Dynamic modelling of biochemical reaction networks and sampling methods for constraint-based models

Shirin Fallahi

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2021

UNIVERSITY OF BERGEN

Dynamic modelling of biochemical reaction networks and sampling methods for constraint-based models

Shirin Fallahi

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 05.02.2021

Year: 2021

Title: Dynamic modelling of biochemical reaction networks and sampling methods for constraint-based models

Name: Shirin Fallahi

Print: Skipnes Kommunikasjon / University of Bergen

Preface

This dissertation is submitted as a partial fulfillment of the requirements for the degree of Doctor of Philosophy (PhD). This study is carried out at the Department of Mathematics, University of Bergen. Parts of the work in the thesis have been carried out in collaboration with the Department of Biological science, University of Bergen, Department of Biological and Environmental Sciences, University of Gothenburg and Department of Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg. The work is supported by the Research Council of Norway through grant 248840, dCod 1.0, as part of the Center of Digital Life Norway.

The subject of the thesis is dynamic Modelling of biochemical reaction networks and sampling methods for constraint-based models. Paper A and B are mathematical methodology papers where the candidate is the first author of paper A and the second author of paper B. Papers C and D are mathematical modelling papers motivated by two biological applications. The candidate is the first author of papers C and D.

Advisory committee:

- Guttorm Alendal (University of Bergen, Department of Mathematics)
- Hans Julius Skaug (University of Bergen, Department of Mathematics)
- Anders Goksøyr (University of Bergen, Department of Biological science)

Outline

This thesis consists of two parts. Part I provides the motivation and the background for the four papers presented in part II.

Ch. 1 gives an introduction to the biological context and two applications in biology that inspired the mathematical models in papers C and D. The structure of a general reaction network is defined in Ch. 2. A detailed description of the constraint-based modelling of the reaction network is presented in Ch. 3. A summary of paper A is also given in Ch. 3. This chapter contains both deterministic and stochastic formulations of the constraint-based model. Ch. 3 also provides a description of the sampling algorithms associated with each formulation to study the probability distributions of the reaction rates in the network. The kinetics functions to describe the reaction rates are presented in Ch. 4. Each kinetics function is derived from a set of ordinary differential equations for the concentrations in the network. A summary of paper B is also given in Ch. 4. Using the kinetics functions to construct a dynamic model of the network leads to the problem of parameter estimation and identifiability of the kinetic parameters. In Ch. 5 we have introduced these concepts and provided a summary of their applications in two dynamic models of papers C and D. Ch. 5 includes also a description of the local and global sensitivity analysis of concentrations simulated by a dynamic model to changes in the parameter values. Finally in part I, an overview of the papers is given in Ch. 6.

List of papers

Paper A:

Shirin Fallahi, Hans J. Skaug, Guttorm Alendal. *A comparison of Monte Carlo sampling methods for metabolic network models*. PLOS ONE, **15**:7 (2020) doi: 10.1371/journal.pone.0235393

Paper B:

Håvard G. Frøysa, **Shirin Fallahi** and Nello Blase. *Evaluating model reduction under parameter uncertainty*. BMC Systems Biology, **12**:79 (2018) doi: 10.1186/s12918-018-0602-x

Paper C:

Shirin Fallahi, Marie Mlnaíková, Charlotte Alvord, Guttorm Alendal, Håvard G. Frøysa, Torbjörn Lundh and Malin C. Celander. *A new conceptual toxicokinetic model to assess synergistic mixture effects between the aromatic hydrocarbon β-naphthoflavone and the azole nocodazole on the CYP1A biomarker in a fish cell-line*. Revision submitted to: Environmental Science and Technology, September 2020.

Paper D:

Shirin Fallahi, Rune Kleppe, Anders Goksøyr, Guttorm Alendal. *A mathematical model of allosteric regulation in the cytosolic fatty acid synthesis pathway*.

Acknowledgements

I have received support and guidance from many people throughout the writing of this dissertation and believe that this accomplishment would not have been possible without their invaluable help. I am so grateful for all the help and support I got.

First, I would like to thank my supervisors Guttorm Alendal, Hans J. Skaug and Anders Goksøy. A special thanks to Guttorm and Hans for your patience, motivation and guidance. I am truly grateful for the valuable comments and immense knowledge I got from Hans while writing paper A in this thesis. A special thanks also to Guttorm and Anders for introducing me to my co-authors Runne Kleppe, Malin C. Celander and Torbjörn Lundh from the University of Bergen, University of Gothenburg and Chalmers University of Technology. It was a great pleasure for me to work with you on the papers and I thank you for your guidance and support.

Next, I want to thank my co-authors of the papers specially Håvard G. Frøysa and Nello Blaser. I enjoyed working with you on paper B of this thesis. A special thanks to Håvard for his invaluable help and great discussions. I would express a very special thanks to the whole dCod 1.0 project. It was a great pleasure working with you and having great scientific discussions in our biweekly meetings. As a member of dCod project I got the opportunity to participate in many events, workshops and conferences and I am very thankful for this.

I would further like to thank the Department of Mathematics at the University of Bergen, for all of the opportunities I was given to carry out this PhD thesis. I really enjoyed working in such a friendly environment. A special thanks to Kristian Gundersen for making coffee in the coffee breaks and to Anna Oleynik for her valuable advice.

Last but not least, a big thanks to my husband Pouya Rahmani for all his advice, patience and kindness, I am so grateful. A special thanks to my parents for all your support and kindness. Finally, I would like to thank all my family and friends for their support.

Shirin Fallahi
September 2020

Contents

Part I

Background

Chapter 1

Introduction and Biological Motivation

Mathematical modelling of biochemical reaction networks is the main subject of this thesis. In this chapter, we give an overview of some biological motivation and background that inspired the mathematical modelling in the thesis.

1.1 Motivation

Mathematical modelling of molecular biology has extended the traditional interaction diagrams to mechanistic mathematical models, becoming a valuable tool in molecular cell biology. There are different applications in biology that demonstrate how mathematical modelling has been a valuable tool. Here we present a brief description of two applications that are the motivation behind the mathematical modelling in the thesis.

1.1.1 The dCod 1.0 project

When starting my PhD, I joined the dCod 1.0 project [2], which ran from 2016 to 2020 and has been funded by the Research council of Norway through the Center for Digital Life Norway [1]. During the work on the thesis, dCod 1.0 has been the main scientific community and all co-authors of the papers included in the thesis are members of this project. The main goal of the project is *to create a deeper understanding of cods' adaptations and reactions to stressors in the environment* by employing the competencies in environmental toxicology, biology, bioinformatics and mathematics. Studying the response of the cod liver and modelling fatty acid metabolism [4] has been of prime interest in the dCod 1.0, where precision cod liver slices have been widely used in the studies [8; 16; 23; 24; 105].

Yadetie et al. [104] found that the exposure of Atlantic cod (Gadus morhua) to environmental contaminants, such as PCB153, increases the levels of enzymes involved in the fatty acid synthesis of the cod liver. The motivation of my PhD project is to study the influence of changes in the level of enzymes through constructing a mathematical model of fatty acid synthesis metabolism [3]. In paper D of this thesis, a model for fatty acid metabolism is presented. It would be of considerable interest if we were able to validate the suggested model by using experimental data and potentially use it to predict the behaviour of the fatty acid metabolism in cod liver under various changes in the level of the enzymes caused by different environmental contaminants.

1.1.2 The FORMAS project

During my PhD period, I joined the FORMAS project, which has been run by one of the collaborators of the dCod 1.0 project in Sweden. The purpose of this project is *to develop new mathematical forecast models to obtain synergistic assessment factors that can be applied in chemical risk assessments and biomonitoring programs*.

Fish are typically exposed to mixtures of environmental pollutants in their natural habitats. Exposure to chemical mixtures can cause different additive or non-additive mixture effects, such as synergistic mixture effects. Two toxic chemicals that can enter the fish body are the aromatic hydrocarbon and AhR agonist β-naphtoflavone (BNF) and the benzimidazole and microtubule disassembling drug nocodazole (NOC). These two chemicals have a synergistic mixture effect on AhR-CYP1A signalling, which is central to the chemical detoxification pathway [45]. Using experimental data from a previous study [37; 61], I have created a model in paper C to study the toxicokinetic interactions between BNF and NOC on the AhR-CYP1A signalling pathway in fish and/or a fish cell line.

1.2 Biochemical reaction networks and reconstruction

Knowledge about the chemical reactions that take place inside the cells of a living system is required to study the above mentioned phenomena. The *metabolism* of an organism is known as the set of all chemical reactions that are carried out by the organism. Metabolism is typically divided into several *metabolic* pathways in which one chemical is converted to another chemical through a reaction. These chemicals are known as *metabolites*.

The process of identifying the set of all metabolites and reactions in the metabolism of an organism is known as network reconstruction [68]. A *genome* of an organism is the set of all the organisms' genes from which the set of all the metabolites and reactions can be obtained. The *genome scale reconstructions* of an organism is the process of reconstructing the whole metabolism of the organism. All the available biochemical, genetic and genomic (BiGG) information is required in a reconstruction of a metabolic reaction network [67]. In the reconstruction process, the BiGG information is collected and organised in a mathematical model that could serve as a platform to answer biological questions.

The first metabolic network reconstructed was for the *E.coli* bacteria, with the findings published in 1990 [68]. In 2000, Edwards and Palsson [22] published the first genome scale reconstruction of E. coli, and this was improved by Orth et al. [64] in 2011. Orth et al. [64] attempted to model the whole metabolism, which includes 1,136 metabolites and 2,251 reactions.

Genome scale reconstruction of human metabolism is of great interest in the field as metabolism plays a key role in many major human diseases. The first genome scale reconstruction of the human metabolic network, Recon 1, is published in 2007 by Duarte et al. [19], which accounts for 2,766 metabolites and 3,741 reactions. In 2013, Thiele et al. [88] published an improved reconstruction of human metabolism, Recon 2, with a focus on the expansion of metabolic reaction coverage. Recon 2 consists of 5,063 metabolites and 7,440 reactions. Recon 3D, published in 2018, represents the

most comprehensive human metabolic network model to date, accounting for 10,600 reactions involving 5,835 metabolites [13].

The fatty acid synthesis reaction network from the Kegg database, shown in Fig. 1.1, is an example of a reconstructed metabolic pathway for Zebrafish, which is the most similar species to cod. In paper D, I have created a dynamic model of a reduced version of this metabolic reaction network.

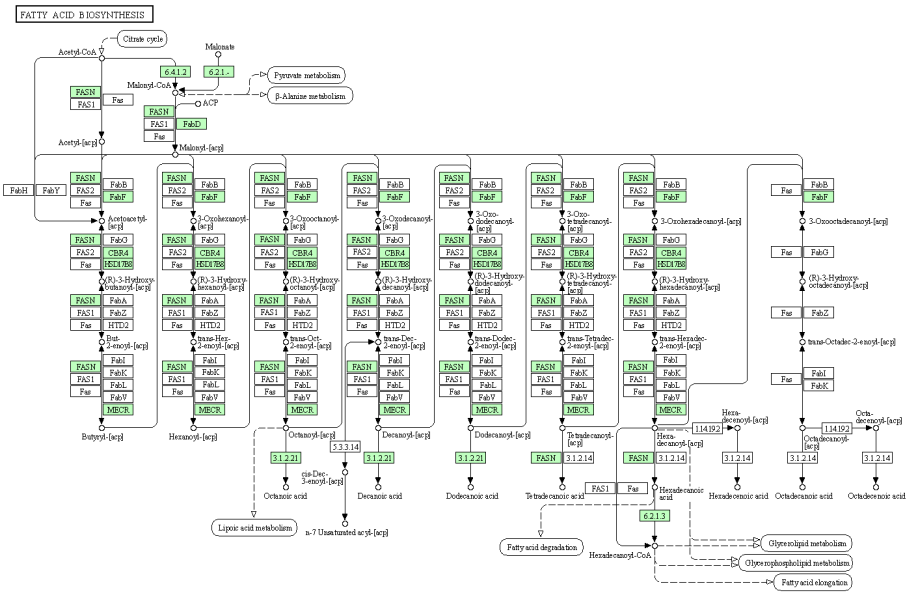


Figure 1.1: The fatty acid synthesis metabolic pathway from KEGG [3]. The relevant reactions for Zebrafish are marked by green.

In the thesis, we assume that the metabolic network is given, and the reconstruction process of a chemical network is not in focus.

1.2.1 Metabolites and chemical reactions

A reaction network is composed of chemical reactions that occur between metabolites. The six chemical elements of carbon (C), oxygen (O), nitrogen (N), hydrogen (H), phosphate (P) and sulfur (S) are the only elements that form metabolites [68, p. 155]. Below we list the chemical formulas for the metabolites considered in the mathematical model of the fatty acid synthesis metabolic pathway [3] in paper D.

- citrate, $C_6H_5O_7^{-3}$
- acetyl-CoA, $C_{23}H_{38}N_7O_{17}P_3S$
- malonyl-CoA, $C_{23}H_{38}N_7O_{19}P_3S$
- palmitoyl-CoA, $C_{37}H_{66}N_7O_{17}P_3S$

We denote the metabolites by the mathematical notation X_i as their chemical formula is not the main focus of our mathematical modelling of a chemical reaction network.

There are only three different categories of chemical reactions that occur between metabolites [68, p. 154]. Assume that C, P and A denote a primary metabolite, a phosphate group and a co-factor such as ATP, respectively. The three categories of reactions are thus:

- a *reversible conservation*
$$CP \rightleftharpoons PC \tag{1.1}$$
- a *bi-molecular association*
$$C + P \rightleftharpoons CP \tag{1.2}$$
- a *co-factor-coupled reaction*
$$C + AP \rightleftharpoons CP + A. \tag{1.3}$$

A metabolite is called a reactant or substrate if it is on the left hand side of a reaction, and it is called a product if it is on the right hand side of the reaction. As an instance, the third category of reactions in our models is presented as

$$X_1 + X_2 \rightleftharpoons X_3 + X_4. \tag{1.4}$$

where X_i denotes the metabolite i.

1.3 Systems biology

To model and analyse the metabolism of a living organism requires system thinking, modelling and simulation in biology as a data-rich field [67]. A new field known as systems biology emerged in the late 20^{th} century. The field has shifted the way modelling is conducted from the component level to the system level to answer new questions that only arise in the system level. This shift has occurred due to the ability to generate genome-wide data sets that include the detailed list of biological components and their interactions [47], as well as the emergence of computers with high computational powers.

The data sets in biology are typically large and have usually been processed and analysed statistically using methods in the bioinformatics field [54]. However, knowledge from both bioinformatics and mathematics is required to construct a biochemical reaction network and build a mathematical model of the network. Systems biology studies biological systems in a systematic way by first perturbing the systems and then monitoring the system response, integrating this data and finally constructing mathematical models to describe the structure of the biological system and its response to different perturbations [47]. COBRA 2.0 [82] and Raven 2.0 [100] are the toolboxes that have been widely used to help with this analysis. In the thesis, we focus on the mathematical modelling for metabolic reaction networks and the methods to analyse such models.

Chapter 2

Structure properties of reaction networks

In this chapter, we define the *stoichiometric matrix* **S**, which is the mathematical representation of the structure of a chemical reaction network. This matrix is the central part of the mathematical models of chemical reaction networks discussed in the next chapters of this thesis. The main reference for this chapter is the textbook "Systems biology" by Palsson [68].

2.1 The stoichiometric matrix

Metabolism is essentially a large network of coupled chemical reactions. Assume that the chemical reaction network is composed of n irreversible reactions in which m metabolites are involved. Any reaction R_j, $j = 1, \ldots, n$ can thus be written as

$$R_j : \quad \sum_{i=1}^{m} s_{ij}^{-} X_i \rightarrow \sum_{i=1}^{m} s_{ij}^{+} X_i \tag{2.1}$$

where X_i is the metabolite i. The s_{ij}^{-} and s_{ij}^{+} are non-negative integers denoting the substrate and product stoichiometric coefficients, respectively [56].

