the perturbed values of all fluxes used for parameter estimation.

Designation	Perturbed Fluxes	Perturbed Values
P1	v_1	2
P2	v_2	0.2
P3	v_3	0.5

3 Results:

3.1 Estimation of flux 1:

Table 2: Table showing the ability of different formulations mentioned above to estimate enzyme kinetic parameters for different fluxes based on the perturbations given in the above table.

Flux	Data Type	Formulation (b)		Formulation (c)	
		$\epsilon_x(\%)$	$\epsilon_v(\%)$	$\epsilon_x(\%)$	$\epsilon_v(\%)$
flux 1	no noise	30	5	33.33	0
	noisy	x	x	0	33.33
flux 2	no noise	90	30	0	31
	noisy	85	40	0	30
flux 3	no noise	15	10	17	0
	noisy	x	x	17	0

3.2 Estimation of flux 2:

3.3 Estimation of flux 3:

4 Appendix

4.1 Formulation a):

$$\min_{\mathbf{x},\mathbf{p},\mathbf{v}} \ \|\mathbf{v} - \mathbf{v}^*\| \tag{18a}$$

$$\text{st } N(\mathbf{x},\mathbf{p}) - v_j D(\mathbf{x},\mathbf{p}) = 0 \qquad \forall\ j \in \mathbb{R}^p \tag{18b}$$

$$x_i^*(1-\epsilon) \le x_i \le x_i^*(1+\epsilon) \qquad \forall\ i \in \mathbb{R}^{mp} \tag{18c}$$

$$v_j^*(1-\epsilon) \le v_j \le v_j^*(1+\epsilon) \qquad \forall\ j \in \mathbb{R}^p \tag{18d}$$

$$\mathbf{p}_{min} \le \mathbf{p} \le \mathbf{p}_{max} \tag{18e}$$

4.2 Formulation b):

In this formulation, in Equation (19), the flux and concentrations are constrained as earlier with an uncertainty term.

$$\min_{\mathbf{x},\mathbf{p},\mathbf{v}} \ w_1 \|\mathbf{v} - \mathbf{v}^*\| + w_2 \|\mathbf{x} - \mathbf{x}^*\| \tag{19a}$$

$$\text{st } N(\mathbf{x},\mathbf{p}) - v_j D(\mathbf{x},\mathbf{p}) = 0 \qquad \forall\ j \in \mathbb{R}^p \tag{19b}$$

$$x_i^*(1-\epsilon) \le x_i \le x_i^*(1+\epsilon) \qquad \forall\ i \in \mathbb{R}^{mp} \tag{19c}$$

$$v_j^*(1-\epsilon) \le v_j \le v_j^*(1+\epsilon) \qquad \forall\ j \in \mathbb{R}^p \tag{19d}$$

$$\mathbf{p}_{min} \le \mathbf{p} \le \mathbf{p}_{max} \tag{19e}$$

4.3 Formulation d):

Due to the inability of Equation (18) to get feasible results, we modified the problem to include the L2-norm of the concentrations as well in the objective function in Equation (20a). In Equation (20) the flux and the concentrations are free variables, i.e., $\mathbf{x} \in [\mathbf{0}, +\infty]$.

$$\min_{\mathbf{x},\mathbf{p},\mathbf{v}} w_1 \|\mathbf{v} - \mathbf{v}^*\| + w_2 \|\mathbf{x} - \mathbf{x}^*\| \tag{20a}$$

$$\text{st } N(\mathbf{x}, \mathbf{p}) - v_j D(\mathbf{x}, \mathbf{p}) = 0 \qquad \forall\, j \in \mathbb{R}^p \tag{20b}$$

$$0 \leq x_i \leq +\infty \qquad \forall\, i \in \mathbb{R}^{mp} \tag{20c}$$

$$0 \leq v_j \leq +\infty \qquad \forall\, j \in \mathbb{R}^p \tag{20d}$$

$$\mathbf{p}_{min} \leq \mathbf{p} \leq \mathbf{p}_{max} \tag{20e}$$

After testing the problem in Equation (20) above for the acetate uptake flux v_1, we found that all solutions using IPOPT exceeded the maximum number of iterations (set at 5000) and never produced an optimal solution. Hence, we use a modified form of the bounds in Equation (20) above in Equation (19).

References

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.

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.