Article

Assessing and Resolving Model Misspecifications in Metabolic Flux Analysis

Rudiyanto Gunawan [1,2,*] and Sandro Hutter [1,2]

[1] Institute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, 8093 Zurich, Switzerland
[2] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
* Correspondence: rudi.gunawan@chem.ethz.ch; Tel.: +41-44-633-2134

Abstract:

Background: Metabolic flux analysis (MFA) is an indispensable tool in metabolic engineering. The simplest variant of MFA relies on an overdetermined stoichiometric model of the cell's metabolism under the pseudo-steady state assumption, to evaluate the intracellular flux distribution. Despite its long history, the issue of model error in the overdetermined MFA, particularly misspecifications of the stoichiometric matrix, has not received much attention.

Method: We evaluated the performance of statistical tests from linear least square regressions, namely Ramsey RESET test, F-test and Lagrange multiplier test, in detecting model misspecifications in the overdetermined MFA, particularly missing reactions. We further proposed an iterative procedure using the F-test to correct such an issue.

Result: Using Chinese hamster ovary and random metabolic networks, we demonstrated that: (1) a statistically significant regression does not guarantee high accuracy of the flux estimates, (2) the removal of a reaction with a low flux magnitude can cause disproportionately large biases in the flux estimates, (3) the F-test could efficiently detect missing reactions, and (4) the proposed iterative procedure could robustly resolve the omission of reactions.

Conclusion: Our work demonstrated that statistical analysis and tests could be used to systematically assess, detect and resolve model misspecifications in the overdetermined MFA.

Keywords: metabolic flux analysis, model misspecification, constraint-based model, stoichiometric model, Chinese hamster ovary cell culture

1. Introduction

The ability of biological systems to produce highly complex molecules at high enantiomeric excess has pushed metabolic engineering that relies on directed alterations of the cell's biochemical reactions through recombinant DNA technology, to the center stage of biotechnology [1,2]. The understanding of cellular metabolism and its manipulation encompass much of the research activities in modern biotechnology. Mathematical modeling of cellular metabolism, particularly constraint-based or stoichiometric modeling, has been playing an important role, not only in the analysis of metabolic phenotypes (metabolic flux distribution), but also in the design and optimization of metabolic pathways to enhance productivity or to synthesize new desired products. The most widely used model-based analysis in metabolic engineering is the metabolic flux analysis (MFA), which comprises methods for determining intracellular metabolic fluxes. MFA employs a stoichiometric model of the metabolic reaction network based on the mole balance equation of the intracellular metabolites under a pseudo-steady state assumption [1,3].

A simple strategy of MFA, from here on referred to as the overdetermined MFA, uses a reduced stoichiometric model of the cell's metabolism such that the estimation of the metabolic fluxes is mathematically well-posed, i.e. the flux estimation involves an (over-)determined system. For larger and more realistic metabolic models, the flux estimation in the MFA often becomes underdetermined as the number of unknown fluxes exceeds the number of balance equations. There has been a flurry of activity in the development of MFA methods for large metabolic models based on linear programming (notably flux balance analysis) [4], which goes hand in hand with the creation of genome-scale metabolic models. While the MFA strategies above are predominantly based on measurements of extracellular concentration of metabolites, a different class of MFA techniques that rely on data from ^{13}C isotopic labeling experiments, emerged and matured over the past two decades. The experimental, analytical and computational procedures for ^{13}C-based MFA have now been standardized [5].

Because of the experimental and/or computational complexity in the application of MFA using genome-scale metabolic models and ^{13}C isotopic labeling experiments, the overdetermined MFA continues to be used in practice, thanks to its simple formulation and numerical implementation [6,7]. A common criticism of the overdetermined MFA is the use of a reduced (incomplete) description of the metabolic network. The accuracy of the flux estimates is often a concern in the application of the overdetermined MFA. Not to mention, the formulation of an appropriate reduced order model for a given set of extracellular species measurements in a particular organism is challenging. Past studies have established several guidelines for good practices in the overdetermined MFA. For example, analytical conditions that guarantee the ability of the stoichiometric model to balance, i.e. the measured rates of extracellular species concentrations can be balanced, and ensure the existence of a unique solution for the intracellular metabolic fluxes, have been formulated [8]. In consideration of data noise (measurement errors), statistical tests on the goodness of fit could be used to assess the consistency between the data and the stoichiometric model [9,10]. The accuracy of the flux estimates could also be quantified by computing the corresponding confidence intervals [11] or by propagating known errors in the measurements to the flux estimates [12]. Finally, the significance of observed changes in the flux distribution between conditions or strains, could and should be established by standard statistical tests (e.g., t-test) [13].

The test on goodness of fit or statistical significance of regression in the overdetermined MFA could fail because of several reasons, including (1) incorrect assumptions on the characteristics of data noise (e.g. on the mean and variance of the noise) and (2) model specification error [14,15]. In such a scenario, the resulting flux estimates of the overdetermined MFA may have large inaccuracy or bias. Procedures for detecting and locating gross measurement errors and missing or incorrectly specified components have previously been proposed based on either the improvement of goodness of fit upon removal of a measured variable [9] or the directionality of the residual vectors [14]. Despite its long history, the assessment, detection and rectification of misspecifications of the stoichiometry matrix in the overdetermined MFA have not received much attention. In a recent study, Sokolenko et al. provided a procedure for detecting model error through *in silico* generation of flux profiles and the common statistical *t*-tests [16].

In this work, we adapted statistical analysis and tools for model misspecifications commonly used in the field of linear least square regression, to address the issue of missing reactions in the overdetermined MFA. We posited that the simplification of the stoichiometric model, either manually or using a numerical algorithm [17], to generate an overdetermined flux estimation problem in the overdetermined MFA may inadvertently remove important metabolic reactions. In this study, we illustrated how an omission of reaction(s) could lead to biases in the flux estimates, and evaluated the performance of several statistical tests including Ramsey's RESET test, F-test and Lagrange multiplier test, to detect such specification errors. Finally, we proposed an iterative procedure based on the F-test to resolve the issue of missing reactions. We demonstrated the ability of the aforementioned model misspecification tests and correction procedure, by applying them to the flux analysis of Chinese hamster ovary metabolism and *in silico* metabolic networks.

2. Materials and Methods

2.1. Metabolic Flux Analysis

The MFA is based on the mole balance equation for the intracellular metabolites under a pseudo steady state assumption, as follows:

$$\frac{d\mathbf{c}}{dt} = \mathbf{Sv} = \mathbf{0} \tag{1}$$

where $\mathbf{c}$ is the vector of m metabolite concentrations, $\mathbf{v}$ denotes for the vector of n metabolic fluxes, and $\mathbf{S}$ denotes the $m\times n$ stoichiometric matrix. The fluxes describe either the rate of reactions that consume or produce the metabolites, or the rate of the transport of metabolites between the cell and the extracellular environment or between different intracellular compartments. In the typical formulation of MFA, some, if not all, of the exchange fluxes – the fluxes of metabolites in and out of the cell – could be estimated from the measurements of extracellular species concentrations. The task at hand is to estimate the unknown internal metabolic fluxes from the exchange fluxes. For such a purpose, we begin with partitioning the stoichiometric matrix $\mathbf{S}$ and the flux vector $\mathbf{v}$ into:

$$\mathbf{Sv} = [\mathbf{S_E} \quad \mathbf{S_I}]\begin{bmatrix}\mathbf{v_E}\\ \mathbf{v_I}\end{bmatrix} = \mathbf{S_E v_E} + \mathbf{S_I v_I} = \mathbf{0} \tag{2}$$

where the subscripts $\mathbf{E}$ and $\mathbf{I}$ refer to the exchange and internal fluxes, respectively. Thus, given the values of exchange fluxes $\mathbf{v_E}$, the estimation of unknown internal fluxes $\mathbf{v_I}$ reduces to solving the following linear equation:

$$\mathbf{S_I v_I} = -\mathbf{S_E v_E} \tag{3}$$

The estimation of the internal metabolic fluxes $\mathbf{v_I}$ as stated in Eq. (3) could be casted as a linear least square regression problem:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{4}$$

where $\mathbf{y}$ denotes the vector of measured response variables, $\mathbf{X}$ denotes the (non-random) design matrix containing the values of explanatory variables, $\boldsymbol{\beta}$ denotes the unknown parameter vector, and $\mathbf{e}$ denotes the vector of measurement errors (noise). The ordinary least square (OLS) estimate of the parameters $\boldsymbol{\beta}$ is given by the minimum of the following error function:

$$\boldsymbol{\Phi}_{\text{OLS}}(\mathbf{b}) = (\mathbf{y} - \mathbf{Xb})^{\text{T}}(\mathbf{y} - \mathbf{Xb}) \tag{5}$$