The stoichiometric matrix $\mathbf{S} = \{S_{ij}\} \in \mathbb{R}^{m \times n}$ of the chemical reaction network is formed by the net production of metabolite i in reaction j defined as $S_{ij} = s_{ij}^{+} - s_{ij}^{-}$. The stoichiometric matrix **S** is denoted as

$$\mathbf{S} = \begin{array}{c} \begin{matrix} R_1 & \ldots & R_n \end{matrix} \\ \begin{bmatrix} S_{11} & \ldots & S_{1n} \\ \vdots & \ddots & \vdots \\ S_{m1} & \ldots & S_{mn} \end{bmatrix} \end{array} \begin{matrix} X_1 \\ \vdots \\ X_m \end{matrix} \tag{2.2}$$

where each column corresponds to a reaction in the network and each row corresponds to a metabolite involved in different reactions. In the remainder, we drop the labels X_i and R_j from the matrix **S** for simplicity.

In Eq. 2.1 we describe a general case where the metabolite X_i can be present in both sides of the reaction R_j and both stoichiometric coefficients s_{ij}^{+} and s_{ij}^{-} are, in this case, non-zero. Note that a reversible reaction is presented as two irreversible reactions in Eq. 2.1 and that one can instead model it directly as a reversible reaction.

In a chemical reaction network the metabolite X_i is typically present in one side of the reaction R_j, and this leads to $S_{ij} = s_{ij}^+$ or $S_{ij} = -s_{ij}^-$. Furthermore, a typical metabolic pathway is an open reaction network that includes exchange reactions. We have $s_{ij}^+ = 0$ or $s_{ij}^- = 0$ for all i if the exchange reaction R_j is leaving or entering the pathway. In a large metabolic network, a few numbers of metabolites are typically involved in each reaction. Therefore, most of the s_{ij}^+ and s_{ij}^- are zero and this makes **S** a sparse matrix. In a genome scale metabolic network, the most common number of metabolites participating in a reaction is four [68, p. 165].

The topological properties of a reaction network and its possible dynamical behaviours can be determined purely by the matrix **S**, see e.g. the PhD thesis by Håvard G. Frøysa [30].

2.1.1 Reaction maps

The standard way of visualising the network in Eq. 2.1 is a directed graph in which the nodes are the metabolites and the edges between the nodes are the reactions between the metabolites. The matrix **S** represents this directed graph which is known as *reaction map* [68].

The linear maps are networks in which all reactions have only one input and one output. For a linear reaction map, the columns of **S** have only two entries corresponding to the metabolites that the reaction connects [68, p. 165]. A simple example network is shown in Fig. 2.1 to illustrate a linear reaction map. The network consists of four metabolites and seven reactions. It includes a reversible reaction and a split and is open since it has the entering reaction R_1 and two leaving reactions R_6 and R_7.

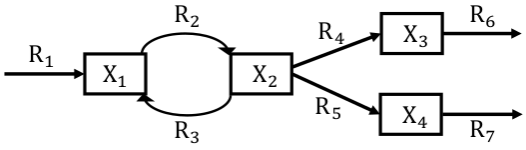


Figure 2.1: An example reaction network consisting of a reversible reaction and a split. Each node is a metabolite and each arrow is a reaction. The reactions are given in Eq. 2.3

.

Assuming for simplicity that all non-zero s_{ij}^+ and s_{ij}^- are equal to one, the reactions of the network can be represented on the form Eq. 2.1 as

$$\begin{aligned}
R_1 &: \quad \longrightarrow X_1 \\
R_2 &: \quad X_1 \longrightarrow X_2 \\
R_3 &: \quad X_2 \longrightarrow X_1 \\
R_4 &: \quad X_2 \longrightarrow X_3 \\
R_5 &: \quad X_2 \longrightarrow X_4 \\
R_6 &: \quad X_3 \longrightarrow \\
R_7 &: \quad X_4 \longrightarrow
\end{aligned} \tag{2.3}$$

where R_1 is an entering reaction, R_2, R_3, R_4 and R_5 are the internal reactions and the reactions R_6 and R_7 are the leaving reactions. The two reactions R_2 and R_3 form a reversible reaction together. The 4×7 stoichiometrix matrix of the network is given as

$$\mathbf{S} = \begin{array}{c} \begin{matrix} R_1 & R_2 & R_3 & R_4 & R_5 & R_6 & R_7 \end{matrix} \\ \begin{bmatrix} 1 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix} \end{array} \begin{matrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{matrix} \tag{2.4}$$

A reaction map of a network is nonlinear if it includes the reactions with more than one input or output. One can count the number of metabolites involved in a reaction by adding up the number of non-zero entries in the corresponding column of $\mathbf{S}$ [68, p. 165]. An example network adopted from Rao et al. [76] is given in Fig. 2.2 to illustrate a nonlinear reaction map. The network consists of four metabolites and five reactions that interact through complexes, which are defined in the next section.

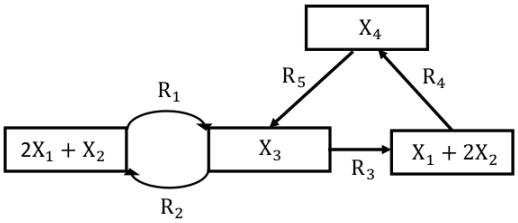


Figure 2.2: Example of a nonlinear reaction network from Rao et al.[76]. Each node is a complex consisting of potentially several metabolites and each arrow is a reaction. The reactions are given in Eq. 2.5.

In this network, we have more than one metabolite on each side of four reactions, and this leads to a nonlinear reaction map. In addition, all the non-zero s_{ij}^{+} and s_{ij}^{-} are not necessarily one for this example network. The reactions of the network can be represented on the form Eq. 2.1 as

$$\begin{aligned} R_1 &: \quad 2X_1 + X_2 \longrightarrow X_3 \\ R_2 &: \quad X_3 \longrightarrow 2X_1 + X_2 \\ R_3 &: \quad X_3 \longrightarrow X_1 + 2X_2 \\ R_4 &: \quad X_1 + 2X_2 \longrightarrow X_4 \\ R_5 &: \quad X_4 \longrightarrow X_3 \end{aligned} \tag{2.5}$$

where two reactions R_1 and R_2 are together a reversible reaction. The network represents a closed system as there is no reaction entering or leaving the network. The 4×5 stoichiometric matrix of the network becomes

$$\mathbf{S} = \begin{array}{c} \begin{matrix} R_1 & R_2 & R_3 & R_4 & R_5 \end{matrix} \\ \begin{bmatrix} -2 & 2 & 1 & -1 & 0 \\ -1 & 1 & 2 & -2 & 0 \\ 1 & -1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \end{array} \begin{matrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{matrix} . \tag{2.6}$$

2.2 Complex factorisation of the stoichiometric matrix

In a typical metabolic reaction network, there is more than one metabolite on either side of a reaction, as shown in the example network of Fig. 2.2. A complex C_k is any combination of metabolites that occur on either side of the reactions for a network [76]. In this section, we explain the complex factorisation of the stoichiometric matrix $\mathbf{S}$ as this is the core concept in paper B.

A set of complexes is defined as the union of all the different left- and righthand sides of the reactions in a network. For instance, the network in Fig. 2.2 has four complexes shown in the boxes. The complexes corresponding to the network are

$$C_1 : 2X_1 + X2, \qquad C_2 : X3, \qquad C_3 : X_1 + 2X2, \qquad C_4 : X4 \tag{2.7}$$

where two complexes C_2 and C_4 are composed of only one metabolite. The expression of the complexes in terms of the metabolites is formalised by the *complex stoichiometric matrix* $\mathbf{Z} = \{Z_{ik}\} \in \mathbb{N}_0^{m\times c}$ where m is the number of metabolites and c is the number of complexes [76]. Each column of $\mathbf{Z}$ corresponds to a complex, and each row is associated with a metabolite presented in different complexes. The entry Z_{ik} is the number of molecules of metabolite X_i presented in one unit of the complex C_k. All entries of $\mathbf{Z}$ are non-negative integers since a complex cannot contain negative numbers of metabolites. The 4×4 complex stoichiometric matrix corresponding to the complexes of Eq. 2.7 is

$$\mathbf{Z} = \begin{array}{c} \begin{matrix} C_1 & C_2 & C_3 & C_4 \end{matrix} \\ \begin{bmatrix} 2 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{matrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{matrix} . \tag{2.8}$$

The complex stoichiometric matrix $\mathbf{Z}$ equals the identity matrix when all complexes are composed of only one metabolite.

Since the complexes are the left- and righthand sides of the reactions in a reaction map, they can be considered as nodes that are connected through the reactions in the network. Such a network is called the *complex graph* [76]. The reactions in a complex graph can be rewritten such that the left- and righthand sides of all reactions are individual complexes. The network in Fig. 2.2 is a complex directed graph and its associated reactions can be

$$\begin{aligned} R_1 &: \quad C_1 \longrightarrow C_2 \\ R_2 &: \quad C_2 \longrightarrow C_1 \\ R_3 &: \quad C_2 \longrightarrow C_3 \\ R_4 &: \quad C_3 \longrightarrow C_4 \\ R_5 &: \quad C_4 \longrightarrow C_2 \end{aligned} \tag{2.9}$$

where the complexes, C_k, are defined in Eq. 2.7.

A complex graph can be defined by its *incidence matrix* $\mathbf{B} = \{B_{kj}\}$ [9]. This is a $c \times n$ matrix where c is the number of complexes and n is the number of reactions in the network. The B_{kj} equals -1 if the complex C_k is consumed in reaction R_j and 1 if the complex C_k is produced in reaction R_j, while otherwise 0. The 4×5 incidence matrix

associated with the example network in Fig. 2.2 is given as

$$\mathbf{B} = \begin{array}{c} \begin{matrix} R_1 & R_2 & R_3 & R_4 & R_5 \end{matrix} \\ \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \end{array} \begin{matrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{matrix} . \tag{2.10}$$

Note that in Rao et al. [76] and paper B, only the internal reactions are included in the incidence matrix **B**. However, it is also possible to include the exchange reactions in **B**.

Having the complex stoichiometric matrix **Z** and the incidence matrix **B**, the complex factorisation of the stoichiometric matrix **S** in Eq. 2.2 is given by

$$\mathbf{S} = \mathbf{ZB}. \tag{2.11}$$

Note that this identity holds only if the potential exchange reactions in **S** have been included in **B**. For the matrix **S** in Eq. 2.6, it can be simply verified that

$$\mathbf{S} = \mathbf{ZB} = \begin{bmatrix} 2 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}. \tag{2.12}$$

Chapter 3

Dynamic modelling of metabolic reaction networks and Monte Carlo sampling

So far we have defined the stoichiometric matrix **S** in Eq. 2.2, which can be used to explore the structural behaviour of a chemical reaction network. In this chapter, we discuss how it can be used to construct a mathematical dynamic model of a chemical reaction network. Several modelling techniques are available for metabolic networks. In this thesis, we discuss two widely used approaches classified as (i) constraint-based modelling, which involves stoichiometry and (ii) mechanistic modelling, which involves both stoichiometry and kinetic parameters. Models in the first category are known as static models while the second category presents the dynamic models [75].

To construct a dynamic model of a metabolic reaction network, we first assume *spatial homogeneity*, that is, that the metabolites are equally distributed throughout the volume. The second assumption is the *continuum hypothesis* in which we assume that there are a great many reactant molecules. This assumption allows us to approximate the discrete changes in molecule numbers by the continuous changes in concentration **x** [48, p. 19].

Assume we have a metabolic reaction network as Eq. 2.1, which consists of m metabolites X_i and n reactions R_j. We define $\mathbf{x} = \begin{bmatrix} x_1 & \dots & x_m \end{bmatrix}^T \in \mathbb{R}^m_{\geq 0}$ as a vector consisting of metabolite concentrations. The x_i is the concentration of the metabolite X_i, and it is non-negative as the concentration of a metabolite cannot be negative. Let the vector $\mathbf{v} = \begin{bmatrix} v_1 & \dots & v_m \end{bmatrix}^T \in \mathbb{R}^n$ be the vector of fluxes (reaction rates) such that v_j is the flux associated with the reaction R_j. Note that if the reaction R_j is reversible, its associated rate v_j can acquire a negative value; otherwise, v_j must be non-negative.

3.1 Dynamic modelling

Dynamics deals with changes i.e. the evolution in time of a system. The dynamic quantities are assumed to be the concentration of the metabolites, **x**, and the fluxes, **v**. It is further assumed that the metabolite concentrations, **x**, changes only by the fluxes, **v**, and these changes satisfy the mass balance defined by matrix **S**. The changes in the metabolite concentrations are then modelled by a set of ordinary differential equations

$$\frac{d\mathbf{x}}{dt} = \mathbf{S}\mathbf{v} \tag{3.1}$$

which is the core equation in the dynamical modelling of metabolic networks [68]. Note that in Eq. 3.1 the reaction fluxes $\mathbf{v}$ have not been restricted and no assumption has been made for them. In the next chapter, we discuss different kinetic functions $\mathbf{v} = \mathbf{v}(\mathbf{x})$ that define the fluxes $\mathbf{v}$ as a function of the metabolite concentrations $\mathbf{x}$. Defining the fluxes as a function of the metabolite concentrations, the Eq. 3.1 can be written as

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}) \tag{3.2}$$

where $\mathbf{f}(\mathbf{x})$ is possibly a nonlinear function given by $\mathbf{S}$ and $\mathbf{v} = \mathbf{v}(\mathbf{x})$. In this chapter, we look at the dynamical model in the form of Eq. 3.1 and discuss some widely used methods to analyse the fluxes in biochemical networks that are modelled through Eq. 3.1.

3.1.1 Constraint-based modelling (CBM) and flux balance analysis (FBA)

The *constraint-based modelling* (CBM) approach aims to describe the potential behaviour of an organism by enforcing the physio-chemical constraints on the metabolic network such as mass balance, energy balance, and flux limitations [21; 22; 29; 50; 74; 93–97]. In this type of modelling the main assumption is that the organism reaches the steady state under any given environmental condition

$$\mathbf{S}\mathbf{v}_{ss} = \mathbf{0} \tag{3.3}$$

where the system is at rest and the metabolite concentrations remain constant in Eq. 3.1. The Eq. 3.3 is a set of m algebraic equations with n number of unknown fluxes $\mathbf{v}_{ss}$. In a typical metabolic network the number of metabolites, m, is less than the number of reactions, n, leading to multiple steady state solutions for the set of equations in Eq. 3.3. To find a single vector of fluxes $\mathbf{v}_{ss}$, the feasible solution space of Eq. 3.3 should be more constrained.

The other assumption in the CBM approach is the capacity constraint on the steady state fluxes

$$\mathbf{v}^{lb} \leq \mathbf{v}_{ss} \leq \mathbf{v}^{ub} \tag{3.4}$$

where $\mathbf{v}^{lb}$ and $\mathbf{v}^{ub}$ are vectors of lower and upper bounds on the reaction fluxes. Note that the value of $\mathbf{v}^{lb}$ must be zero in the case of irreversible reactions. These lower and upper bounds are experimentally available for a limited number of fluxes, and the rest are arbitrarily set to large values. The set of constraints in Eq. 3.4 reduces the set of feasible steady state fluxes $\mathbf{v}_{ss}$; however, it is not sufficient in general to avoid multiple steady state solutions.

Flux variability analysis (FVA) [58] is a widely used computational tool to find the maximum and minimum attainable values of each flux that still satisfy both the steady sate constraint in Eq. 3.3 and the capacity constraints in Eq. 3.4. In FVA, for each flux v_j ($j = 1, \ldots, n$) two constrained optimisation problems are constructed as

$$\begin{aligned} \max_{\mathbf{v}_{ss}}/\min_{\mathbf{v}_{ss}} \quad & (v_j)_{ss} \\ \text{s.t.} \quad & \mathbf{S}\mathbf{v}_{ss} = \mathbf{0} \\ & \mathbf{v}^{lb} \leq \mathbf{v}_{ss} \leq \mathbf{v}^{ub}. \end{aligned} \tag{3.5}$$

The optimisation problems can be cast as Linear Programming (LP) problems and solved by the Simplex method [92]. An application of the FVA is to reduce the size of the solution space by setting the upper and lower bounds in Eq. 3.4 to the minimum and maximum flux values provided by FVA.