By invoking the first order necessary condition for optimality ($d\boldsymbol{\Phi}_{\text{OLS}}/d\mathbf{b} = 0$), the OLS parameter estimate is given by:

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{6}$$

According to the Gauss-Markov theorem [18], when the measurement errors are additive and uncorrelated with zero mean and constant variance (i.e. $\text{Cov}(\mathbf{e}) = \sigma^2\mathbf{I}$), $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ gives the minimum variance unbiased estimate (MVUE) of $\boldsymbol{\beta}$ among all linear estimators, with the following variance-covariance matrix:

$$\text{Cov}\left(\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{Cov}(\mathbf{e})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{7}$$

If the measurement errors are correlated and/or have unequal variance with a known variance-covariance matrix $\text{Cov}(\mathbf{e})$, then one could resort to the generalized least square (GLS) formulation by minimizing the following error function

$$\boldsymbol{\Phi}_{\text{GLS}}(\mathbf{b}) = (\mathbf{y} - \mathbf{Xb})^{\text{T}}\text{Cov}(\mathbf{e})^{-1}\,(\mathbf{y} - \mathbf{Xb}) \tag{8}$$

Naturally, the GLS estimation requires the variance-covariance matrix $\text{Cov}(\mathbf{e})$ to be invertible. When $\text{Cov}(\mathbf{e}) = \mathbf{LL}^T$ is invertible, the GLS is equivalent to the OLS regression for the linear problem $\mathbf{L}^{-1}\mathbf{y} = \mathbf{L}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{L}^{-1}\mathbf{e}$. The GLS parameter estimate $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ is therefore given by:

$$\widehat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}^T\text{Cov}(\mathbf{e})^{-1}\mathbf{X})^{-1}\mathbf{X}^T\text{Cov}(\mathbf{e})^{-1}\mathbf{y} \tag{9}$$

Here, $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ is the MVUE of $\boldsymbol{\beta}$ with the following variance-covariance matrix:

$$\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}\right) = (\mathbf{X}^T \mathrm{Cov}(\mathbf{e})^{-1}\mathbf{X})^{-1} \tag{10}$$

By drawing parallels between the linear equation of the MFA in Eq. (3) and the least square regression problem in Eq. (4), we could set $\mathbf{y} = -\mathbf{S}_\mathbf{E}\mathbf{v}_\mathbf{E}$, $\mathbf{X} = \mathbf{S}_\mathbf{I}$, and $\boldsymbol{\beta} = \mathbf{v}_\mathbf{I}$, and use the OLS or GLS formulation, whichever appropriate, to obtain the flux estimate $\hat{\mathbf{v}}_\mathbf{I}$. Note that the existence of an optimal parameter estimate for the OLS (or GLS) requires the matrix $\mathbf{X}^T\mathbf{X}$ (or $\mathbf{X}^T\mathrm{Cov}(\mathbf{e})^{-1}\mathbf{X}$) to be invertible, i.e. the matrix $\mathbf{X}$ needs to have a full column rank. In the context of the overdetermined MFA, the rank condition puts a constraint on the dimension of the matrix $\mathbf{S}_\mathbf{I}$, such that the number of unknown internal fluxes $\mathbf{v}_\mathbf{I}$ should not exceed the number of metabolite species in the mole balance equation. The rank condition also necessitates the reactions to be linearly independent, i.e. each of the reactions could not be written as a linear combination of the other reactions. Upon violation of the above rank condition, for example when the number of reactions exceeds that of the species, the non-zero degrees of freedom in the mole balance equation allow the existence of many solutions (not all physically or biologically feasible), a problem that is addressed by strategies such as the flux balance analysis [4]. In this work, we focused on the MFA with the matrix $\mathbf{S}_\mathbf{I}$ having a full column rank.

2.2. Model Misspecification

Because of the rank condition on $\mathbf{S}_\mathbf{I}$, one often faces the challenge of choosing a small number of reactions – for example from the genome scale metabolic models – to include in the mole balance equation used in the overdetermined MFA. Despite the long history of MFA, the impact of an incorrect stoichiometric matrix specification, particularly due to the omission of reactions, on the accuracy of the flux estimate has not received much attention. Recently, Sokolenko et al. used the GLS framework, statistical *t*-test, and simulated flux values to show that model errors could lead to gross deviations in flux estimates that are not statistically significant [16]. In the field of linear least square regression, the impact of a model misspecification on the parameter estimates is well studied, and several tests have previously been developed to detect model misspecifications. In this study, we evaluated the performance of several of such tests, including Ramsey RESET test, F-test and Lagrange-Multiplier test, in detecting specification errors of the stoichiometric matrix in the overdetermined MFA.

2.2.1. Effects of Missing Reactions

In the following, we considered the problem of missing or omitted reactions in the stoichiometric matrix. We assume that the metabolic network model in the MFA given in Eq. (1) is incomplete, and that the true mole balance is governed by

$$[\mathbf{S}_\mathbf{E} \quad \mathbf{S}_\mathbf{I} \quad \mathbf{S}_\mathbf{O}]\begin{bmatrix}\mathbf{v}_\mathbf{E}\\ \mathbf{v}_\mathbf{I}\\ \mathbf{v}_\mathbf{O}\end{bmatrix} = \mathbf{0} \tag{11}$$

where $\mathbf{S}_\mathbf{O}$ contains the stoichiometric coefficients of the omitted reactions and $\mathbf{v}_\mathbf{O}$ denotes the vector of metabolic fluxes of the omitted reactions. The least square problem of estimating the unknown metabolic fluxes $\mathbf{v}_\mathbf{I}$ and $\mathbf{v}_\mathbf{O}$ from the measurements of the exchange fluxes $\mathbf{v}_\mathbf{E}$ is given by:

$$-\mathbf{S}_\mathbf{E}\mathbf{v}_\mathbf{E} = \mathbf{S}_\mathbf{I}\mathbf{v}_\mathbf{I} + \mathbf{S}_\mathbf{O}\mathbf{v}_\mathbf{O} + \mathbf{u} \tag{12}$$

where $\mathbf{u}$ denotes the vector of measurement errors for the true model. Without loss of generality, in order to illustrate the impact of omitting reactions, we could consider the OLS estimate of the internal fluxes $\mathbf{v}_\mathbf{I}$ using the mole balance in Eq. (2), as follow:

$$\hat{\mathbf{v}}_\mathbf{I} = \left(\mathbf{S}_\mathbf{I}{}^T\mathbf{S}_\mathbf{I}\right)^{-1}\mathbf{S}_\mathbf{I}{}^T(-\mathbf{S}_\mathbf{E}\mathbf{v}_\mathbf{E}) \tag{13}$$