Flux balance analysis (FBA) [65] is a successful CBM method in which an optimality criterion, typically based on evolutionary arguments, is introduced to identify a unique set of flux estimations satisfying constraints in Eqs. 3.3 and 3.4. The first papers leading up to the FBA method were published in the 1980s [28; 69; 70; 101]. In the recent decade, the FBA has become a popular modelling approach for flux analysis of metabolic networks, even in the genome scale [50; 65]. A reason could be that the FBA tries to find a biologically meaningful and feasible vector of the fluxes by using a limited amount of information about the network and providing valuable information about the behaviour of the system.

In the FBA, the objective function to be optimised as a function of the fluxes can take different forms depending on various biological functionalities, e.g. the maximising growth rate or ATP production of an organism [68]. A common choice of the objective function is the biomass yield corresponding to maximising cell growth [26] mathematically described as

$$\sum_{j=1}^{n} c_j (v_j)_{ss} = \mathbf{c}^T \mathbf{v}_{ss} \tag{3.6}$$

where the known coefficient c_j determines the contribution of flux v_j to the biomass objective function. With $\mathbf{v}^{lb}$, $\mathbf{v}^{ub}$ and $\mathbf{c}$ specified, the constrained optimisation problem can be cast as a Linear Programming (LP) problem

$$\begin{aligned} \max_{\mathbf{v}_{ss}} \quad & \mathbf{c}^T \mathbf{v}_{ss} \\ \text{s.t.} \quad & \mathbf{S}\mathbf{v}_{ss} = \mathbf{0} \\ & \mathbf{v}^{lb} \leq \mathbf{v}_{ss} \leq \mathbf{v}^{ub} \end{aligned} \tag{3.7}$$

which is a well studied optimisation problem that may be solved by the Simplex method [92]. If $\mathbf{v}^{lb} \leq \mathbf{0}$ the solutions are guaranteed to exist, and if $\mathbf{v}^{ub}$ is finite the solution is guaranteed to be bounded. In the case of degeneracy, the LP problem in Eq. 3.7 will not have a unique solution, and alternative optimal solutions (AOS) will be obtained. This phenomenon happens often when applying FBA on genome scale metabolic networks, and there are several approaches to addressing this. The most common one is to solve a secondary optimisation, e.g. to minimise the total flux of the solution [68].

A challenge in FBA is to choose an appropriate objective function to reflect the most likely physiological state of the organism. The flux solutions obtained from FBA are also biased to the choice of the objective function [43]. There are alternative approaches to analysing the fluxes of a metabolic network, see e.g. the textbook by Palsson [67]. One of these is to sample from the flux solution space formed by Eqs. 3.3 and 3.4, which is explained in the next section.

3.2 Sampling of constraint-based metabolic models

An approach to analyse the metabolic fluxes is the *uniform random sampling* of the fluxes' solution space. This method provides a statistically meaningful number of flux solutions that are uniformly distributed over the entire feasible solution space defined by Eqs. 3.3 and 3.4. This approach also provides an unbiased assessment of the flux solution space for a metabolic network since it does not enforce any extra constraints on the model. A Monte Carlo (MC) sampling algorithm is typically used for sampling purposes. In paper A we have discussed and compared four different sampling algorithms to sample the flux solution space at genome scale.

3.2.1 Deterministic formulation and Monte Carlo sampling algorithms

The FBA method provides a single point estimation of the metabolic fluxes. Suppose that we want to study the probability distribution of the fluxes and their mean value. Typically, for metabolic networks with a large number of fluxes, the analytical form of the joint probability density function of fluxes is unavailable. Hence, Monte Carlo approximations are often used [53]. An MC algorithm draws a large number of flux samples which are then used to estimate the mean flux values and marginal probability distribution of the fluxes.

In paper A, we distinguish between a deterministic and stochastic formulation of a given metabolic model and the associated flux measurements. The deterministic formulation of a given metabolic network is

$$\begin{aligned} &\mathbf{S}\mathbf{v}_{ss} = \mathbf{0} \\ &\mathbf{A}\mathbf{v}_{ss} = \mathbf{b} \\ &\mathbf{v}^{lb} \leq \mathbf{v}_{ss} \leq \mathbf{v}^{ub} \end{aligned} \tag{3.8}$$

where the steady state flux measurements are encoded in the data vector $\mathbf{b} \in \mathbb{R}^n$. The matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a diagonal matrix where A_{jj} is one in the presence of data for v_j and otherwise zero. The deterministic formulation in Eq. 3.8 describes the exact steady state phase of the network considering the limited capacity of the fluxes and also integrates the available experimental values of the fluxes in the model by fixing them at the given values.

The flux solution space of Eq. 3.8 forms a closed convex polytope P which is the n-dimensional analogue to the three dimensional polyhedron [38]. An example network and the concept of a metabolic model's deterministic formulation is illustrated in Fig. 3.1 (b and c). For a more detailed description of the figure and different setups, read about the figure in paper A.

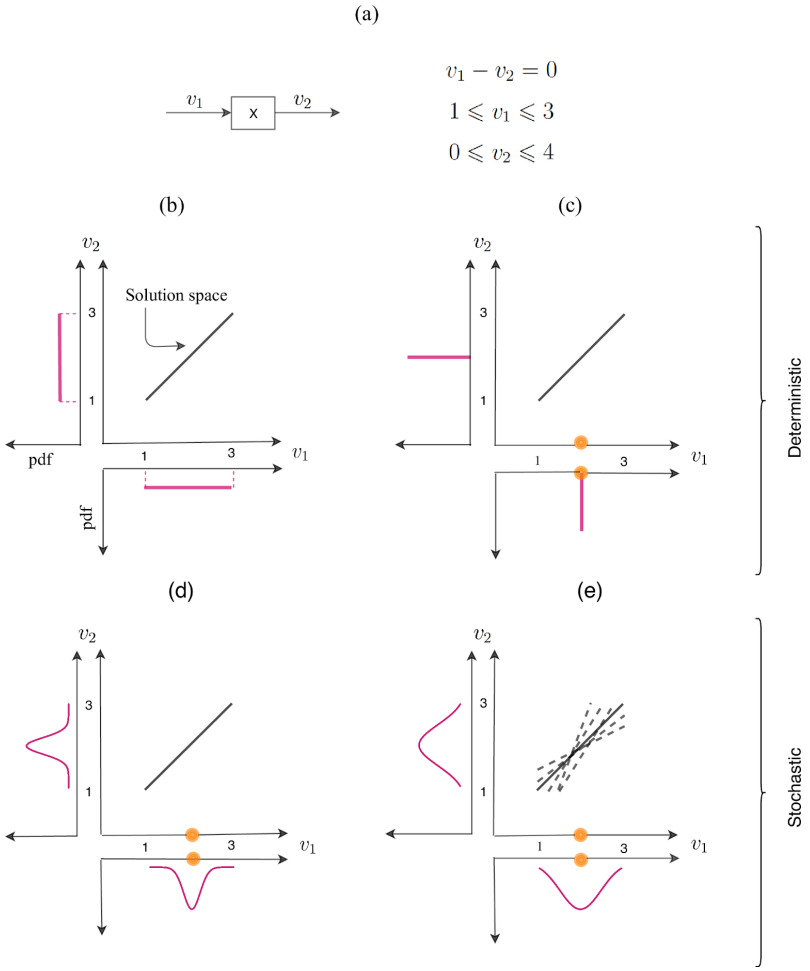


Figure 3.1: Fig. 1 of paper A. Different experimental setups and associated solution space and sampling probability density function (pdf) $\mathrm{p}(\mathrm{v})$ *(pink curve). (a): Example metabolic network and corresponding mathematical model. (b): Deterministic formulation without measurements. (c) and (d): Flux measurement of* v_1 *(orange circle) available in deterministic setup (c) and stochastic setup (d). (e): Relaxed steady state assumption and flux measurement of* v_1 *in stochastic setup.*

For low dimensional polytopes ($\mathrm{n} < 15$) a MC rejecting sampling algorithm can be used [102]. However, this method becomes inefficient, and advanced MC sampling algorithms are required when the dimension of the polytope, which equals the number of the reactions in the metabolic network, is high. The more commonly used sampling algorithm is the Hit-and-Run (HR) [6; 90], which is a Markov Chain Monte Carlo (MCMC) approach. The HR algorithm was applied by Almaas et al. [5] for the first time to sample the flux solution space of the bacterium Escherichia coli metabolic network. The HR algorithm generates a new sample by selecting a random direction. The HR has the Markovian nature i.e. each future sample is dependent only on the current sample. Convergence to the target distribution is guaranteed for a MCMC sampling approach, see, for instance, [36]. The standard HR algorithm samples efficiently from the high

dimension polytope so long as the polytope is heterogeneous in scales of the fluxes. For a typical metabolic network, the associated polytope is narrow in some flux directions as the orders of the flux scales' magnitude might be very small. In the presence of narrow corners, the HR algorithm must take small steps, which makes a full exploration of the solution space impossible in finite time.

To address this issue, the Artificial Hit-and-Run (ACHR) [51] and optimised general parallel sampler (OPTGP) were developed [60] based on the standard HR algorithm. The core idea of the ACHR is to use optimal direction choices in HR (where the direction of the next sample is chosen at random) to allow for larger steps along the elongated directions. To do so, it estimates the centre of the solution space in a warm-up phase and revises the estimated centre continuously with further sampling. The estimated centre is then used to inform the direction of further sampling. The ACHR algorithm, by choosing the optimal direction choice, requires fewer steps to explore the solution space compared to the standard HR. The OPTGP is an improvement of the ACHR as it generates multiple short chains from the estimated centre and picks only the last point of the chain as a sample. In this way, it increases the randomness of exploring the solution space. It allows also for parallel sampling, thereby generating a large number of samples in a shorter time compared to ACHR. Both ACHR and OPTGP are, however, non-Markovian and their convergence is not guaranteed.

Recently, rounding procedures have been proposed to remove the heterogeneity issue of the solution space, thus a modified version of HR is used [17; 40]. The Coordinate Hit-and-Run with rounding (CHRR) [40] is an algorithm in which first the solution space is converted to a rounded closed convex polytope, and then a variant of the HR algorithm is used to sample the rounded polytope. In this algorithm, the Markovian nature of the traditional HR is preserved and convergence of the CHRR is guaranteed.

The ACHR algorithm is available in both the COnstrained Based Reconstruction and Analysis (COBRA) toolbox [82] (in Matlab) and COBRApy (in Python) [20]. The OPTGP is available only in the COBRApy, and CHRR is only provided in the COBRA toolbox. In the Matlab implementations of both ACHR and CHRR, the solution space of the CBM is first reduced using the FVA as described in Sec. 3.1.1; the sampling algorithm is then applied to the reduced solution space.

3.2.2 Stochastic formulation and Monte Carlo sampling algorithms

In the deterministic formulation given by Eq. 3.8, we do not account for the uncertainty of the flux measurements if we fix the fluxes at their measured values, and this might lead to overconfidence in the results. In Van den Meersche et al. [91] the uncertainties corresponding to the experimental values are encoded in a stochastic formulation of a metabolic network model

$$\begin{aligned} \mathbf{S}\mathbf{v}_{ss} &= \mathbf{0} \\ \mathbf{A}\mathbf{v}_{ss} &= \mathbf{b} + \boldsymbol{\epsilon} \\ \mathbf{v}^{lb} &\leq \mathbf{v}_{ss} \leq \mathbf{v}^{ub} \end{aligned} \tag{3.9}$$

where the uncertainties of the flux data are encoded thorough considering the noise term $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma)$. The diagonal matrix $\Sigma = \mathtt{diag}(\sigma_1, \ldots, \sigma_n)$ represents the variances of

flux data. The equations in Eq. 3.9 encode an exact steady state assumption, bounded fluxes, and flux observations with related experimental noise (Fig. 3.1 d). In Van den Meersche et al. [91], a variant of the Metropolis algorithm [79], known as mirror algorithm, has been used to sample from a proposed *truncated multivariate normal* (TMVN) distribution. Van den Meersche et al. [91] provided a function in R [87] to produce a set of samples of fluxes $\mathbf{v}$ in this framework.

Another assumption that can be relaxed in Eq. 3.9 is the exact steady state assumption to agree with the stochastic nature of a cell. In 2016, it was shown that metabolites can accumulate or deplete in a metabolic network [66], and recently MacGillivray et al. [57] studied metabolic networks under the relaxed steady state assumption through the so-called RAMP model. In 2019, Heinonen et al. [41] presented a stochastic formulation of a metabolic network model as

$$\begin{aligned} S\mathbf{v}_{ss} &= \mathbf{0} + \boldsymbol{\beta} \\ A\mathbf{v}_{ss} &= \mathbf{b} + \boldsymbol{\epsilon} \\ \mathbf{v}^{lb} &\leq \mathbf{v}_{ss} \leq \mathbf{v}^{ub} \end{aligned} \tag{3.10}$$

where $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \Gamma)$ is a vector of disturbances around the steady state assumption $S\mathbf{v}_{ss} = \mathbf{0}$. The equations in Eq. 3.10 encode a relaxed steady state assumption, bounded fluxes, and flux observations with related experimental noise (Fig. 3.1 e).

In Heinonen et al. [41], a Bayesian flux model is formulated based on the Bayes rule to study the probability distribution of the fluxes. This framework is called the *Bayesian metabolic flux analysis* (BMFA). Currently available information about the fluxes are represented as a *prior* probability density, $\mathrm{p}(\mathbf{v}_{ss})$. The available measurements $\mathbf{b}$ for the fluxes are linked to the fluxes $\mathbf{v}_{ss}$ from the model via a *likelihood*, $\mathrm{p}(\mathbf{b} \mid \mathbf{v}_{ss})$, which is the conditional density of $\mathbf{b}$ given $\mathbf{v}_{ss}$. The BMFA makes an inference about $\mathbf{v}_{ss}$ through the posterior distribution of fluxes obtained via Bayes formula

$$\mathrm{p}(\mathbf{v}_{ss} \mid \mathbf{b}) \propto \mathrm{p}(\mathbf{b} \mid \mathbf{v}_{ss})\mathrm{p}(\mathbf{v}_{ss}). \tag{3.11}$$

Applying this rule, a TMVN posterior distribution for the fluxes has been presented, and the Gibbs algorithm [34] has been used to sample the posterior distribution. The BMFA is implemented in the COBRA toolbox by Heinonen et al. [41].

3.2.3 A comparison of sampling algorithms in paper A

In paper A [25], both available deterministic and stochastic frameworks are reviewed for the first time, and a rigorous comparison of the corresponding sampling algorithms has been performed. A description of six sampling algorithms HR, ACHR, OPTGP, CHRR, mirror and Gibbs is provided. The first four algorithms are suitable to sample the polytope formed by the deterministic formulation in Eq. 3.8, and the last two algorithms are appropriate to sample from the posterior distribution of the fluxes modelled by the stochastic formulations in Eqs. 3.9 and 3.10.

The sampling algorithms ACHR, OPTGP, CHRR and Gibbs are applied on ten genome scale metabolic networks that were obtained from the BiGG database [52]. The 20,000 samples were generated for each flux in each model, with a thinning parameter of 1000 in each sampling algorithm. A sample generated by a MCMC

algorithm is guaranteed to be representative of the true flux distribution only if the sample chain defined as $v^{(1)}, \ldots, v^{(20000)}$ has converged (in distribution). We have assessed the convergence of each algorithm by computing the percentage of flux chains that passed four different convergence diagnostics: the Raftery and Lewis diagnostic [73], Geweke test [35], IPSRF test [11] and Hellinger distance test [10]. The results from four convergence diagnostics across four algorithms and ten models are shown in Fig. 3.2. We see that the CHRR algorithm is the method with the least convergence problems, while the ACHR method had convergence problems for many models.

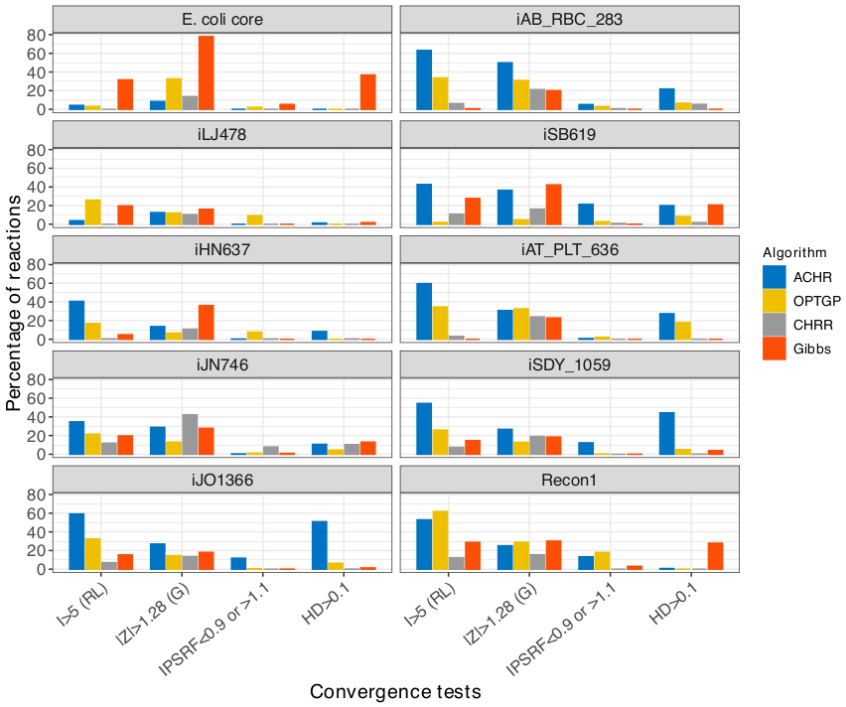


Figure 3.2: Fig. 2 of paper A. Four convergence diagnostics across four algorithms and ten models. The vertical axis shows the proportions of flux chains in each model rejected by the different convergence tests: Raftery and Lewis (RL), Geweke (G), IPSRF and Hellinger distance (HD) on the horizontal axis.

Convergence for individual flux chains has also been inspected, and the result shows that it is not necessarily the same flux chains that failed to converge according to the different diagnostics. Therefore, it is suggested that a combination of convergence diagnostics be applied to make a certain decision about sampling convergence.

We compared the CHRR against each of three other algorithms in terms of sample means and standard deviations, and the result is shown in Fig. 3.3. The reason for selecting the CHRR algorithm as the reference algorithm is that the CHRR algorithm has a Markovian nature and its convergence is guaranteed. The Pearson correlation [18] of the mean and the standard deviation between the two algorithms are calculated and reported in Fig. 3.3. In general, the four algorithms returned very similar sample means and standard deviations, as can be seen from the fact that the points in the plot lie along

the identity line. The Gibbs sampler deviated the most from CHRR, which probably is due to the informative prior distribution imposed on some of the models.

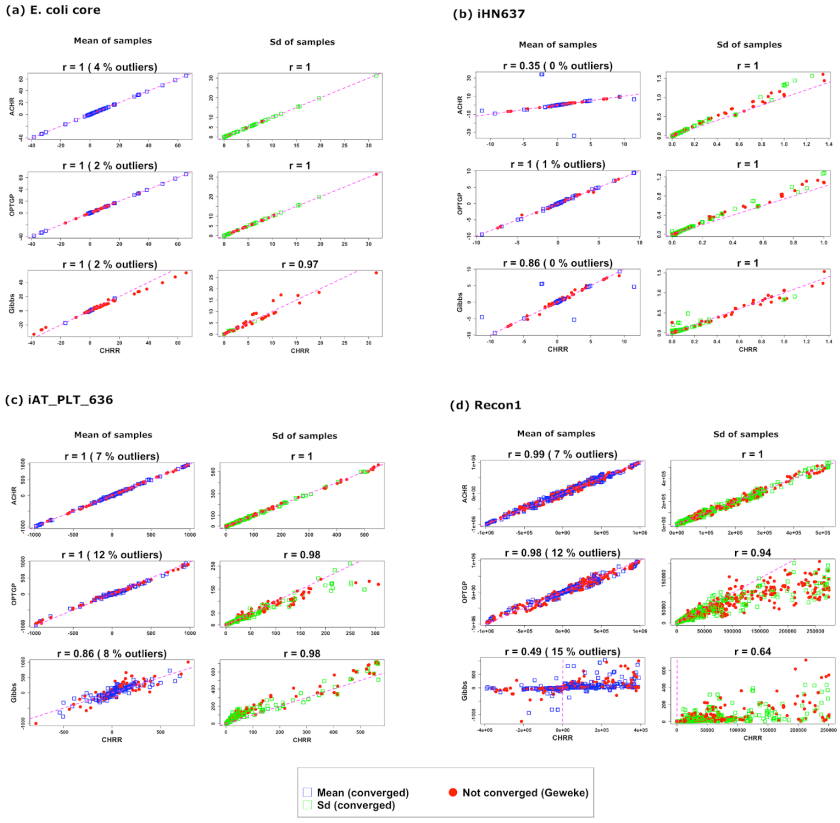


Figure 3.3: Fig. 3 of paper A. Scatter plot of sample means (blue) and standard deviations (green) for ACHR, OPTGP and the Gibbs sampler (vertical axis) against CHRR (horizontal axis) for four models. The Pearson correlation is shown on top of each scatter plot, and the values shown in parenthesis are the proportion of removed outliers. The sample means and standard deviations marked in red correspond to the reactions for which at least one of the two algorithms in a comparison failed the Geweke test.

We define the efficiency of each algorithm in generating independent samples per time unit for each individual flux as

$$\mathrm{E} = \frac{\mathrm{ESS}}{\text{Run time}} \tag{3.12}$$

where the effective sample size (ESS) is the number of independent draws from the flux target distribution [33]. The sampling algorithms are compared in terms of sampling efficiency by calculating E in Eq. 3.12. Fig. 3.4 shows the cumulative distribution functions for the efficiency measure E of the different metabolic models separately for each sampler. It indicates that the CHRR method outperforms the three other algorithms by generating the highest number of independent samples per time unit for each flux.

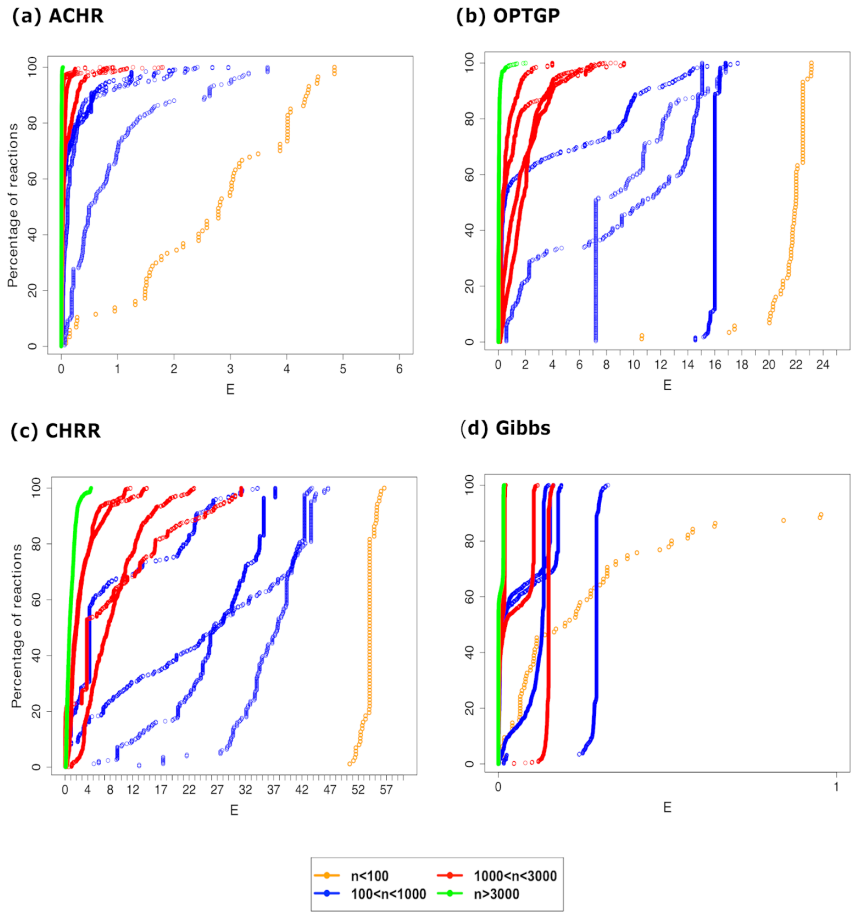


Figure 3.4: Fig. 6 of paper A. Assessing sampling efficiency across four algorithms and ten models. The vertical axis shows the proportions of reactions being less than a given value of the efficiency measure E on the horizontal axis. The ten different curves represent the ten models which are categorized in four groups according to their number of reactions (see legend).

The mirror algorithm used by Van den Meersche et al. [91] for the stochastic formulation in Eq. 3.9 was attempted on the ten models, but we were only able to successfully run it for the E. coli core model. In the mirror algorithm, there is a hyper parameter that defines the jump length in the Metropolis Hasting algorithm to propose a new sample. A typical feature of a genome scale metabolic network is the large variation in flux ranges. Therefore, a small jump length is needed for the fluxes with a small range, and a large jump length is needed for the fluxes with a large range. The best choice of jump lengths as a hyper parameter in the mirror algorithm was not trivial, as a large step length made the sampling algorithm very inefficient and a small jump length led to almost no movement from the initial flux vector. Thus an efficient sampling algorithm that allows the stochastic formulation in Eq. 3.9 to be applied at the genome scale is lacking.

Chapter 4

Kinetics and model reduction

In the papers of the thesis, we have not used the FBA approach for modelling; instead, we have used kinetic modelling since we are interested in modelling the time-varying changes in the metabolite concentrations vector $\mathbf{x}$. To do so, we need to know the reaction rates (fluxes), $\mathbf{v}$, which depend on the metabolite concentrations and physico-chemical conditions (e.g. temperature, pH) [48, p. 19]. It is assumed that the physico-chemical environment is fixed, and the fluxes $\mathbf{v}$ can be described by a function of metabolite concentrations $\mathbf{x}$. The dynamic model in Eq. 3.1 can be rewritten as

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}), \quad \text{where} \quad \mathbf{f}(\mathbf{x}) = \mathbf{S}\mathbf{v} \quad \text{and} \quad \mathbf{v} = \mathbf{v}(\mathbf{x}). \tag{4.1}$$

The function $\mathbf{v}(\mathbf{x})$ is called *kinetics*. The dynamical model in Eq. 4.1 is a set of ordinary differential equations that describes the dynamical state of the metabolite concentrations in a metabolic network. The kinetics was studied for the first time in 1864 by Cato M. Guldberg and Peter Waage [39]. They proposed the law of mass action, which has been used extensively to model the reaction rate of an elementary reaction [55]. The law of mass action is that the rate of a chemical reaction is proportional to the product of the concentration of the reactant metabolites. The law of mass action may not be applied directly in the papers since we do not model all the intermediates and co-factors of the chemical reactions. However, it has been used to derive several of the kinetics discussed in the next section. The main reference for this chapter is the textbook "Mathematical modeling in systems biology: an introduction" by Ingalls [48].

4.1 Kinetics

In the dynamical model of a reaction network, we may use different kinetics for various reactions based on the type of reactions. In this section, we provide a detailed description of the most common functions $\mathbf{v}(\mathbf{x})$ as the kinetics of reactions. In papers B, C and D we have used several of them.

4.1.1 Michaelis-Menten kinetics

The majority of reactions that occur in a cell are catalysed by enzymes (known as proteins). The enzymes typically bind to the reactants in a reaction and facilitate their

conversion to the reaction products [48, p. 47]. An enzyme-catalysed reaction with a single-substrate S can be written as

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} P + E \tag{4.2}$$

where E is an enzyme, ES an intermediate and P a product. The kinetics of an enzyme-catalysed reaction is often described by the Michaelis-Menten [49].

Let [S], [E], [ES] and [P] denote the concentrations of S, E, ES and P over time. Applying the law of mass action [55], a dynamic model describing the changes in the concentrations can be formed by a set of ordinary differential equations

$$\frac{d[S]}{dt} = -k_1[S][E] + k_{-1}[ES] \tag{4.3}$$

$$\frac{d[E]}{dt} = -k_1[S][E] + k_{-1}[ES] + k_2[ES] \tag{4.4}$$

$$\frac{d[ES]}{dt} = k_1[S][E] - k_{-1}[ES] - k_2[ES] \tag{4.5}$$

$$\frac{d[P]}{dt} = k_2[ES] \tag{4.6}$$

where k_1, k_{-1} and k_2 are the forward rate constant, backward rate constant and catalytic rate constant, respectively. These constants are called the kinetic parameters, and they must be positive. Adding the Eqs. 4.4 and 4.5 corresponding to the concentrations of E and ES we get

$$\frac{d}{dt}([E] + [ES]) = 0 \tag{4.7}$$

which implies that the total amount of the enzyme is conserved in the reaction. This leads to a conservation law

$$[E] + [ES] = E_T \tag{4.8}$$

where $E_T \in \mathbf{R}_{\geq 0}$ is the total concentration of the enzyme. In the analysis of the system, we further applied the quasi steady state assumption for the intermediate, ES. In the quasi steady state assumption, we assume that all dynamic reactions involving ES occur on the fast time-scale, and the intermediate ES comes rapidly to its steady state concentration. Applying this assumption leads to the algebraic equation

$$k_1[S][E] - k_{-1}[ES]^{qss} - k_2[ES]^{qss} = 0. \tag{4.9}$$

In solving this equation for $[ES]^{qss}$ and substituting [E] by $E_T - [ES]^{qss}$ from Eq. 4.8 we obtain

$$[ES]^{qss} = \frac{E_T[S]}{\frac{k_{-1}+k_2}{k_1} + [S]}. \tag{4.10}$$

Inserting the expression for $[ES]^{qss}$ from Eq. 4.10 in Eq. 4.6, the production rate for the product P of the chemical reaction in Eq. 4.2 becomes

$$\frac{d[P]}{dt} = \frac{k_2 E_T S}{\frac{k_{-1}+k_2}{k_1} + S}. \tag{4.11}$$

The assumptions taken here essentially simplify the chemical reaction in Eq. 4.2 to the reaction

$$S \longrightarrow P. \tag{4.12}$$

The flux ν of this simplified reaction is given by Eq. 4.11. By defining $V_{max} = k_2 E_T$ and $K_M = \frac{k_{-1}+k_2}{k_1}$, the flux ν form Eq. 4.11 can be rewritten as a function of the substrate concentration [S]

$$\nu = \frac{d[P]}{dt} = \frac{V_{max}[S]}{K_M + [S]}. \tag{4.13}$$

This reaction rate is known as the Michaelis-Menten rate law and, it is shown in Fig. 4.1.

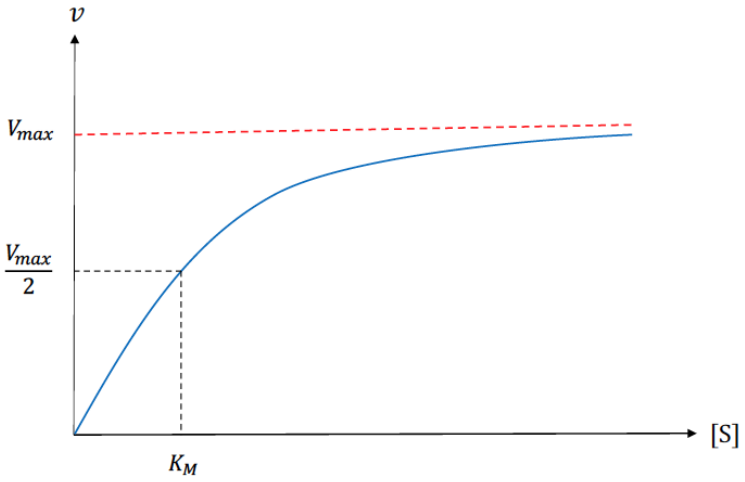


Figure 4.1: Michaelis-Menten kinetics given by Eq. 4.13 for the single substrate reaction given by Eq. 4.12. The flux value ν (vertical axis) is a function of the substrate concentration [S] (horizontal axis). The flux ν equals to $V_{max}/2$ for $[S] = K_M$.