The derivation below could easily be rewritten using the GLS estimation whenever appropriate. Substituting $-\mathbf{S}_\mathbf{E}\mathbf{v}_\mathbf{E}$ from the true model in Eq. (12) to Eq. (13), we obtain

$$\hat{\mathbf{v}}_{\mathbf{I}} = \left(\mathbf{S}_{\mathbf{I}}{}^{T}\mathbf{S}_{\mathbf{I}}\right)^{-1}\mathbf{S}_{\mathbf{I}}{}^{T}(\mathbf{S}_{\mathbf{I}}\mathbf{v}_{\mathbf{I}} + \mathbf{S}_{\mathbf{O}}\mathbf{v}_{\mathbf{O}} + \mathbf{u}) \tag{14}$$

$$\hat{\mathbf{v}}_{\mathbf{I}} = \mathbf{v}_{\mathbf{I}} + \left(\mathbf{S}_{\mathbf{I}}{}^{T}\mathbf{S}_{\mathbf{I}}\right)^{-1}\mathbf{S}_{\mathbf{I}}{}^{T}\mathbf{S}_{\mathbf{O}}\mathbf{v}_{\mathbf{O}} + \left(\mathbf{S}_{\mathbf{I}}{}^{T}\mathbf{S}_{\mathbf{I}}\right)^{-1}\mathbf{S}_{\mathbf{I}}{}^{T}\mathbf{u} \tag{15}$$

Therefore, even when the measurement error $\mathbf{u}$ has zero mean, the flux estimate $\hat{\mathbf{v}}_{\mathbf{I}}$ may no longer be unbiased, i.e. the expected value of $\hat{\mathbf{v}}_{\mathbf{I}}$ may not equal to $\mathbf{v}_{\mathbf{I}}$. In particular, the term $\left(\mathbf{S}_{\mathbf{I}}{}^{T}\mathbf{S}_{\mathbf{I}}\right)^{-1}\mathbf{S}_{\mathbf{I}}{}^{T}\mathbf{S}_{\mathbf{O}}\mathbf{v}_{\mathbf{O}}$ on the right hand side of Eq. (15) gives the specification bias caused by the missing reactions. Therefore, the specification bias scales with the degree of correlations between the stoichiometry of the accounted reactions in $\mathbf{S}_{\mathbf{I}}$ and that of the omitted reactions in $\mathbf{S}_{\mathbf{O}}$ (from $\mathbf{S}_{\mathbf{I}}{}^{T}\mathbf{S}_{\mathbf{O}}$), and with the magnitude of the omitted reaction fluxes $\mathbf{v}_{\mathbf{O}}$. As illustrated in Case Study I below, the omission of a single reaction could lead to a large bias in the flux estimate.

The derivation above shows how the omission of – or failure to include – reactions in the stoichiometric model could cause a bias in the flux estimate in the overdetermined MFA. In addition to the flux bias, missing reactions would also result in a lower variance for the OLS estimate $\hat{\mathbf{v}}_{\mathbf{I}}$, i.e. the variance of $\hat{\mathbf{v}}_{\mathbf{I}}$ in Eq. (14) is smaller than the variance of the OLS estimates for the true model in Eq. (12) [19]. Under certain conditions, the mean square error of the OLS estimate $\hat{\mathbf{v}}_{\mathbf{I}}$ may also be lower than that for the true model [19]. On the contrary, the inclusion of irrelevant reactions in the stoichiometry matrix – **S** contains reactions that are non-existent in the actual system – would not introduce any bias to the OLS estimate of the fluxes. However, having additional reactions would artificially increase the variance and mean square error of the flux estimate [19].

2.2.2. Model Misspecification Tests

There exist several tests to detect model misspecification in a linear least square regression. In this study, we focus on tests of the misspecification of regression mean – whether $\mathbf{X\beta}$ is a good description of the response variable $\mathbf{y}$. While tests for checking the validity of constant variance or normality of the error variables in the least square regression also exist, we do not deal with these so-called misspecifications of higher moments. Interested readers are referred to the article by Long and Trivedi [20]. The first misspecification test under evaluation is the Ramsey RESET (Regression Equation Specification Error Test) test, which does not require any information on the possible missing variables (reactions) to formulate the hypothesis test. In the context of the overdetermined MFA, this scenario corresponds to when the stoichiometry of the missing reactions is unknown. The p order RESET test is based on the following linear least square problem:

$$\mathbf{y} = \mathbf{X\beta} + \alpha_1\hat{\mathbf{y}}^2 + \cdots + \alpha_p\hat{\mathbf{y}}^{p+1} + \mathbf{e}^* \tag{16}$$

where $\hat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\hat{\mathbf{y}}^{\text{p}}$ is the vector of p powers of $\hat{y}_i$ elements. The premise of the RESET test is that the contribution from the missing variables could be approximated by the powers of $\hat{\mathbf{y}}$ (by way of a Taylor series expansion). The null hypothesis in the RESET test is that the coefficients α_i's are zero, i.e. H$_0$: $\alpha_1 = \cdots = \alpha_\text{p} = 0$. The null hypothesis will be rejected if the following condition on the test statistic S_{RESET} is satisfied:

$$S_{RESET} = \frac{\left(\boldsymbol{\Phi}_{\text{OLS}}\left(\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right) - \boldsymbol{\Phi}_{\text{OLS}}\left(\widetilde{\boldsymbol{\beta}}_{\text{OLS}}, \widetilde{\boldsymbol{\alpha}}_{\text{OLS}}\right)\right)/p}{\boldsymbol{\Phi}_{\text{OLS}}\left(\widetilde{\boldsymbol{\beta}}_{\text{OLS}}, \widetilde{\boldsymbol{\alpha}}_{\text{OLS}}\right)/(N-K-p)} > F_{p,N-K-p}(a) \tag{17}$$

where $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is given in Eq. (6), $\widetilde{\boldsymbol{\beta}}_{\text{OLS}}$ and $\widetilde{\boldsymbol{\alpha}}_{\text{OLS}}$ are the OLS estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ for the regression problem in Eq. (16), N is the length of the response vector $\mathbf{y}$, K is the dimension of the unknown parameter $\boldsymbol{\beta}$, and $F_{p,N-K-p}(a)$ is the $(1-a)$th percentile point of the Snedecor's F distribution with p and $N-K-p$ degrees of freedom. In essence, the inequality in Eq. (17) is fulfilled when the additional parameters $\boldsymbol{\alpha}$ in Eq. (16) leads to a (statistically) significantly better data fitting than the linear equation in Eq. (4). A rejection of the null hypothesis is therefore taken as a strong evidence supporting for the existence of a model functional misspecification.

The next two misspecification tests apply to the scenario where the stoichiometry of the candidate missing reactions are known. Given the current extensive knowledge on metabolic reactions, including the complete maps of metabolic network for many major organisms and the availability of extensive and curated biochemical reaction databases (e.g. KEGG [21], MetaCyc [22], RHEA [23]), one could efficaciously put together a list of possible missing reactions in the overdetermined MFA. Here, we consider the least square regression problem:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{e}^{*} \tag{18}$$

where $\mathbf{Z}$ denotes the design matrix of the candidate missing variables – the stoichiometric matrix of the possible missing reactions in the overdetermined MFA. The first test in this category is based on a similar null hypothesis as in the Ramsey RESET test. The null hypothesis H_0: $\boldsymbol{\alpha} = \mathbf{0}$ will be rejected if the following condition on the test statistic S_{F-test} is fulfilled:

$$S_{F-test} = \frac{\left(\boldsymbol{\Phi}_{\mathrm{OLS}}\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right) - \boldsymbol{\Phi}_{\mathrm{OLS}}\left(\bar{\boldsymbol{\beta}}_{\mathrm{OLS}}, \bar{\boldsymbol{\alpha}}_{\mathrm{OLS}}\right)\right)/o}{\boldsymbol{\Phi}_{\mathrm{OLS}}\left(\bar{\boldsymbol{\beta}}_{\mathrm{OLS}}, \bar{\boldsymbol{\alpha}}_{\mathrm{OLS}}\right)/(N-K-o)} \geq F_{o,N-K-o}(a) \tag{19}$$

where $\bar{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\bar{\boldsymbol{\alpha}}_{\mathrm{OLS}}$ are the OLS estimate for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ for the regression problem in Eq. (18) and o is the number of possible missing variables (i.e. the dimension of $\boldsymbol{\alpha}$). We refer the model misspecification test above as, for the lack of a better word, the F-test. Note that the existence of a unique solution of $\bar{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\bar{\boldsymbol{\alpha}}_{\mathrm{OLS}}$ requires that the combined matrix $[\mathbf{X} \quad \mathbf{Z}]$ has a full column rank.

The third and last misspecification test in this work is derived from the Lagrange multiplier (LM) test, as proposed by Davidson and Mackinnon [24]. The premise of the LM test is to examine whether there exists a significant correlation between the residuals of the proposed linear model in Eq. (4) and the part of the matrix $\mathbf{Z}$ in Eq. (18), denoted by denoted by $\mathbf{Z}^c$, that remains after removing the linear influence by $\mathbf{X}$. In addition, the LM test also incorporates a heteroscedasticity-consistent (HC) covariance matrix, to accommodate the situation when the variance of the measurement error is not constant. Following the procedure formulated in Long and Trivedi [20], we first compute the $\mathbf{Z}^c$ as follow

$$\mathbf{Z}^{\mathbf{c}} = \mathbf{Z} - \breve{\mathbf{X}}\left(\breve{\mathbf{X}}^{T}\breve{\mathbf{X}}\right)^{-1}\breve{\mathbf{X}}^{T}\mathbf{Z} \tag{20}$$

where $\breve{\mathbf{X}} = [\mathbf{1} \quad \mathbf{X}]$ and $\mathbf{1}$ is the vector of 1s of an appropriate dimension. The test statistic is given by:

$$S_{LM} = \hat{\mathbf{e}}^{T}\mathbf{Z}^{\mathbf{c}}\left(\frac{N-K}{N-K-o}\mathbf{Z}^{\mathbf{c}^{T}}\widehat{\boldsymbol{\Omega}}\mathbf{Z}^{\mathbf{c}}\right)^{-1}\mathbf{Z}^{\mathbf{c}^{T}}\hat{\mathbf{e}} \tag{21}$$

where $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ denotes the OLS residuals and $\widehat{\boldsymbol{\Omega}}$ is a diagonal matrix with the squares of the standard errors as the diagonal elements ($\widehat{\Omega}_{i,i} = \hat{e}_i^2$). The null hypothesis H_0: $\mathbf{Z}^{\mathbf{c}^{T}}\hat{\mathbf{e}} = \mathbf{0}$ is rejected if $S_{LM} \geq \chi_o^2(a)$, where $\chi_o^2(a)$ denotes the $(1-a)$th percentile point of the chi-square distribution with o degrees of freedom. The rejection of the null hypothesis due to a (statistically) significant correlation between the residual $\hat{\mathbf{e}}$ and $\mathbf{Z}^{\mathbf{c}}$ is taken as a strong evidence for the existence of a model misspecification.

2.2.3. Resolving Model Misspecification

The rejection of the null hypothesis in the tests above does not immediately point to the identity of the missing or omitted reactions. Given the set of possible missing reactions, we could nevertheless apply the F-test or LM test to determine whether the addition of a reaction or a set of reactions among the possible missing reactions would significantly improve the linear regression. Reactions that are positively identified by the tests, could therefore be added into the stoichiometric matrix. The procedure above can be repeated until no more reactions can be added. More precisely, we propose the following iterative procedure for correcting misspecifications in the stoichiometric matrix for the overdetermined MFA using the F-test:

1. Given the exchange fluxes $\mathbf{v}_E$, the stoichiometric matrices $\mathbf{S}_E$ and $\mathbf{S}_I$, and the possible missing reaction stoichiometric matrix $\mathbf{S}_A$, we formulate the linear least square regression problem with $\mathbf{y} = -\mathbf{S}_{\mathbf{E}}\mathbf{v}_{\mathbf{E}}$, $\mathbf{X} = \mathbf{S}_{\mathbf{I}}$, and $\boldsymbol{\beta} = \mathbf{v}_{\mathbf{I}}$.
2. Compute S_{F-test} using $\mathbf{Z}$ constructed from every k-tuple combination of the columns (reactions) of $\mathbf{S}_A$.

3. Identify the k-tuple combination(s) satisfying $S_{F-test} \geq F_{o,N-K-o}(\alpha)$ and move the corresponding columns from **S_A** to **S_I**.
4. Repeat steps 2-3 until no more reactions can be moved from **S_A** to **S_I**, that is until the remaining set of k-tuple reaction combinations satisfying $S_{F-test} \geq F_{o,N-K-o}(\alpha)$ is empty.

The procedure above is written generally for any k-tuple combination of reactions. In the case study, we performed the procedure, first using $k = 1$ (single reactions) and using $k = 1$ followed by $k = 2$ (pairs of reactions). If desired, the LM test could also be used in place of the F-test.

2.3. In silico Metabolic Network Models and Data Generation

2.3.1. Chinese hamster ovary model

In the demonstration of the specification bias caused by missing reactions and the iterative procedure for resolving the stoichiometric matrix misspecification, we employed a metabolic network model previously created for the flux analysis of Chinese hamster ovary (CHO) batch cultivation data [16,25]. The model describes the concentration of 49 metabolites, out of which 34 are transported in and out of the cell by exchange fluxes. The stoichiometric matrix $\mathbf{S_I}$ has a dimension of 49 metabolites and 47 internal fluxes with a full column rank. Figure 1 illustrates the CHO metabolic network model that corresponds to the stoichiometric matrix $\mathbf{S_I}$ in the MFA. The complete list of reactions is given in Supplementary Table S1.

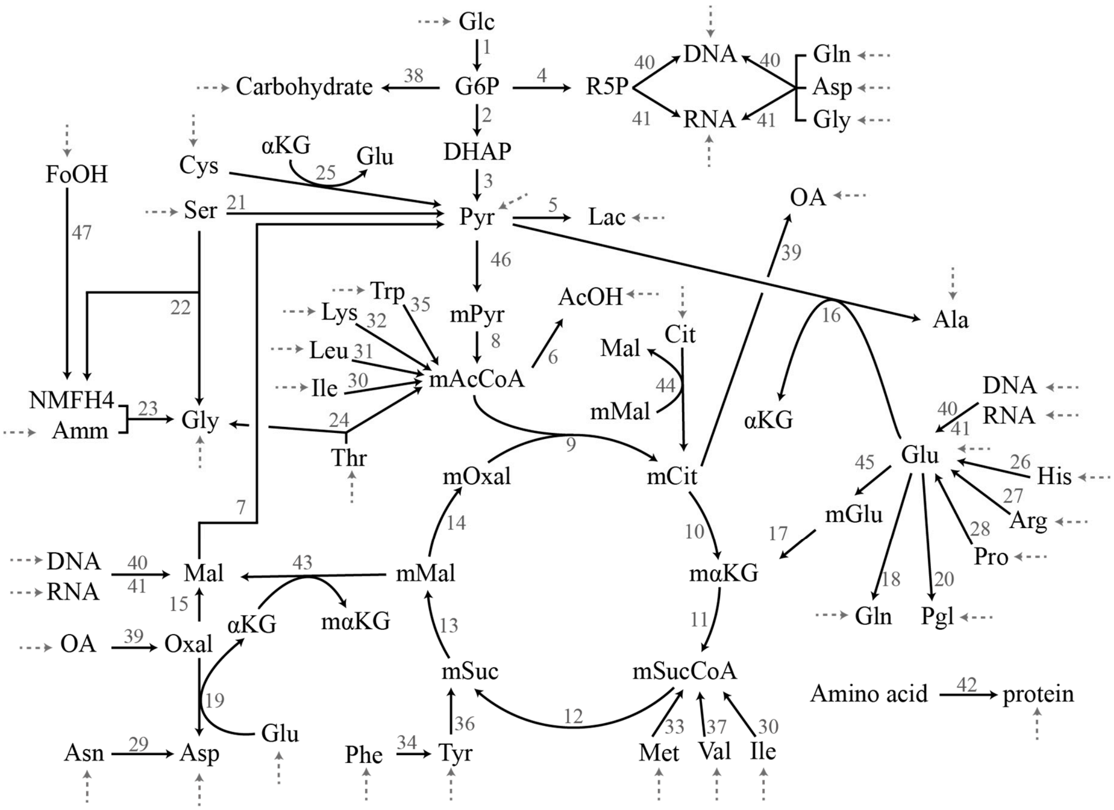

Figure 1. Chinese hamster ovary metabolic network model in Case Study I and III (adapted from [16,25]). The dashed arrows indicate the exchange (uptake) fluxes.