The rate for the consumption of S in the model for the simplified reaction is

$$\frac{d[S]}{dt} = -\frac{V_{max}[S]}{K_M + [S]}. \tag{4.14}$$

In the Michaelis-Menten rate law, the enzyme information is intrinsic in the constants V_{max} and K_M. The V_{max} is the maximal reaction rate (flux) with the following relationship

$$\lim_{[S]\to+\infty} \nu([S]) = V_{max} \tag{4.15}$$

where $\nu([S])$ is the flux given in Eq. 4.13. This implies that the flux ν approaches to the limiting rate V_{max} as the substrate concentration [S] increases. The constant K_M known as the Michaelis constant is the concentration of substrate S at which the flux ν is half of V_{max}. Both Michaelis-Menten kinetics parameters are shown in Fig. 4.1. The V_{max} and K_M values for some reactions can be found in databases such as BRENDA [81]

and SABIO-RK [103]. As discussed in the next chapter, the value of these parameters should in general be estimated having the experimental data of the concentrations or the fluxes.

Several of the reactions in the yeast glycolysis example model of paper B, and the reactions in the fatty acid synthesis model of paper D, are modelled using the original or slightly different version of the Michaelis-Menten kinetics.

4.1.2 Kinetics of competitive inhibition

Enzymes can be regulated by other molecules that may increase or decrease the activity of the enzyme. The molecules that decrease the enzyme's activity are called inhibitors. There are three types of inhibition: competitive inhibition, non-competitive inhibition, and uncompetitive inhibition [48, p. 55-63]. Here we describe the kinetics for a competitively inhibited enzyme. A competitive inhibitor is a molecule that binds to the active site of the enzyme and blocks the binding of the substrate to the enzyme. The reaction scheme, including both the catalytic and inhibition processes, can be described as

$$\begin{aligned} S + E &\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} P + E \\ I + E &\underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} EI \end{aligned} \tag{4.16}$$

where I is the inhibitor and EI is the enzyme-inhibitor complex. Denoting the concentration of I and EI by [I] and [EI], we have

$$\frac{d[S]}{dt} = -k_1[S][E] + k_{-1}[ES] \tag{4.17}$$

$$\frac{d[E]}{dt} = -k_1[S][E] + k_{-1}[ES] + k_2[ES] - k_3[I][E] + k_{-3}[EI] \tag{4.18}$$

$$\frac{d[ES]}{dt} = k_1[S][E] - k_{-1}[ES] - k_2[ES] \tag{4.19}$$

$$\frac{d[P]}{dt} = k_2[ES] \tag{4.20}$$

$$\frac{d[EI]}{dt} = k_3[I][E] - k_{-3}[EI]. \tag{4.21}$$

We treat the concentration [I] as a fixed quantity, and this is justified based on the assumption that the inhibitor, I, is far more abundant than the enzyme such that the formation of EI does not significantly change [I] [48, p. 56]. In addition, the quasi steady state assumption is applied for the intermediate, ES, and the enzyme-inhibitor complex, EI. In the quasi steady state assumption, we assume that all dynamic reactions involving, for example, ES occur on a fast time scale, and the intermediate ES comes rapidly to its steady state concentration. Following this assumption for both ES and EI, we replace the ordinary differential equations given in Eq. 4.19 and 4.21 with algebraic equations

$$k_1[S][E] - k_{-1}[ES]^{qss} - k_2[ES]^{qss} = 0 \tag{4.22}$$

$$k_3[I][E] - k_{-3}[EI]^{qss} = 0. \tag{4.23}$$

Using the algebraic equations in Eqs. 4.22 and 4.23 and the conservation law $[E] + [ES]^{qss} + [EI]^{qss} = E_T$ yields

$$[ES]^{qss} = \frac{E_T[S]}{K_M + \frac{K_M[I]}{K_i} + [S]} \tag{4.24}$$

where $K_M = \frac{k_{-1}+k_2}{k_1}$ is the Michaelis constant and the $K_i = k_{-3}/k_3$ is the dissociation constant for inhibitor binding. Inserting the expression for $[ES]^{qss}$ from Eq. 4.24 in Eq. 4.20, the production rate for the product P of the chemical reaction in Eq. 4.16 is

$$\frac{d[P]}{dt} = \frac{V_{max}[S]}{K_M(1 + \frac{[I]}{K_i}) + [S]} \tag{4.25}$$

where $V_{max} = k_2 E_T$ is the maximal reaction rate.

The obtained reaction rate essentially simplified the chemical reaction of Eq. 4.16 to the reaction given in 4.12. The flux v for the chemical reaction in Eq. 4.12 in the presence of the inhibitor can then be rewritten as

$$v = \frac{d[P]}{dt} = \frac{V_{max}[S]}{K_M(1 + \frac{[I]}{K_i}) + [S]}. \tag{4.26}$$

Note that this rate is the Michaelis-Menten rate law given in Eq. 4.13 for $[I] = 0$. The limiting reaction rate V_{max} is not affected in competitive inhibition as shown in Fig. 4.2. Instead, the inhibitor increases the substrate concentration [S] required to reach the half-maximal rate, as the effective Michaelis constant of the inhibited reaction is $K_M(1 + \frac{[I]}{K_i})$ [48]. All these properties are shown in the plot of the rate law in Fig. 4.2 for zero, low and high concentrations of the inhibitor.

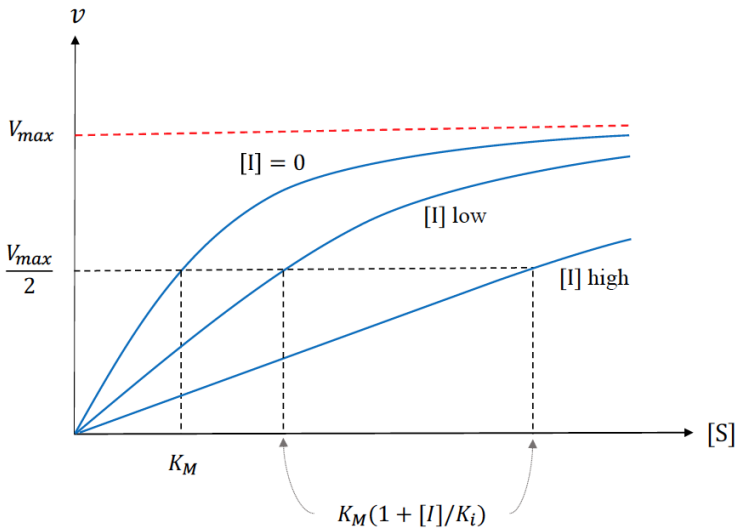


Figure 4.2: Reaction rate law in Eq. 4.26 for an enzyme catalyzed reaction in the presence of a competitive inhibitor, I, given by Eq. 4.16. The flux value v (vertical axis) is a function of the substrate concentration [S] (horizontal axis). The reaction rate is plotted for three different levels of the inhibitor, and it can be seen that the effective Michaelis constant of the inhibited reaction, $K_M(1 + \frac{[I]}{K_i})$*, increases with the inhibitor concentration [I].*

In paper D, some of the reactions of the fatty acid synthesis pathway are competitively inhibited enzyme catalysed reactions, and their rates are modelled using the rate law described by Eq. 4.26.

4.1.3 Hill kinetics

A generalised form of the Michaelis-Menten kinetics is the Hill kinetics proposed in 1910 by A.V. Hill [44] as

$$v = \frac{d[P]}{dt} = \frac{V_{max}[S]^n}{(K_A)^n + [S]^n} \tag{4.27}$$

where V_{max}, K_A and n are positive constants. The Hill kinetics has the same property as the Michaelis-Menten kinetics as for all n

$$\lim_{[S]\to+\infty} v([S]) = V_{max}. \tag{4.28}$$

The constant K_A takes the role of the K_M in the Michaelis-Menten kinetics since for $[S] = K_A$ we get

$$v(K_A) = \frac{V_{max}}{2} \tag{4.29}$$

for all n. The dimensionless form of the Hill kinetics can be written as

$$\frac{v}{V_{max}} = \frac{(\frac{[S]}{K_A})^n}{1 + (\frac{[S]}{K_A})^n}. \tag{4.30}$$

This is sketched in Fig. 4.3 for four different values of n. Note that the Hill kinetics for $n = 1$ gives the Michaelis-Menten kinetics.

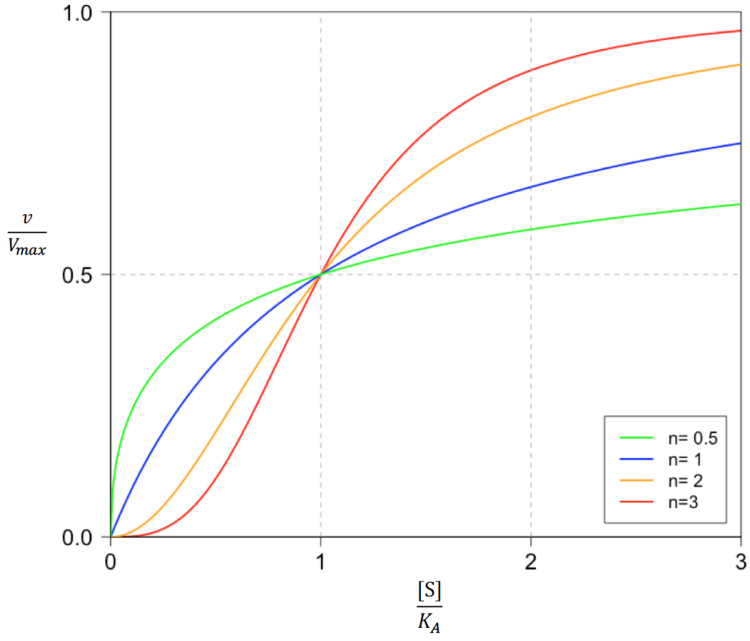


Figure 4.3: The flux v as a function of the substrate concentration $[S]$ for the non-dimensional Hill kinetics given by Eq. 4.30 and four different n. The flux v equals to $V_{max}/2$ for $[S] = K_A$ and all n.

A slightly modified version of the Hill kinetics is used to model the rate law for a reaction in the model of paper C, which describes the interaction between the two chemicals β-naphthoflavone (BNF) and nocodazole (NOC) and their influence on cytochrome P450 1A (CYP1A) enzymes. We have also applied the Hill kinetics to describe the rate laws of several reactions in the fatty acid synthesis model of paper D.

4.1.4 Zero, first and second order kinetics

The last kinetics functions that we discuss in the thesis are the zero, first and second order kinetics. These kinetics functions are used to model some fluxes in papers C and D, and they are also used in the small example model in paper B. The first order kinetics is

$$v = k \tag{4.31}$$

where k is a constant. In this type of kinetics, the flux is a constant that does not change with time. The first order kinetics can be used for the saturated fluxes or entering fluxes to a reaction network from the outside. The first order kinetics can be described by

$$v = kx \tag{4.32}$$

where k is a constant and x is the substrate concentration. In the first order kinetics, the flux v is a linear function of the substrate concentration x. This type of kinetics

is common to use when the substrate X is far from being saturated and the limiting factor of the reaction. This case can be seen in Fig. 4.1 for the Michaelis-Menten kinetics, where for low substrate concentrations the flux is a linear function of the substrate concentrations and can be approximated by first order kinetics in the linear regime. Using first order kinetics decreases the number of parameters as it only has one parameter, while the Michaelis-Menten kinetics in Eq. 4.13 has two parameters.

The common form of the second order kinetics is

$$v = kx_1x_2 \tag{4.33}$$

where k is a constant and x_1 and x_2 are the concentrations of the two substrates X_1 and X_2. In the second order kinetics, the flux is proportional to the product of the two concentrations, which obeys the mass action law.

The parameters k in the Eqs. 4.31, 4.32 and 4.33 are called *kinetic parameters*. In paper C, we estimate the first and second order kinetic parameters in a model of the interaction between two chemicals BNF and NOC and their influence on CYP1A enzymes using real data from two experiments [37; 61]. More about the parameter estimation is provided in the next chapter.

4.2 Model reduction

The dynamic model in the form of Eq. 4.1 incorporates a high level of detail about the chemical reaction network compared to the CBM approach discussed in chapter 3. This type of modelling approach involves both the network stoichiometry and kinetic parameters. The degree of the model complexity can easily increase if we model each reaction with all its characteristics, especially when we use advanced kinetics to describe the reaction rates. A complex model of a biochemical reaction network typically has a large number of parameters. Moreover, a small number of the kinetic parameters can usually be identified from the literature, while the remaining parameters should be estimated. To do so, large datasets including the metabolite concentrations are required, and often not all the metabolite concentrations can be measured. Therefore, it might be beneficial to apply some model reduction techniques to reduce the model to a simplified model that can still describe the dynamical features of the network. To avoid oversimplification, a model reduction technique should reduce the model in a controlled manner by requiring that the output of the reduced model and full (original) model be sufficiently close using a defined measure.

In the dynamic model given by Eq. 4.1, the metabolite concentrations, reactions and kinetic parameters are involved. A common model reduction technique is the model order reduction approach, which reduces the dimension of the concentration vector $\mathbf{x}$ as the state space. For a general introduction about the model order reduction approach, see e.g. Schilders et al. [83]. In paper D, we have not included the concentration of the co-factors and intermediates of the reactions, and this can actually be viewed as a model order reduction. Leaving the co-factors and intermediates of the reactions out of the model is a model order reduction discussed in Feliu and Wiuf [27]. Removing the intermediates from the model also leads to a simplification in the reaction kinetics. Applying simplified kinetics such as zero, first and second order kinetics, as discussed

in Sec. 4.1.4, can also be seen as a model reduction technique that can be used to decrease the number of kinetic parameters.

Consider we have a dynamic model of a chemical reaction network given by Eq. 4.1 and want to reduce the model complexity. There are different approaches to reducing the complexity of the model, with the most common ones being lumping, sensitivity analysis and timescale analysis [63; 72; 84]. Here we focus on the model reduction approach suggested by Rao et al. [76] as paper B in the thesis is an extension of this approach. The approach in Rao et al. [76] is applicable for a large class of the kinetics, including the ones discussed in Sec. 4.1.

The model reduction procedure in Rao et al. [76] includes a method to reduce the model and a measure to compare the reduced and full model. First, it specifies a set $\mathcal{M}_I$ of compounds that are important and should be preserved during the model reduction. The important compounds can, for instance, be the ones that are possible to measure experimentally. Then all the complexes (see Sec. 2.2) of the reaction network are classified into two groups. One category contains all the complexes that include at least one of the compounds in the set $\mathcal{M}_I$, and those complexes will not be subjected to reduction. The second category is composed of the complexes considered for reduction since they do not contain any of the important compounds in $\mathcal{M}_I$. The reduction is then based on the assumption that the model reaches an asymptotically stable steady state that can be obtained by integrating the system for a sufficient amount of time. The complex is reduced by fixing its concentration to the corresponding steady state value of the full model, which can be done for any number of complexes. Having the possibility to reduce any given set of complexes, an iterative method for selecting the complexes for reduction is presented in Rao et al. [76]. The method reduces one complex at a time by choosing the one that leads to the smallest error defined as

$$I_T(\mathbf{x}_r, \mathbf{x}_f) = \sum_{i \in \mathcal{M}_I} \frac{1}{Tn(\mathcal{M}_I)} \int_0^T \left| 1 - \frac{x_{ir}(t)}{x_{if}(t)} \right| dt \tag{4.34}$$

where T is the time interval, $n(\mathcal{M}_I)$ is the number of compounds in $\mathcal{M}_I$, x_{ir} and x_{if} are the concentrations of compound number i in the reduced and full model, respectively. The vector of the compound concentrations in the reduced and full model are denoted by the vectors $\mathbf{x}_r$ and $\mathbf{x}_f$, respectively. This error integral gives the average relative difference between the concentration of all compounds in $\mathcal{M}_I$ in the full and reduced model over the given time interval $[0, T]$. Ultimately, the method stops when an error threshold is reached.