2.3.2. Random Metabolic Models

For a large scale evaluation of the model misspecification tests, we generated *in silico* data $\mathbf{y} = -\mathbf{S_E v_E}$ using randomly generated stoichiometric matrices $\mathbf{S_I}$ of various sizes (m = 50 to 200 metabolites and n = 55 to 220 reactions) and with different numbers of exchange fluxes (between 25 and 100). More precisely, for each data vector $\mathbf{y}$, we employed the RMBNToolbox (MATLAB) [26] to generate a random stoichiometric matrix $\mathbf{S_I}$ with the desired dimensions and a full column rank, where each species participated in at least one reaction. For the specified number of exchange fluxes m_E, we assigned the first m_E metabolites (rows) of $\mathbf{S_I}$ as the species whose exchange fluxes were measured. Accordingly, we partitioned the matrix $\mathbf{S_I}$ into

$$\mathbf{S_I v_I} = \begin{bmatrix} \mathbf{S_{I,E}} \\ \mathbf{S_{I,NE}} \end{bmatrix} \mathbf{v_I} = -\mathbf{S_E v_E} = \mathbf{y} \tag{22}$$

where $\mathbf{S_{I,E}}$ corresponds to the first m_E rows of the measured exchange fluxes and $\mathbf{S_{I,NE}}$ corresponds to the remaining rows. We subsequently generated the internal flux vector $\mathbf{v_I}$ using a linear combination of the kernel of $\mathbf{S_{I,NE}}$ with uniform random coefficients between -1 and 1. Finally, we contaminated the computed $\mathbf{y} = -\mathbf{S_E v_E} = \mathbf{S_I v_I}$ with independent realizations of Gaussian distributed random numbers with zero mean and for different coefficients of variation (CoV of 1 to 20%). The misspecification tests were applied to the *in silico* generated data $\mathbf{y}$, using the design matrix $\mathbf{X}$ constructed by removing reactions randomly (between 2 to 20 reactions) from the matrix $\mathbf{S_I}$.

The performance of each misspecification test was judged based on the rate of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). For the Ramsey RESET test, the true positives and false negatives were determined by applying the test to the generated data using the stoichiometric matrix $\mathbf{S_I}$ with some of the reactions (columns) randomly removed. In this scenario, a rejection of the null hypothesis corresponded to a TP, while a non-rejection was a FN error. Meanwhile, the numbers of FP and TN of the RESET test were computed based on rejections and non-rejections of the null hypothesis, respectively, when applying the test using the full stoichiometric matrix $\mathbf{S_I}$ (no missing reactions). The data generation and the RESET test were repeated for 1000 times, each using a different randomly generated stoichiometric matrix $\mathbf{S_I}$.

In order to determine the TP and FN rates of the F-test and LM test, we generated *in silico* data $\mathbf{y}$ as described above and applied the tests using the stoichiometric matrix $\mathbf{S_I}$ with a set of its reactions randomly removed. Here, the matrix $\mathbf{Z}$ came from the set of actual reactions that were removed from the matrix $\mathbf{S_I}$. On the other hand, for the determination of the FP and TN rates, we repeated the above tests but using a matrix $\mathbf{Z}$ containing a distinct set of linearly independent reactions of the same size as the set of the omitted reactions. The above procedure was again repeated for 1000 times. Finally, in all of these tests, we ensured that the rank condition for the OLS estimation was always satisfied.

3. Results

3.1. Case study I: Specification bias

In the first case study, we considered the CHO metabolic model in Figure 1 with the measured exchange flux values and standard deviations reported previously [16]. We employed the GLS regression to obtain the estimate of $\mathbf{v_I}$, and used the resulting $\mathbf{\hat{v}_{I,GLS}}$ (see Supplementary Table S2) for the calculations below. Table 1 gives the minimum, median, mean and maximum absolute specification bias for the omission of single reactions, one at a time, from the stoichiometric matrix $\mathbf{S_I}$. Here, we only removed reactions that would not create an orphan species, i.e. a species that does not participate in any reaction. For each reaction removal, we also generated 10,000 vectors of *in silico* data of $\mathbf{y} = \mathbf{S_I v_I}$ using the full $\mathbf{S_I}$ matrix, and contaminated the data with independent Gaussian random noise with the variance-covariance matrix constructed from the reported standard deviations [16]. For each data vector, we evaluated the significance of regression by the analysis of variance (ANOVA) [27] using the reduced $\mathbf{S_I}$ matrix, i.e. the matrix $\mathbf{S_I}$ with a missing column (reaction). The averages of the p values from the ANOVA are given in Table 1. Here, we took p value of 0.05 as the threshold to reject the GLS regression – any p values higher than the threshold indicate a poor regression outcome.

The individual removal of roughly 3/4 of the reactions (26 out of 36 reactions) still produced a significant regression with $p < 0.05$. On average, the median, mean and maximum specification biases in the flux estimates were higher for the removal of reactions that caused a poor regression ($p > 0.05$). The two highest p values came from the removal of reactions with the two largest fluxes, and each expectedly had large specification biases. There were nevertheless exceptions, where a poor regression resulted from removing a reaction with a moderately low flux value (e.g. reactions 21 and 22). On the other hand, many of the cases with a significant regression ($p < 0.05$), were associated with high maximum specification biases. In fact, the case with the lowest p value (i.e. the most significant regression) had a mean bias of ~38% and a maximum bias of above 1000%. Equally important, the removal of several reactions with a low flux magnitude led to large mean and maximum flux biases (mean bias >150%), as highlighted in Table 1. Therefore, while a poor regression generally points to a model misspecification problem or a violation of the assumption on measurement noise, a statistically significant regression does not guarantee a small specification bias in the flux estimates. In addition, removing reactions with a low flux magnitude can cause disproportionately large specification biases in the flux estimate. These observations clearly motivate the use of a more systematic assessment of the model misspecification issue in the overdetermined MFA.

Table 1. Case study I: Specification bias in the CHO model

			Absolute Specification Bias (%)[d]			
Reaction[a]	$\hat{v}_{l,GLS}$	***p* value[c]**	**min**	**median**	**mean**	**max**
25	-0.02	0.00±0.00	0.00	0.41	2.73	54.1
19	0.03	0.00±0.00	0.00	0.39	2.48	48.8
10	-1.46	0.00±0.00	0.00	0.15	1.96	11.6
45	-0.21	0.00±0.00	0.00	2.04	18.3	269
17	-0.21	0.00±0.00	0.00	2.11	19.0	280
31	-0.24	0.00±0.00	0.00	2.83	24.9	361
27	0.34	0.00±0.00	0.00	2.12	15.6	229
14	12.50	0.00±0.00	0.00	1.31	33.3	855
9	12.50	0.00±0.00	0.00	1.31	33.3	855
46	15.04	0.00±0.00	0.00	0.88	38.1	1020
8	15.04	0.00±0.00	0.00	0.88	38.1	1020
37	0.27	0.00±0.00	0.00	5.86	54.1	753
12	17.42	0.00±0.00	0.02	1.28	43.1	1190
11	17.84	0.00±0.00	0.02	1.09	44.0	1220
13	18.06	0.00±0.00	0.02	1.66	46.7	1230
30	-0.27	0.00±0.00	0.00	6.87	63.8	889
24	-0.38	0.00±0.00	0.00	6.67	60.9	860
26	0.27	0.00±0.00	0.00	4.19	36.1	509
35	0.13	0.00±0.00	0.00	3.12	28.8	399
33	0.22	0.00±0.00	0.00	11.7	124	2060
32[b]	-1.18	0.01±0.01	0.00	21.0	196	2770
29[b]	0.99	0.01±0.01	0.00	13.9	170	3840
34[b]	0.47	0.01±0.01	0.00	16.2	152	2110
36[b]	0.87	0.02±0.01	0.00	23.1	217	3020
23[b]	-1.34	0.02±0.01	0.00	17.1	177	2470
28[b]	1.03	0.02±0.01	0.00	17.6	158	2220
15	12.31	0.05±0.02	0.00	14.3	121	2210
4	1.24	0.05±0.02	0.00	5.25	65.1	2420
21	-6.81	0.13±0.03	0.00	53.9	475	6980
16	19.26	0.15±0.03	0.00	61.4	573	8100
18	-21.53	0.20±0.04	0.00	61.6	632	8830
43	19.52	0.46±0.04	0.00	74.0	477	8050
22	7.24	0.47±0.04	0.00	121	988	14700
7	19.63	0.52±0.04	0.00	21.8	114	2360
2	157.77	0.97±0.01	0.00	1.28	803	21600
3	315.55	0.97±0.01	0.00	1.28	803	21600

[a] The reaction numbers refer to the CHO metabolic network in Figure 1.

[b] The omission of reactions with a low flux could cause large specification biases in the flux estimate.

[c] The significance of regression was assessed by ANOVA. The average p value (mean ± standard error) was computed for 10,000 GLS regressions using independently generated *in silico* data.

[d] The minimum, median, mean and maximum biases were computed over the remaining reaction fluxes in the model.

3.2. Case study II: Stoichiometric model misspecification tests

We evaluated the ability of the Ramsey RESET test, F-test and LM test in detecting the issue of stoichiometric matrix misspecification in the overdetermined MFA, particularly the existence of omitted or missing reactions from $\mathbf{S}_\mathbf{I}$. As outlined in Materials and Methods, we determined the rates of TP, TN, FP and FN using randomly generated pairs of data $\mathbf{y} = -\mathbf{S}_\mathbf{E}\mathbf{v}_\mathbf{E}$ and stoichiometric matrices $\mathbf{S}_\mathbf{I}$ with missing reactions. For the F- and LM tests, we used the information on the actual missing reactions, as well as a distinct set of reactions, as the design matrix of the missing variables $\mathbf{Z}$. The results using the baseline MFA problems with 100 metabolites, 60 unknown internal reactions, 50 measured exchange reactions, and 2, 5 or 10 missing reactions for different noise levels (1 to 20% CoV), are summarized in Table 2. Note that by definition, the TP and FN rates sum to 1 and so do the FP and TN rates.

In general, Table 2 shows that as the level of measurement noise increase (higher CoV), the TP rates expectedly drops. We also observed that the smaller the set of missing reactions, the poorer were the TP rates. This trend was expected since with fewer missing reactions, the reduced $\mathbf{S}_\mathbf{I}$ was closer to the true system and could more accurately capture the flux balance. Therefore, for the tests to correctly detect a misspecification of $\mathbf{S}_\mathbf{I}$, the missing reactions would need to cause a significant deterioration in the data fitting, a scenario that is less likely to occur as the number of missing reactions becomes lower. Meanwhile, the FP rates were not a strong function of the noise level. The FP rates improved with a lower number of missing reactions, albeit only slightly. The results in Table 2 illustrate the poor performance of the RESET test with low rates of TP and high rates of FP. Increasing the order of the RESET test from $p = 1$ to 2 improved the TP rates at the cost of increased FP rates. The LM test has the lowest FP rates among all tests, with moderate TP rates. The best overall performer was the F-test with >80% TP rates in almost all settings (except for 2 missing reactions at 20% CoV) and <15% FP rates. The TP rates of the F-test were nearly 1 for the highest number of missing reactions.

Considering the performance values in Table 2, we therefore recommend the F-test for detecting model misspecification in the overdetermined MFA. The F-test however requires the stoichiometry of the candidate missing reactions as an input. With the extensive knowledge on metabolic reactions available in the literature and in online public databases, such a requirement may not be overly limiting. The results from further evaluations of the F-test performance for metabolic networks of different sizes ($m = 50$ and 200 metabolites, and $n = 55$ and 220 reactions, respectively) is summarized in Table 3. With larger networks, detecting the same number of missing reactions became more difficult, as expected. At the largest network size ($m = 200$), the rate of correctly detecting a misspecification with 2 missing reactions went slightly lower than 60%. Fortunately, the TP rates for detecting 5 or more missing reactions were still high (>88%). Not to mention, the FP rates depended weakly on the size of the networks and the number of missing reactions, and remained relatively low between 10 to 15% in most of the cases in our study (see also Supplementary Table S3).

Table 2. Case study II: Performance of model misspecification tests (values represent rates)

`					RESET test (p = 1)				RESET test (p = 2)				F-test				LM test			
m	n_{v_I}	n_{v_E}	n_{v_O}	CoV	TP	FN	FP	TN	TP	FN	FP	TN	TP	FN	FP	TN	TP	FN	FP	TN
100	**60**	**50**	**2**	**0.01**	0.18	0.82	0.56	0.44	0.33	0.67	0.75	0.25	0.86	0.14	0.09	0.91	0.68	0.32	0.11	0.89
				0.05	0.28	0.72	0.57	0.43	0.44	0.56	0.78	0.22	0.82	0.19	0.09	0.91	0.67	0.33	0.14	0.86
				0.1	0.32	0.69	0.58	0.42	0.51	0.49	0.76	0.24	0.82	0.19	0.10	0.90	0.66	0.34	0.16	0.84
				0.2	0.42	0.58	0.56	0.44	0.69	0.31	0.81	0.19	0.71	0.29	0.08	0.92	0.60	0.41	0.18	0.82
			5	**0.01**	0.11	0.89	0.57	0.43	0.33	0.67	0.76	0.25	0.99	0.01	0.14	0.87	0.71	0.29	0.07	0.93
				0.05	0.12	0.88	0.54	0.46	0.34	0.67	0.73	0.27	0.98	0.02	0.12	0.88	0.73	0.27	0.06	0.94
				0.1	0.19	0.81	0.54	0.46	0.41	0.59	0.75	0.25	0.97	0.03	0.13	0.87	0.71	0.29	0.11	0.90
				0.2	0.29	0.71	0.55	0.45	0.58	0.42	0.82	0.19	0.93	0.07	0.11	0.89	0.70	0.30	0.12	0.88
			10	**0.01**	0.11	0.89	0.57	0.43	0.40	0.60	0.73	0.27	1.00	0.00	0.11	0.89	0.47	0.53	0.00	1.00
				0.05	0.13	0.87	0.57	0.43	0.42	0.58	0.76	0.24	1.00	0.00	0.10	0.90	0.48	0.52	0.01	0.99
				0.1	0.16	0.84	0.54	0.46	0.47	0.53	0.75	0.26	1.00	0.00	0.13	0.87	0.48	0.52	0.01	0.99
				0.2	0.26	0.74	0.57	0.43	0.57	0.43	0.79	0.21	0.99	0.01	0.12	0.88	0.44	0.56	0.01	0.99

Table 3. Case Study II: Additional misspecification tests using the F-test (values represent rates)

m	n_{v_I}	n_{v_E}	n_{v_O}	CoV	TP	FN	FP	TN
50	**30**	**25**	**2**	0.01	0.86	0.14	0.11	0.89
				0.05	0.82	0.18	0.10	0.90
				0.1	0.75	0.25	0.09	0.91
				0.2	0.69	0.31	0.09	0.91
			5	0.01	0.99	0.01	0.10	0.90
				0.05	0.98	0.02	0.10	0.90
				0.1	0.97	0.03	0.10	0.90
				0.2	0.92	0.08	0.11	0.89
			10	0.01	1.00	0.00	0.10	0.90
				0.05	1.00	0.00	0.09	0.91
				0.1	1.00	0.00	0.09	0.91
				0.2	0.99	0.02	0.11	0.90
200	**120**	**100**	**2**	0.01	0.76	0.24	0.11	0.89
				0.05	0.73	0.27	0.10	0.90
				0.1	0.67	0.33	0.07	0.93
				0.2	0.58	0.42	0.10	0.90
			5	0.01	0.97	0.03	0.16	0.84
				0.05	0.95	0.05	0.11	0.89
				0.1	0.94	0.07	0.13	0.87
				0.2	0.88	0.12	0.13	0.88
			10	0.01	1.00	0.00	0.15	0.85
				0.05	0.99	0.01	0.16	0.84
				0.1	1.00	0.01	0.13	0.87
				0.2	0.98	0.02	0.15	0.85
			20	0.01	1.00	0.00	0.14	0.86
				0.05	1.00	0.00	0.14	0.86
				0.1	1.00	0.00	0.15	0.86
				0.2	1.00	0.00	0.14	0.86