The model reduction method presented in Rao et al. [76] computes the error integral in Eq. 4.34 for a single set of parameters and does not account for parameter uncertainty. In paper B [31] we present a new way to evaluate model reductions that takes parameter uncertainty into account. We encode the parameter uncertainty by assuming that we have a given number of parameter sets instead of a single parameter set. Then all possible reduced models are obtained for all parameter sets. Based on Eq. 4.34 we define the symmetric error measure

$$E_T(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2} \left(I_T(\mathbf{x}_1, \mathbf{x}_2) + I_T(\mathbf{x}_2, \mathbf{x}_1) \right). \tag{4.35}$$

For any two models, we can compute this measure without favouring one of them.

Using the measure in Eq. 4.35 as the dissimilarity measure, the full (original) model and reduced models for all the parameter sets are clustered using the single linkage clustering. A dendrogram [46], like the ones in Fig. 4.5, is then used to visualise the clustering results. Distribution throughout the dendrogram is then obtained for a given reduction considering different parameter sets, and this is visualised by colouring. Furthermore, a reduction is evaluated through the test statistics of a Kolmogorov-Smirnov test [15] for the distribution of the reduced and full model. The reductions with a test score lower than a threshold are considered as consistent with the full model. Among the consistent models, the best reduced model is the one that uses the most reduction.

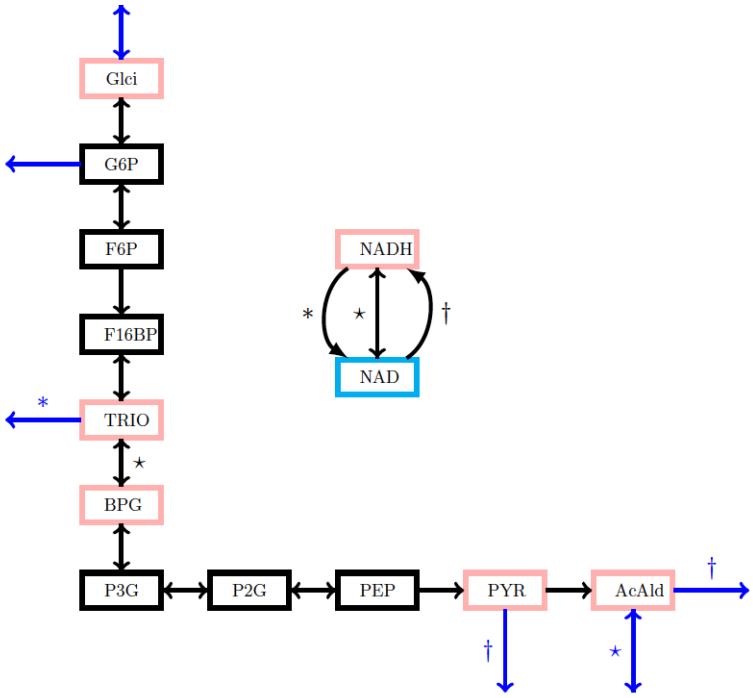


Figure 4.4: Fig. 2 of paper B, the yeast glycolysis reaction network from Rao et al. [76]. The nodes are the compounds and the arrows are the reactions. The important compounds that are not subjected to reduction are specified by pink rectangles and the candidate compounds for reduction are specified by the black rectangles. See the original figure in paper B for a detailed description of the figure.

The suggested method in paper B has been tested on a kinetic model of the yeast glycolysis reaction network shown in Fig. 4.4, which is taken from Rao et al. [76]. The resulting dendrograms are shown in Fig. 4.5, where in each panel we visualise the result for a sample of 100 parameter sets using a log-normal distribution. The reference value of each parameter has been used as the expectation of the parameter, and the reference value is further divided by a scaling factor to define the deviation. It can be seen that for high parameter uncertainty (scaling 3) all 100 full models are evenly distributed between the reduced models. This confirms that when the parameters in the

model are highly uncertain, the model can be reduced maximally without increasing the uncertainty in the model dynamics, whereas for low parameter uncertainty (scaling 100) all the full models are clustered tightly implying that the full model should not be much reduced. This analysis shows that the parameter uncertainty plays an important role in the model reduction to obtain a suitable model complexity in balance with the parameter uncertainty.

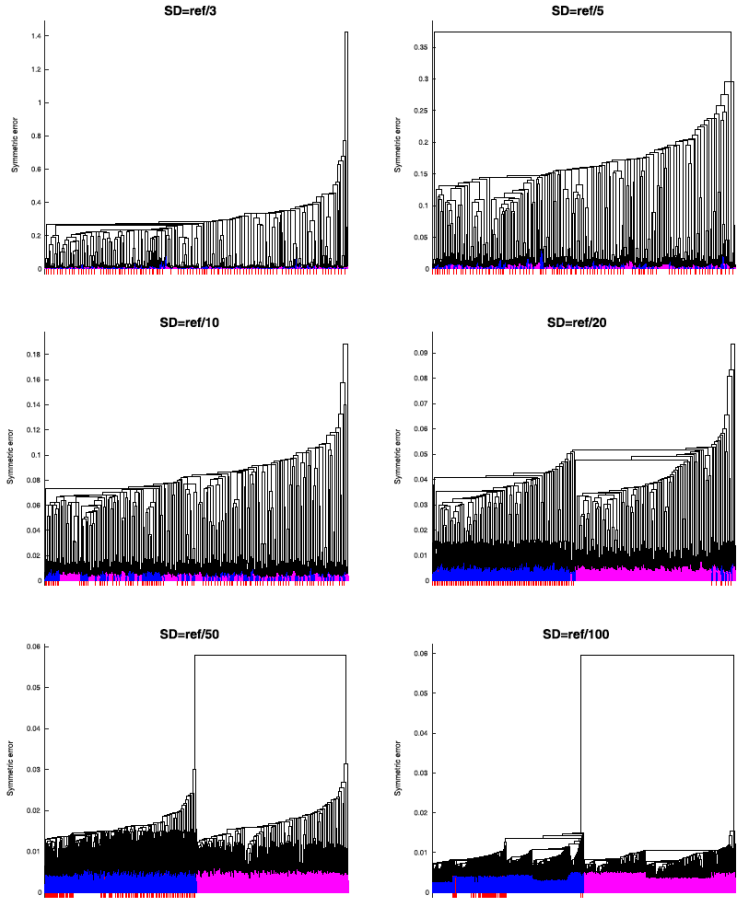


Figure 4.5: Figure 5 of paper B. Dendrogram from single linkage clustering of all the model reductions for the example network in Fig 4.4. The 100 parameter sets were sampled from a log-normal distribution with the expectation equals the reference value of the parameter (denoted by ref in the plot) and the standard deviation equals to the reference value divided by six different scaling factors given on top of each panel. The full models are colored in red and the models where F16BP was reduced are given in purple and all other modes are blue.

As mentioned earlier, the method by Rao et al. [76] reduces one complex at a time in the model reduction procedure. In paper B, we further show that, for a small example network, a smaller error might be obtained by reducing two complexes at the same time.

Chapter 5

Parameter estimation, sensitivity and identifiability analysis

The dynamic model in Eq. 4.1 for a chemical network is constructed through defining the kinetics $\mathbf{v}(\mathbf{x})$. Some of this kinetics are discussed in the previous chapter, where we observed that each kinetics has one or more than one kinetic parameter. Therefore, the dynamic model given by Eq. 4.1 is dependent on the kinetic parameters, and rewriting the model as follows shows this dependency explicitly

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}, \boldsymbol{\theta}) \tag{5.1}$$

where $\boldsymbol{\theta}$ is a vector of the kinetic parameters. The value of the kinetic parameters for some reactions in a controlled experiment can be obtained from the BRENDA [81] and SABIO-RK [103]. However, kinetic parameters must in general be estimated if a dynamic model of a chemical reaction network has been made. In paper C we construct a dynamic model of a given reaction network and estimate the kinetic parameters.

The typical experimental data is the concentration measurements for some of the metabolites in the network. The distance between the simulated concentration of those metabolites by the model and their experimental value is calculated using a cost function. The cost function is a function of the parameters $\boldsymbol{\theta}$ denoted by $\psi(\boldsymbol{\theta})$. The parameters in the model can then be estimated by fitting the model to the data by minimising $\psi(\boldsymbol{\theta})$. A common cost function ψ is a weighted sum-of-squares error [86]; however, it can take other forms, see e.g. Zheng et al. [107]. We start the chapter by describing the parameter estimation of a proposed model in paper C.

5.1 Parameter estimation in paper C

In paper C, we construct a dynamic model for a pathway shown in Fig. 5.1. In this pathway, we illustrate how the chemicals β-naphthoflavone (BNF) and nocodazole (NOC) activate the aryl hydrocarbon receptor (AhR), which results in the induction of cytochrome P450 1A (CYP1A) enzymes' synthesis. Both BNF and NOC bind to the CYP1A enzymes, where BNF is metabolised by CYP1A enzymes and NOC inhibits CYP1A enzymes. Thus, chemical NOC occupies the CYP1A enzymes in the elimination pathway, which delays the biotransformation of BNF.

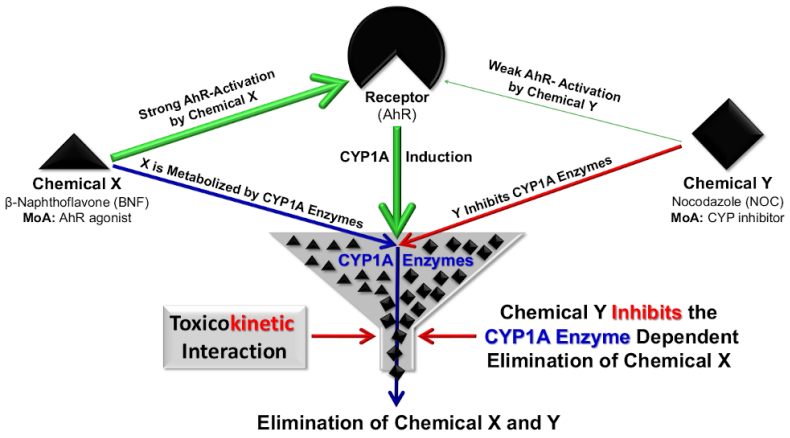


Figure 5.1: Figure 1 of paper C. The pathway to illustrate the toxicokinetic interaction between chemical X (β-naphthoflavone (BNF) black triangles) and chemical Y (nocodazole (NOC) black squares) and their effect on the aryl hydrocarbon receptor (AhR) which leads to induction of the cytochrome P450 1A (CYP1A) enzymes. For a more detailed description of the pathway see the original figure in paper C.

The reactions associated with the pathway in Fig. 5.1 can be presented in the form of Eq. 2.1 as

$$\begin{aligned}
R_1: &\quad BNF + CYP1A \longrightarrow CYP1A_{OBNF} \\
R_3: &\quad NOC + CYP1A \longrightarrow CYP1A_{ONOC} \\
R_2: &\quad CYP1A_{OBNF} \longrightarrow CYP1A + BNF_b \\
R_4: &\quad CYP1A_{ONOC} \longrightarrow CYP1A + NOC_b \\
R_5: &\quad CYP1A \longrightarrow
\end{aligned} \tag{5.2}$$

where $CYP1A_{OBNF}$ and $CYP1A_{ONOC}$ are the CYP1A enzymes occupied by BNF and NOC molecules, respectively. The BNF_b and NOC_b are the biotransformed BNF and NOC molecules, respectively.

A dynamic model of the chemical reactions given by Eq. 5.2 is proposed in paper C to describe the changes in the concentrations of BNF, NOC, free CYP1A enzymes and CYP1A enzymes occupied by BNF or NOC. Note that as the changes in the concentrations of the biotransformed BNF and NOC were not our prime interest; we have not included them in the model. To construct the model, we have used the mass action law, first order kinetics and a modified version of the Hill kinetics discussed in Sec. 4.1. See the manuscript for more about the model equations.

The data used in paper C were obtained from a previous study using PLHC-1 cells that had been treated with the carrier vehicle and different doses of BNF (0.1 and 1 μM) and NOC (1, 10 and 25 μM), alone or mixed, and measured at five different times (6, 12, 24, 48 and 72h) [37; 61]. The measurements for the free CYP1A enzymes were the diagnostic ethoxyresorufin-O-deethylase (EROD) activity, which is assumed to be proportional to the concentration of free CYP1A enzymes with a proportionality constant equal to one. The measurements from various treatments are shown in Fig. 5.2.

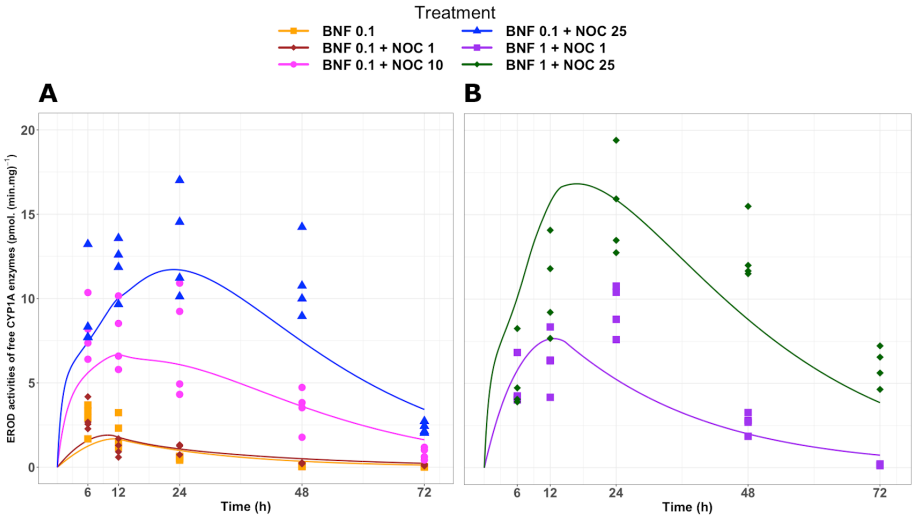


Figure 5.2: Figure 2 of paper C. Fitted values for EROD activity of free CYP1A enzymes. The solid line shows the EROD activity simulated from the model fitted to the measured EROD activity in six different treatments. For a detailed description of the figure and different treatments see the original figure in paper C.

The parameters of the proposed model are estimated by fitting the simulated concentration of the free CYP1A enzymes from the model against its observations represented by EROD data (Fig 5.2). The estimated values for the parameters were those values that minimise a squared logarithmic error (SLE) [89] between the concentration of the free CYP1A enzymes from the model and its observations. This cost function has been used as we did not want large errors to be significantly more penalised than small ones.

In Fig. 5.2, we see that the model captured the dynamic behaviour of the EROD activities of the free CYP1A enzymes. The model could also describe changes in the concentrations of BNF, NOC and CYP1A enzymes occupied by BNF or NOC qualitatively. We refer you to the paper C to see the plots. The model can be further refined if the data for the concentration of more compounds in the pathway are available. This motivates the need to design the experiments that provide time-course measurements of the concentrations of BNF and/or NOC.

5.2 Global sensitivity analysis (GSA)

When developing a dynamical model in the form of Eq. 4.1 for a reaction network, a real challenge could be the lack of experimental data. In this case, the value of kinetic parameters, $\boldsymbol{\theta}$, are usually obtained from existing databases for kinetic parameters or extracted from the literature directly. Typically, a large variation can be seen in the value of a single parameter from different sources. Large variations in the value of parameters might lead to significant uncertainty in the model output. The Global Sensitivity Analysis (GSA) is a tool to describe how the uncertainty in the model parameters can influence the uncertainty in the model output. The most influential parameters on the model output can be identified through the GSA, and we can try to increase the accuracy

in the value of these parameters to reduce the model uncertainty. The GSA can also provide information about the non-influential parameters on the model output to set them at small values or simplify the model by excluding them [108]. There are various types of methods to perform the GSA for a dynamic model, see e.g. Pianosi et al. [71].