3.3. Case study III: Resolving model misspecification

In the last case study, we evaluated the performance of the proposed iterative procedure to resolve stoichiometric matrix misspecifications in the overdetermined MFA (see Materials and Methods). Here, we returned to the flux analysis of the CHO metabolic network in Figure 1. For the performance assessment, we created 100 different stoichiometric matrices $\mathbf{S}_{\mathbf{I,true}}$ by randomly removing a number n_{extra} of columns from the stoichiometric matrix $\mathbf{S}_{\mathbf{I}}$ of the CHO model. For each $\mathbf{S}_{\mathbf{I,true}}$, we generated an artificial data vector $\mathbf{y} = \mathbf{S}_{\mathbf{I,true}}\mathbf{v}_{\mathbf{I,true}}$ using the GLS flux estimate (see

Supplementary Table S2), and contaminated the data vector with independent Gaussian noise with zero mean and the variance-covariance matrix as in Case Study I. The data generation procedure was repeated for 100 times. For each data vector, we then created a reduced matrix $\mathbf{S}_{\mathbf{I,red}}$ by randomly removing a number n_{omit} of reactions from $\mathbf{S}_{\mathbf{I,true}}$. The reactions removed in the creation of $\mathbf{S}_{\mathbf{I,true}}$ and $\mathbf{S}_{\mathbf{I,red}}$ were subsequently combined in the matrix $\mathbf{S}_{\mathbf{A}}$. In other words, the set of candidate missing reactions $\mathbf{S}_{\mathbf{A}}$ had equal fractions of the actual omitted reactions and the extra reactions that were not used in the *in silico* data generation. Finally, we applied the strategy for resolving model misspecification to each data vector using the matrix $\mathbf{S}_{\mathbf{I,red}}$ as the reduced stoichiometric matrix and the matrix $\mathbf{S}_{\mathbf{A}}$ as the candidate missing reaction matrix. The strategy was implemented using two settings: (1) $k = 1$ and (2) $k = 1$ followed by $k = 2$.

Table 4 gives the number of reactions in $\mathbf{S}_{\mathbf{A}}$ that were not positively identified by the iterative procedure to be included in the stoichiometric matrix. As an indication of good performance, the number of omitted reactions (extra reactions) that remained should be low (high). The results in Table 4 demonstrated that the proposed procedure using $k = 1$ was able to correctly detect and incorporate almost all of the omitted reactions, while keeping the incorrect inclusion of extra reactions low. As expected, performing an additional run with $k = 2$ after finishing the procedure with $k = 1$ led to a higher incorporation rate of the omitted reactions. But, such a strategy came at the cost of a higher rate of incorrect addition of the extra reactions. Because of the small size of the CHO model and the number of missing reactions considered, a higher k (e.g. $k = 2$) led to the incorporation of all omitted and extra reactions (see Supplementary Table S4). Considering the trade-off above, we thus recommend using a simple implementation with $k = 1$ to resolve the issue of stoichiometric matrix misspecifications in the overdetermined MFA.

Table 4. Case study III: Iterative procedure for resolving model misspecification in the CHO model

			Number of remaining reactions[a]	
k	n_{extra}	n_{omit}	**Extra reactions**	**Omitted reactions**
1	3	3	2.82±0.38	0.99±0.10
	5	5	4.13±0.63	1.34±0.46
	8	8	5.89±0.83	2.21±0.48
1 then 2	5	5	3.66±0.59	0.97±0.17
	8	8	5.03±0.70	1.00±0.29

[a]The number of remaining reactions (mean ± standard error) corresponds to the average over 100 generations of the stoichiometric matrix $\mathbf{S}_{\mathbf{I,true}}$, of the median number across 100 *in silico* data simulations.

4. Discussion

Metabolic flux analysis is an indispensable tool for elucidating and understanding the metabolic phenotype of cells, with numerous applications in biotechnology and biomedical fields. The core of the MFA is the stoichiometric model of the metabolic network, which under pseudo-steady state assumption, enforces a constraint on the distribution of the metabolic flux values. The power of MFA methods stems from its ability to provide estimates on the intracellular metabolic fluxes, from measurements of extracellular species concentrations or ^{13}C isotopic labeling experiments, using only the stoichiometry on the metabolic reactions. However, because of its reliance on the stoichiometric model, the accuracy of the flux predictions should therefore depend sensitively on the veracity of the stoichiometric matrix in the MFA. Despite its obvious importance, the impact, detection and rectification of stoichiometric matrix misspecifications have not received much attention in the past. We aimed to fill this gap for the overdetermined MFA, where the flux estimation problem constitutes an overdetermined linear regression problem.

Statistical analysis of linear least square regression has provided numerous tools for assessing, among other things, the goodness of fit, gross measurement errors, accuracy of flux estimate and error propagation in the overdetermined MFA (as presented in Introduction). In this work, we adapted the statistical analysis and tests for model misspecifications in linear least square regressions. To the best of our knowledge, such analysis and tests have not yet been applied to the flux estimation problem of the overdetermined MFA. Here, we first derived a simple formula to evaluate the specification bias in the flux estimate due to missing reactions in the stoichiometric model. Using a stoichiometric model of the CHO metabolism, we showed that the significance of the regression is not a sufficient indicator of a low flux bias. In particular, the omission of reaction(s) that results in a high bias (error) in the flux estimate, could still produce a statistically acceptable regression. Furthermore, we demonstrated that the removal of reactions with a low flux value could also cause disproportionately large specification biases. In practice, the significance of the regression and the prior information on the magnitude of reaction fluxes are often used for the curation of a reduced order stoichiometric model for the overdetermined MFA. Our findings from the first case study clearly motivated computing the potential flux specification bias during the removal of reaction(s).

Among the statistical tests that we evaluated for detecting stoichiometric matrix misspecifications, the results from random metabolic networks clearly demonstrated the superiority of the F-test. The F-test could provide high TP rates and low FP rates for detecting missing reactions for most problem sizes (except when there were only very few missing reactions). Based on these findings, we proposed an iterative procedure using the F-test to detect and correct missing reactions. In each iteration, we used the F-test to identify a combination of k reactions from the set of candidate missing reactions, that would (statistically) significantly improve the linear regression. The combinations that passed a certain p value threshold, are then incorporated in the stoichiometric model. The application of this iterative procedure to the CHO stoichiometric model indicated that the simple implementation using $k = 1$ gave the most robust performance.

Finally, there exist several obvious limitations in the statistical tests evaluated in this work. First, these tests rely on the assumption of normality for the noise. The validity of this assumption could be checked by standard normality test such as Lilliefors or Shapiro-Wilk test [28], or using a normal probability plot. When the number of measurements is sufficiently large (>30 for non-skewed noise), the normality assumption is typically reasonable, thanks to the central limit theorem. Another limitation of the tests is the imposition on the rank of the matrices, i.e. the set of reaction stoichiometry provides linearly independent vectors. While the rank condition above applies to the standard formulation of any overdetermined MFA, the F-test further require that the set of candidate missing reactions also have linearly independent stoichiometry to the existing reactions in the stoichiometric model.