One of the most established GSA methods is the method of Morris [62], also known as the elementary effect test (EET) [80]. Let $h(\boldsymbol{\theta})$ be the function for which we assess the GSA. In the EET, the mean of N_s local changes is taken as the measure of the global sensitivity of h to changes in the value of i^{th} parameter, θ_i, i.e.

$$\begin{aligned} s_i^G &= \frac{1}{N_s}\sum_{j=1}^{N_s}\left|EE^j\right| \\ &= \frac{1}{N_s}\sum_{j=1}^{N_s}\left|\frac{h(\theta_1{}^j,\ldots,\theta_i{}^j+\Delta_i^j,\ldots,\theta_M{}^j)-h(\theta_1{}^j,\ldots,\theta_i{}^j,\ldots,\theta_M{}^j)}{\Delta_i^j}\right| r_i, \end{aligned} \tag{5.3}$$

where r_i is a scaling factor facilitating a comparison of sensitivities across parameters that may have different units and M is the number of parameters. The sensitivity index s_i^G indicates the sensitivity of h to parameter θ_i considering the direct influence of this parameter. The sensitivity index s_i^G can often times not be computed analytically, and it must be approximated numerically from a sample of the parameter values and the associated model output evaluation. Most of the GSA methods are sample-based methods and require specification concerning the parameters' variability space by providing an interval for each parameter.

There are several sampling strategies that generate samples of the parameter values, see e.g. Campolongo et al. [14]. Here we describe the radial sampling strategy briefly as it has been used in paper D. In radial sampling $2N_s$ points are first sampled from the parameter variability space using the Latin-Hypercube sampling [59]. Then this set of samples will be divided into two sets, where one set includes N_s baseline points and the other includes the N_s auxiliary point. Let us assume $\mathbf{a} = \begin{bmatrix} a_1 & \ldots & a_M \end{bmatrix}$ is a baseline vector and $\mathbf{b} = \begin{bmatrix} b_1 & \ldots & b_M \end{bmatrix}$ is an auxiliary vector. The value of the first parameter in the vector **a** is replaced by the corresponding value from the vector **b** while keeping the rest of the parameters fixed. We repeat this for all parameters such that after each step we go back to the original vector **a** and then change the value of the next parameter [14].

A variant of GSA known as the regional sensitivity analysis (RSA) is a family of methods aimed at identifying regions of the parameter variability space leading to special behaviour in the model output. The idea of the RSA was first proposed by Young et al. [106] and Spear and Hornberger [85]. The RSA method applies the All-At-a-Time (AAT) sampling strategy by simultaneously varying the value of the parameters. To do so, the Latin-Hypercube sampling method [59] is used. Then the generated samples are divided into two groups depending on whether the associated model simulation exhibits the expected behaviour. There are several ways to use those two sets of parameters and make an inference about the regions, which gives us the desired behaviour in the model simulations, see e.g. [71].

5.2.1 Global sensitivity analysis in paper C

For each parameter in the model of paper C, a random sample of 1000 values was first drawn using a log-uniform distribution on the interval from the parameter estimated value divided by 10 to the estimated value multiplied by 10. Then we used the generated values for each parameter to vary it while fixing the remaining parameters at their nominal values. The model described by a set of ODEs is numerically solved for each parameter set using the *ode* function in R [87]. The global sensitivity of the free CYP1A enzymes' EROD activity as the model output is then qualitatively inspected. This has been done by visualising the sensitivity ranges of the EROD activities over time according to changes in the parameter values. We found that to reduce the total uncertainty in the model predictions for the EROD activity of CYP1A enzymes, the parameter corresponding to the rate of CYP1A enzyme degradation should be set at a value of high accuracy. This parameter is associated with the reaction R_5 in Eq. 5.2.

5.2.2 Global sensitivity analysis in paper D

There has been limited modeling of the fatty acid synthesis pathway. However, in 2018, Berndt et al. [7] proposed a dynamic model of the complete central metabolism of liver cells named as HEPATOKIN1, where one of the sub-pathways is fatty acid synthesis. The HEPATOKIN1 is constructed by applying the complex kinetics describing the reaction rates. This led to an extensive model with 1040 kinetic parameters where 272 of them are estimated using the experimental data while all other parameters are fixed at values obtained from the literature. In paper D, we propose a dynamic model of the form of Eq. 4.1 for a simplified version of the fatty acid synthesis' full pathway shown in Fig. 1.1 [3]. The pathway is simplified by lumping the seven cyclic reactions in the full network, and the simplified version is shown in Fig. 5.3.

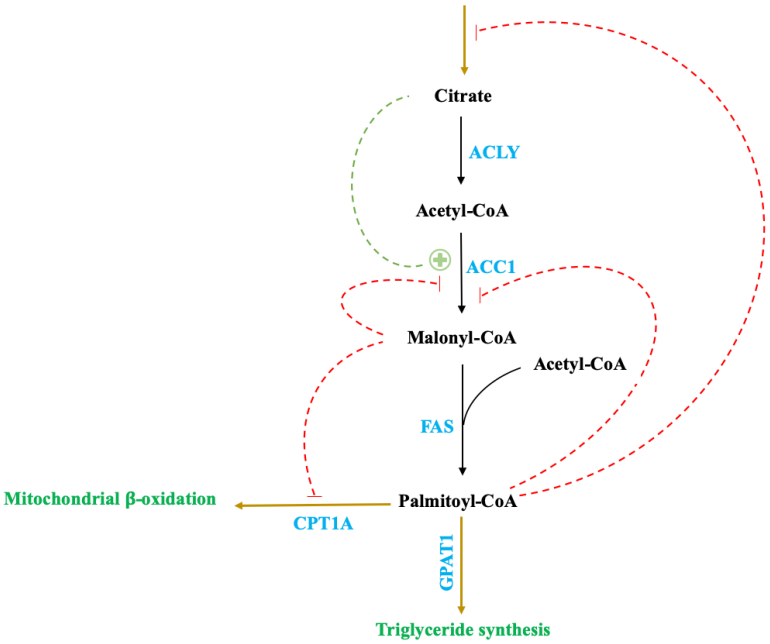


Figure 5.3: Fig. 1 of paper D. A simplified reaction network of fatty acid synthesis with regulatory interactions (dashed lines). The metabolites are citrate, acetyl-CoA, malonyl-CoA, and palmitoyl-CoA. The reactions between the metabolites and exchange reactions are shown by arrows. The enzymes that catalyze the reactions are ACLY, ACC1, FAS, CPT1A and GPAT1 which are marked in blue. For a detailed description of the figure see the original figure in paper D.

In the proposed model, we include the regulatory activities such as enzyme inhibition and activation and have mostly used the Michaelis-Menten kinetics, kinetics for competitive inhibition and Hill kinetics discussed in Sec. 4.1. The proposed model has twenty-one parameters where fourteen of them have been set to the values obtained from the literature and seven must be estimated from the experimental data. To do so we require the time-course data of the concentration of the metabolites included in the model. To our knowledge, this data is not available in the literature. Therefore, our model is a potential model for a fatty acid synthesis network, which must be calibrated to the data once it is available.

In paper D, we have defined the parameter variability space by considering eight different intervals $[0, 0.1]$, $[0, 0.5]$, $[0, 1]$, $[0, 5]$, $[0, 10]$, $[0, 50]$, $[0, 100]$ and $[0, 500]$ for all seven parameters. Trying different intervals, the RSA has been applied to identify the parameter regions in which the associated metabolite concentrations simulated by the model have a higher chance to reach the steady state as the desired behaviour. Based on the RSA, we obtained the intervals from which a modeller should select the initial guess for the parameters in the model fitting process. This will lower down the probability of not reaching the steady state phase by the concentrations of the metabolites.

We further applied the EET method [80] to measure the sensitivity of the metabolite concentrations simulated by the model to changes in the value of the parameters. The sensitivity indices are computed through Eq. 5.3 and the function $\hbar$ is defined as the

metabolite concentrations simulated by the model. The parameter values vary over the intervals obtained from the RSA. By identifying the metabolites for which their concentrations show high sensitivities to change in the value of most parameters, we suggest an experimental design to acquire the most informative data. This data is the concentration measurements of citrate, malonyl-CoA and palmitoyl-CoA to increase the likelihood of gaining a unique value for most of parameters by fitting the model to the data.

5.3 Identifiability and local sensitivity analysis

5.3.1 Structural identifiability

When doing statistical inference to estimate the model parameters, an important question is whether the true parameter values can be determined. To answer this question, we should first assess a model property known as the structural identifiability. This is a property of the model alone and can in principle be assessed before collecting experimental data. The models often contain parameters that are structurally unidentifiable and none of the estimation algorithms succeeds in estimating those parameters. Therefore, the structural identifiability of a model is essential to be assessed and avoid wasting time and resources [98]. Estimating parameters in a model using experimental data only makes sense if the model is structurally identifiable.

Here we use the definition of the structural identifiability presented in the study by Villaverde et al. [99]. For a model of the form of Eq. 5.1 assume that $\mathbf{y} \in \mathbb{R}^{N_y}$ is a vector of the quantities that can be measured experimentally, defined as

$$\mathbf{y}(t) = \mathbf{g}(\mathbf{x}(t), \boldsymbol{\theta}) \tag{5.4}$$

where function $\mathbf{g}$ maps the model variables $\mathbf{x}$ to the measurable output variables $\mathbf{y}$ and is also a function of the parameters $\boldsymbol{\theta}$. The model parameter θ_i is called a structurally locally identifiable parameter if for almost any parameter vector $\boldsymbol{\theta}^*$, there is a neighbourhood $N(\boldsymbol{\theta}^*)$ in which the following holds

$$\widehat{\boldsymbol{\theta}} \in N(\boldsymbol{\theta}^*) \quad \text{and} \quad \mathbf{y}(t, \widehat{\boldsymbol{\theta}}) = \mathbf{y}(t, \boldsymbol{\theta}^*) \Rightarrow \widehat{\theta}_i = \theta_i^*. \tag{5.5}$$

A model is called structurally locally identifiable if all its parameters are structurally locally identifiable. Otherwise, it is structurally unidentifiable [99].

Similar to parameter identifiability, a state $x_j(\tau)$ of the model is said to be observable if it can be determined from the output $\mathbf{y}(t)$, that is, the state $x_j(\tau)$ is observable if for almost every $\mathbf{x}^*(\tau)$ there is a neighbourhood $N(\mathbf{x}^*(\tau))$ in which the following holds

$$\hat{\mathbf{x}}(\tau) \in N(\mathbf{x}^*(\tau)) \quad \text{and} \quad \mathbf{y}(t, \hat{\mathbf{x}}(\tau)) = \mathbf{y}(t, \mathbf{x}^*(\tau)) \Rightarrow \hat{x}_j(\tau) = x_j^*(\tau) \tag{5.6}$$

where $t_0 \leq \tau \leq t \leq t_f$ for a finite time t_f. A model is called observable if all its state variables are observable; otherwise, it is unobservable [99].

In paper D, we assess the structural identifiability of the proposed model for the fatty acid synthesis pathway in Fig. 5.3. This has been done using a framework presented by Villaverde et al. [99] to jointly study the state observability and parametric structural identifiability of the model. For more on this, see the manuscript. This analysis revealed that the experimental concentrations of at least two of the metabolites must be provided to ensure the model observability and structural identifiability of the model parameters.

5.3.2 Practical identifiability and local sensitivity analysis

A concept of practical or data-based identifiability has also been proposed in the literature [42; 77; 78]. A parameter θ_i in a dynamic model of the form of Eq. 5.1 is practically non-identifiable if it is impossible to determine a unique value for it by fitting the model to the data. For some given model, assume that the experimental values are available for the measurable quantities in $\mathbf{y}$ given by Eq. 5.4. By minimising the model predictions for vector $\mathbf{y}$ against its data, we can acquire the estimated parameters $\bar{\boldsymbol{\theta}}$.

As suggested by Gábor et al. [32], the practical identifiability of the parameters $\boldsymbol{\theta}$ can be assessed in two consecutive steps. First, we calculate the sensitivities of the measurable quantities in $\mathbf{y}$ to changes in the estimated parameters $\bar{\boldsymbol{\theta}}$ through calculating the sensitivity matrix $\mathbf{S}^L = \{s^L_{ki}\}$ with

$$s^L_{ki} = \frac{\partial y_k}{\partial \theta_i} |_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}}, \quad k = 1, \ldots, N_y, \quad i = 1, \ldots, M. \tag{5.7}$$

The columns of $\mathbf{S}^L$ correspond to parameters and the rows to the measurable quantities in $\mathbf{y}$. There are several ways to define an overall scoring for individual parameters using the sensitivity matrix $\mathbf{S}^L$, see e.g. Brun et al. [12]. The most common measures could be the root mean squared sensitivity defined as

$$s^{msqr}_i = \sqrt{\frac{1}{N_y} \sum_{k=1}^{N_y} (s^L_{ki})^2}, \quad i = 1, \ldots, M \tag{5.8}$$

or the mean sensitivity given by

$$s^{mean}_i = \frac{1}{N_y} \sum_{k=1}^{N_y} s^L_{ki}, \quad i = 1, \ldots, M. \tag{5.9}$$

The mean sensitivity in Eq. 5.9 provides information on the sign of the average effect a change in a parameter has on the measurable quantities, $\mathbf{y}$.

Using the root mean squared sensitivity given in Eq. 5.8, the parameters with scores below a threshold are classified as non-identifiable parameters as they have no or very little effect on the observed quantities [32]. This type of analysis is known as the local sensitivity analysis (LSA) since it studies the sensitivity of the model output to the local changes in the parameter values.

The parameters may also interfere with each other in a model. Even if a parameter is classified as an influential parameter on the model output, it might still be non-identifiable. This can happen as its possible effect on the output might be compensated by changes in other parameters. This phenomenon is known as collinearity. Let us consider a set of p parameters selected from $\boldsymbol{\theta}$ and build the matrix $\mathbf{S}^L_{sub}$ from their corresponding columns in $\mathbf{S}^L$. This set of parameters are said to be linearly dependent or collinear if there exists a vector $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \ldots & \alpha_p \end{bmatrix}^T$ with $||\boldsymbol{\alpha}|| \neq 0$ such that $\mathbf{S}^L_{sub}\boldsymbol{\alpha} = \mathbf{0}$ [12]. If this relation holds only approximately, the parameters are said to be nearly dependent, and if it does not hold the parameters are independent.

In paper C, we have checked both sources of practical non-identifiability using the estimated value of the parameters. This analysis showed that the practical non-identifiability was not an issue fitting the proposed model to the EROD activity data, and this is a strength of the proposed model of paper C.

Chapter 6

Overview of the papers

In this chapter, we give an overview of the papers in the thesis and their possible connections. The main focus of papers A and B is the mathematical methodology in systems biology, while both papers C and D focus on constructing dynamic models for two different biological phenomena.

Paper A provides an overview of different sampling methods for metabolic network models at the genome scale and compares those methods in terms of convergence, efficiency, and consistency. This is the only paper in the thesis that applies the constraint-based modelling of a metabolic reaction network. A constraint-based model is a static model as it imposes the steady state assumption on the system. In the other three papers, dynamic modelling of metabolic reaction networks has been used. Paper A ranked the Coordinate Hit-and-run with rounding (CHRR) as the best sampling approach. As future work, CHRR can be an alternative approach to study the steady state fluxes of the reactions in the fatty acid synthesis pathway in paper D.

A framework is presented in paper B to evaluate the model reduction under the parameter uncertainty. The method could be potentially used in paper D, e.g. by reducing a dynamic model of the full pathway of fatty acid synthesis when the nominal values of all parameters are available.

Paper C is a biological paper and develops a dynamic model to describe the toxicokinetic interactions between the two chemicals BNF and NOC on the AhR-CYP1A signalling pathway in fish. The model presented in paper C has the potential to be used for other chemical mixtures. For instance, a similar synergistic mixture effect on the CYP1A biomarker has been observed in cells exposed to the polycyclic aromatic hydrocarbon benspyrene in combination with an antifungal imidazole drug clotrimazole (work in progress by Alvord, C.; Lundh, T.; Wiklander, K.; Bernhardsson, A.; Celander, M.C.).