5. Conclusions

In this work, we addressed the misspecification of the stoichiometric matrix in the overdetermined MFA, particularly the omission of reactions. For this purpose, we adapted statistical analysis and tools from linear least square regression to quantify, detect and resolve the issue of missing reactions. In particular, we derived a simple formula to evaluate the flux bias caused by missing reactions in the overdetermined MFA. We further assessed the performance of several model misspecification tests, namely Ramsey RESET test, F-test and Lagrange multiplier test, in detecting missing reactions for overdetermined stoichiometric models. Finally, we proposed an iterative procedure for resolving the issue of missing reactions based on applying the F-test to the candidate missing reactions one at a time and incorporating reactions that pass the significance test. The application of these techniques to CHO metabolic model and random metabolic networks provided several important outcomes. First of all, the significance of the regression, a common metric for assessing the data-model consistency in the overdetermined MFA, could not guarantee a low bias in the flux estimates. In addition, a high flux bias could result from the removal of a reaction with a low flux magnitude that would be deemed unimportant in the construction of the stoichiometric matrix. Therefore, the potential flux bias due to any removal of reaction(s) should be computed during the curation of the stoichiometric model. When the stoichiometry of the candidate missing reactions is known, the F-test provides a robust means, with a high TP rate (nearly 100% for many cases) and a relatively low FP rate (<15%), to detect model misspecifications in the overdetermined MFA. Upon a positive detection of model misspecification in the overdetermined MFA, the proposed iterative procedure in this study gives a systematic approach to effectively and robustly resolve this issue.

Supplementary Materials: The following are available online: Table S1: Metabolic reactions and exchange fluxes in the Chinese hamster ovary metabolic model, Table S2: Intracellular flux estimate of the CHO cell culture, Table S3: Case Study II: Other misspecification tests using the F-test, Table S4: Case Study III: Iterative procedure for resolving model misspecification in the CHO model ($k = 2$), and MATLAB codes of the misspecification tests and the flux analysis of the CHO model.

Acknowledgments: The authors would like to acknowledge financial support from ETH Zurich.

Author Contributions: RG designed the study. RG and SH wrote the MATLAB codes, and generated and analyzed the results. RG and SH wrote the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stephanopoulos, G. Metabolic fluxes and metabolic engineering. *Metab. Eng.* **1999**, *1*, 1–11.
2. Bailey, J. E. Bioprocess Engineering. *Adv. Chem. Eng.* **1991**, *16*, 425–462.
3. Lee, S. Y.; Park, J. M.; Kim, T. Y. Application of metabolic flux analysis in metabolic engineering. *Methods Enzymol.* **2011**, *498*, 67–93.
4. Lewis, N. E.; Nagarajan, H.; Palsson, B. O. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* **2012**, *10*, 291–305.
5. Crown, S. B.; Antoniewicz, M. R. Publishing 13C metabolic flux analysis studies: A review and future perspectives. *Metab. Eng.* **2013**, *20*, 42–48.
6. Naderi, S.; Meshram, M.; Wei, C.; Mcconkey, B.; Ingalls, B.; Budman, H.; Scharer, J. Development of a mathematical model for evaluating the dynamics of normal and apoptotic Chinese hamster ovary cells. *Biotechnol. Prog.* **2011**, *27*, 1197–1205.
7. Nolan, R. P.; Lee, K. Dynamic model of CHO cell metabolism. *Metab. Eng.* **2011**, *13*, 108–24.
8. van der Heijden, R. T. J. M.; Heijnen, J. J.; Hellinga, C.; Romein, B.; Luyben, K. C. A. M. Linear constraint relations in biochemical reaction systems: I. Classification of the calculability and the balanceability of conversion rates. *Biotechnol. Bioeng.* **1994**, *43*, 3–10.
9. Wang, N. S.; Stephanopoulos, G. Application of Macroscopic Balances to the Identification of Gross Measurement Errors. *Biotechnol. Bioeng.* **1983**, *25*, 2177–2208.
10. Goudar, C. T.; Biener, R. K.; Piret, J. M.; Konstantinov, K. B. Metabolic Flux Estimation in Mammalian Cell Cultures. In *Animal Cell Biotechnology: Methods and Protocols*; Pörtner, R., Ed.; Humana Press: Totowa, NJ, 2014; pp. 193–209.
11. Wiechert, W.; Siefke, C.; DeGraaf, a a; Marx, A. Bidirectional reaction steps in metabolic networks: II. Flux estimation and statistical analysis. *Biotechnol. Bioeng.* **1997**, *55*, 118–135.
12. Goudar, C. T.; Biener, R.; Piret, J. M.; Konstantinov, K. B. Metabolic flux estimation in mammalian cell cultures. In *Methods in Biotechnology*; Springer, 2007; Vol. 24, pp. 301–317.
13. Hädicke, O.; Lohr, V.; Genzel, Y.; Reichl, U.; Klamt, S. Evaluating differences of metabolic performances: Statistical methods and their application to animal cell cultivations. *Biotechnol. Bioeng.* **2013**, *110*, 2633–2642.
14. van der Heijden, R. T. J. M.; Romein, B.; Heijnen, J. J.; Hellinga, C.; Luyben, K. C. A. M. Linear constraint relations in biochemical reaction systems: II. Diagnosis and estimation of gross errors. *Biotechnol. Bioeng.* **1994**, *43*, 11–20.
15. Quek, L.-E.; Dietmair, S.; Krömer, J. O.; Nielsen, L. K. Metabolic flux analysis in mammalian cell culture. *Metab. Eng.* **2010**, *12*, 161–71.
16. Sokolenko, S.; Quattrociocchi, M.; Aucoin, M. G. Identifying model error in metabolic flux analysis – a generalized least squares approach. *BMC Syst. Biol.* **2016**, *10*, 1–14.
17. Erdrich, P.; Steuer, R.; Klamt, S. An algorithm for the reduction of genome-scale metabolic network models to meaningful core models. *BMC Syst. Biol.* **2015**, *9*, 1–12.
18. Chipman, J. S. Gauss-Markov Theorem. In *International Encyclopedia of Statistical Science*; 2014; pp. 577–582.
19. Rao, P. Some Notes on Misspecification in Multiple Regressions. *Am. Stat.* **1971**, *25*, 37–39.
20. Long, J. S.; Trivedi, P. K. Some Specification Tests for the Linear Regression Model. *Sociol. Methods Res.* **1992**, *21*, 161–204.

21. Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **2006**, *34*, D354-7.
22. Caspi, R.; Billington, R.; Ferrer, L.; Foerster, H.; Fulcher, C. A.; Keseler, I. M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Mueller, L. A.; Ong, Q.; Paley, S.; Subhraveti, P.; Weaver, D. S.; Karp, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **2016**, *44*, D471–D480.
23. Morgat, A.; Lombardot, T.; Axelsen, K. B.; Aimo, L.; Niknejad, A.; Hyka-Nouspikel, N.; Coudert, E.; Pozzato, M.; Pagni, M.; Moretti, S.; Rosanoff, S.; Onwubiko, J.; Bougueleret, L.; Xenarios, I.; Redaschi, N.; Bridge, A. Updates in rhea-an expert curated resource of biochemical reactions. *Nucleic Acids Res.* **2017**, *45*, D415–D418.
24. Davidson, R.; MacKinnon, J. Heteroskedasticity-Robust Tests in Regression Directions. *Ann. Insee.* **1985**, *59/60*, 183–218.
25. Altamirano, C.; Illanes, a; Casablancas, A.; Gámez, X.; Cairó, J. J.; Gòdia, C. Analysis of CHO cells metabolic redistribution in a glutamate-based defined medium in continuous culture. *Biotechnol. Prog.* **2001**, *17*, 1032–41.
26. Aho, T.; Smolander, O.-P.; Niemi, J.; Yli-Harja, O. RMBNToolbox: random models for biochemical networks. *BMC Syst. Biol.* **2007**, *1*, 22.
27. Montgomery, D. C. *Applied Statistics and Probability for Engineers*; 6th ed.; Wiley, 2003; Vol. 37.
28. Conover, W. J. *Practical nonparametric statistics*; 3rd ed.; Wiley, 1999.