A dynamic model of the fatty acid synthesis pathway is proposed in paper D. The model should be calibrated and validated once the experimental time course concentration values of some of the metabolites are available. The study suggests which steps should be taken from a modeller's side to gain the information needed prior to entering a dialogue with experimentalists concerning the purpose of designing experiments. A common feature of papers C and D is that, in both, a dynamic model of the form of Eq. 4.1 is constructed for two different reaction networks using the kinetics discussed in Section 4.1.

Bibliography

[1] Center for Digital Life Norway. https://digitallifenorway.org/gb/. Accessed: Sep 14, 2020. 1.1.1

[2] dCod 1.0: Decoding the systems toxicology of Atlantic cod. https://www.uib.no/en/dcod. Accessed: Sep 14, 2020. 1.1.1

[3] Fatty acid biosynthesis. https://www.kegg.jp/kegg-bin/show_pathway?org_name=dre&mapno=00061&mapscale=&show_description=show. Accessed: Sep 14, 2020. 1.1.1, 1.1, 1.2.1, 5.2.2

[4] Fatty acid metabolism. https://www.genome.jp/kegg-bin/show_pathway?map01212. Accessed: Sep 14, 2020. 1.1.1

[5] Eivind Almaas, B Kovacs, T Vicsek, Zoltan N Oltvai, and A-L Barabási. Global organization of metabolic fluxes in the bacterium escherichia coli. *Nature*, 427(6977):839–843, 2004. 3.2.1

[6] Claude JP Bélisle, H Edwin Romeijn, and Robert L Smith. Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, 18(2):255–266, 1993. 3.2.1

[7] Nikolaus Berndt, Sascha Bulik, Iwona Wallach, Tilo Wünsch, Matthias König, Martin Stockmann, David Meierhofer, and Hermann-Georg Holzhütter. Hepatokin1 is a biochemistry-based model of liver metabolism for applications in medicine and pharmacology. *Nature communications*, 9(1):1–12, 2018. 5.2.2

[8] Cristina Bizarro, Marta Eide, Daniel J Hitchcock, Anders Goksøyr, and Maren Ortiz-Zarragoitia. Single and mixture effects of aquatic micropollutants studied in precision-cut liver slices of atlantic cod (gadus morhua). *Aquatic Toxicology*, 177:395–404, 2016. 1.1.1

[9] Béla Bollobás. Graph theory. graduate texts in mathematics. 1979. 2.2

[10] Edward L Boone, Jason RW Merrick, and Matthew J Krachey. A hellinger distance approach to mcmc diagnostics. *Journal of Statistical Computation and Simulation*, 84(4):833–849, 2014. 3.2.3

[11] Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998. 3.2.3

[12] Roland Brun, Peter Reichert, and Hans R Künsch. Practical identifiability analysis of large environmental simulation models. *Water Resources Research*, 37(4):1015–1030, 2001. 5.3.2, 5.3.2

[13] Elizabeth Brunk, Swagatika Sahoo, Daniel C Zielinski, Ali Altunkaya, A Draeger, Nathan Mih, F Gatto, A Nilsson, GAP Gonzales, MK Aurich, et al. Recon3d: A resource enabling a three-dimensional view of gene variation in human metabolism. *Nat Biotech,(Accepted)*, 2017. 1.2

[14] Francesca Campolongo, Andrea Saltelli, and Jessica Cariboni. From screening to quantitative sensitivity analysis. a unified approach. *Computer Physics Communications*, 182(4):978–988, 2011. 5.2

[15] William Jay Conover. *Practical nonparametric statistics*. Wiley New York, 1971. 4.2

[16] Karina Dale, Mette Bjørge Müller, Zhanna Tairova, Essa Ahsan Khan, Kristin Hatlen, Merete Grung, Fekadu Yadetie, Roger Lille-Langøy, Nello Blaser, Hans J Skaug, et al. Contaminant accumulation and biological responses in atlantic cod (gadus morhua) caged at a capped waste disposal site in kollevåg, western norway. *Marine environmental research*, 145:39–51, 2019. 1.1.1

[17] Daniele De Martino, Matteo Mori, and Valerio Parisi. Uniform sampling of steady states in metabolic networks: heterogeneous scales and rounding. *PloS one*, 10(4):e0122670, 2015. 3.2.1

[18] Jay L Devore and Kenneth N Berk. Modern mathematical statistics with applications. pages 249–252, 2012. 3.2.3

[19] Natalie C Duarte, Scott A Becker, Neema Jamshidi, Ines Thiele, Monica L Mo, Thuy D Vo, Rohith Srivas, and Bernhard Ø Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782, 2007. 1.2

[20] Ali Ebrahim, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. Cobrapy: constraints-based reconstruction and analysis for python. *BMC systems biology*, 7(1):74, 2013. 3.2.1

[21] Jeremy S Edwards and Bernhard O Palsson. Systems properties of the haemophilus influenzaerd metabolic genotype. *Journal of Biological Chemistry*, 274(25):17410–17416, 1999. 3.1.1

[22] JS Edwards and BO Palsson. The escherichia coli mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, 97(10):5528–5533, 2000. 1.2, 3.1.1

[23] M Eide, OA Karlsen, H Kryvi, PA Olsvik, and A Goksøyr. Precision-cut liver slices of atlantic cod (gadus morhua): An in vitro system for studying the effects of environmental contaminants. *Aquatic Toxicology*, 153:110–115, 2014. 1.1.1

[24] Marta Eide, Halfdan Rydbeck, Ole K Tørresen, Roger Lille-Langøy, Pål Puntervoll, Jared V Goldstone, Kjetill S Jakobsen, John Stegeman, Anders Goksøyr, and Odd A Karlsen. Independent losses of a xenobiotic receptor across teleost evolution. *Scientific reports*, 8(1):1–13, 2018. 1.1.1

[25] Shirin Fallahi, Hans J Skaug, and Guttorm Alendal. A comparison of monte carlo sampling methods for metabolic network models. *PloS one*, 15(7):e0235393, 2020. 3.2.3

[26] Adam M Feist and Bernhard O Palsson. The biomass objective function. *Current opinion in microbiology*, 13(3):344–349, 2010. 3.1.1

[27] Elisenda Feliu and Carsten Wiuf. Simplifying biochemical models with intermediate species. *Journal of the royal society interface*, 10(87):20130484, 2013. 4.2

[28] David A Fell and J Rankin Small. Fat synthesis in adipose tissue. an examination of stoichiometric constraints. *Biochemical Journal*, 238(3):781–786, 1986. 3.1.1

[29] Jochen Förster, Iman Famili, Patrick Fu, Bernhard Ø Palsson, and Jens Nielsen. Genome-scale reconstruction of the saccharomyces cerevisiae metabolic network. *Genome research*, 13(2):244–253, 2003. 3.1.1

[30] Håvard G Frøysa. Estimation and identifiability of kinetic parameters in dynamical models of biochemical reaction networks. 2020. 2.1

[31] Håvard G Frøysa, Shirin Fallahi, and Nello Blaser. Evaluating model reduction under parameter uncertainty. *BMC systems biology*, 12(1):79, 2018. 4.2

[32] Attila Gábor, Alejandro F Villaverde, and Julio R Banga. Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems. *BMC systems biology*, 11(1):1–16, 2017. 5.3.2, 5.3.2

[33] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013. 3.2.3

[34] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984. 3.2.2

[35] John Geweke et al. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, 1991. 3.2.3

[36] Walter R Gilks. Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice*, 1996. 3.2.1

[37] Johanna Gräns, Junko Johansson, Marie Michelová, Britt Wassmur, Elisabeth Norström, Margareta Wallin, and Malin C Celander. Mixture effects between different azoles and β-naphthoflavone on the cyp1a biomarker in a fish cell line. *Aquatic Toxicology*, 164:43–51, 2015. 1.1.2, 4.1.4, 5.1

[38] Branko Grünbaum and Geoffrey C Shephard. Convex polytopes. *Bulletin of the London Mathematical Society*, 1(3):257–300, 1969. 3.2.1

[39] Cato M Guldberg and Peter Waage. Studies concerning affinity. *CM Forhandlinger: Videnskabs-Selskabet i Christiana*, 35(1864):1864, 1864. 4

[40] Hulda S Haraldsdóttir, Ben Cousins, Ines Thiele, Ronan MT Fleming, and Santosh Vempala. Chrr: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics*, 33(11):1741–1743, 2017. 3.2.1

[41] Markus Heinonen, Maria Osmala, Henrik Mannerström, Janne Wallenius, Samuel Kaski, Juho Rousu, and Harri Lähdesmäki. Bayesian metabolic flux analysis reveals intracellular flux couplings. *Bioinformatics*, 35(14):i548–i557, 2019. 3.2.2, 3.2.2, 3.2.2

[42] Stefan Hengl, Clemens Kreutz, Jens Timmer, and Thomas Maiwald. Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, 23(19):2612–2618, 2007. 5.3.2

[43] Helena A Herrmann, Beth C Dyson, Lucy Vass, Giles N Johnson, and Jean-Marc Schwartz. Flux sampling is a powerful tool to study metabolism under changing environmental conditions. *NPJ systems biology and applications*, 5(1):1–8, 2019. 3.1.1

[44] Archibald Vivian Hill. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *j. physiol.*, 40:4–7, 1910. 4.1.3

[45] DE Hinton, H Segner, DWT Au, SW Kullman, and RC Hardman. The toxicology of fishes. *CRC Press Inc., Florida*, pages 328–352, 2008. 1.1.2

[46] Padraig Houlahan and John Scalo. Recognition and characterization of hierarchical interstellar structure. ii-structure tree statistics. *The Astrophysical Journal*, 393:172–187, 1992. 4.2

[47] Trey Ideker, Timothy Galitski, and Leroy Hood. A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, 2(1):343–372, 2001. 1.3

[48] Brian P Ingalls. *Mathematical modeling in systems biology: an introduction*. MIT press, 2013. 3, 4, 4, 4.1.1, 4.1.2, 4.1.2, 4.1.2

[49] Kenneth A Johnson and Roger S Goody. The original michaelis constant: translation of the 1913 michaelis–menten paper. *Biochemistry*, 50(39):8264–8269, 2011. 4.1.1

[50] Kenneth J Kauffman, Purusharth Prakash, and Jeremy S Edwards. Advances in flux balance analysis. *Current opinion in biotechnology*, 14(5):491–496, 2003. 3.1.1, 3.1.1

[51] David E Kaufman and Robert L Smith. Direction choice for accelerated convergence in hit-and-run sampling. *Operations Research*, 46(1):84–95, 1998. 3.2.1

[52] Zachary A King, Justin Lu, Andreas Dräger, Philip Miller, Stephen Federowicz, Joshua A Lerman, Ali Ebrahim, Bernhard O Palsson, and Nathan E Lewis. Bigg models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research*, 44(D1):D515–D522, 2016. 3.2.3

[53] Werner Krauth. Introduction to monte carlo algorithms. In *Advances in Computer Simulation*, pages 1–35. Springer, 1998. 3.2.1

[54] Arthur Lesk. *Introduction to bioinformatics*. Oxford university press, 2019. 1.3

[55] E Wang Lund. Guldberg and waage and the law of mass action. *Journal of Chemical Education*, 42(10):548, 1965. 4, 4.1.1

[56] Timo R Maarleveld, Ruchir A Khandelwal, Brett G Olivier, Bas Teusink, and Frank J Bruggeman. Basic concepts and principles of stoichiometric modeling of metabolic networks. *Biotechnology journal*, 8(9):997–1008, 2013. 2.1

[57] Michael MacGillivray, Amy Ko, Emily Gruber, Miranda Sawyer, Eivind Almaas, and Allen Holder. Robust analysis of fluxes in genome-scale metabolic pathways. *Scientific reports*, 7(1):1–20, 2017. 3.2.2

[58] R Mahadevan and CH Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering*, 5(4):264–276, 2003. 3.1.1

[59] Michael D McKay, Richard J Beckman, and William J Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61, 2000. 5.2

[60] Wout Megchelenbrink, Martijn Huynen, and Elena Marchiori. optgpsampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. *PloS one*, 9(2):e86587, 2014. 3.2.1

[61] M. Michelová. Cocktail effects of -naphthoflavone and nocodazole in the poeciliopsis lucida cell line. *Master-thesis, University of Gothenburg*, 2012. 1.1.2, 4.1.4, 5.1

[62] Max D Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991. 5.2

[63] Miles S Okino and Michael L Mavrovouniotis. Simplification of mathematical models of chemical reaction systems. *Chemical reviews*, 98(2):391–408, 1998. 4.2

[64] Jeffrey D Orth, Tom M Conrad, Jessica Na, Joshua A Lerman, Hojung Nam, Adam M Feist, and Bernhard Ø Palsson. A comprehensive genome-scale reconstruction of escherichia coli metabolism2011. *Molecular systems biology*, 7(1):535, 2011. 1.2

[65] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010. 3.1.1

[66] Tiina M Pakula, Heli Nygren, Dorothee Barth, Markus Heinonen, Sandra Castillo, Merja Penttilä, and Mikko Arvas. Genome wide analysis of protein production load in trichoderma reesei. *Biotechnology for biofuels*, 9(1):132, 2016. 3.2.2

[67] Bernhard Ø Palsson. *Systems biology: properties of reconstructed networks*. Cambridge university press, 2006. ISBN 9780521859035. 1.2, 1.3, 3.1.1

[68] Bernhard Ø. Palsson. *Systems Biology: Constraint-based Reconstruction and Analysis*. Cambridge University Press, 2015. ISBN 9781316239940. 1.2, 1.2.1, 2, 2.1, 2.1.1, 2.1.1, 3.1, 3.1.1, 3.1.1

[69] Eleftherios Terry Papoutsakis. Equations and calculations for fermentations of butyric acid bacteria. *Biotechnology and bioengineering*, 67(6):813–826, 2000. 3.1.1

[70] Eleftherios Terry Papoutsakis and Charles L Meyer. Equations and calculations of product yields and preferred pathways for butanediol and mixed-acid fermentations. *Biotechnology and bioengineering*, 27(1):50–66, 1985. 3.1.1

[71] Francesca Pianosi, Keith Beven, Jim Freer, Jim W Hall, Jonathan Rougier, David B Stephenson, and Thorsten Wagener. Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software*, 79:214–232, 2016. 5.2, 5.2

[72] Ovidiu Radulescu, Alexander N Gorban, Andrei Zinovyev, and Vincent Noel. Reduction of dynamical biochemical reactions networks in computational biology. *Frontiers in genetics*, 3:131, 2012. 4.2

[73] AE Raftery and S Lewis. How many iterations in the gibbs sampler? bayesian statistics 4 (bernardo, jm, berger, jo, dawid, ap and smith, afm, eds), 1992. 3.2.3

[74] Ramprasad Ramakrishna, Jeremy S Edwards, Andrew McCulloch, and Bernhard O Palsson. Flux-balance analysis of mitochondrial energy metabolism: consequences of systemic stoichiometric constraints. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 280(3):R695–R704, 2001. 3.1.1

[75] Karthik Raman and Nagasuma Chandra. Flux balance analysis of biological systems: applications and challenges. *Briefings in bioinformatics*, 10(4):435–449, 2009. 3

[76] Shodhan Rao, Arjan Van der Schaft, Karen Van Eunen, Barbara M Bakker, and Bayu Jayawardhana. A model reduction method for biochemical reaction networks. *BMC systems biology*, 8(1):52, 2014. 2.1.1, 2.2, 2.2, 2.2, 2.2, 2.2, 4.2, 4.2, 4.4, 4.2, 4.2

[77] A Raue, V Becker, Ursula Klingmüller, and J Timmer. Identifiability and observability analysis for experimental design in nonlinear dynamical models. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(4):045105, 2010. 5.3.2

[78] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009. 5.3.2

[79] Gareth O Roberts. Markov chain concepts related to sampling algorithms. *Markov chain Monte Carlo in practice*, 57:45–58, 1996. 3.2.2

[80] Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008. 5.2, 5.2.2

[81] Maurice Scheer, Andreas Grote, Antje Chang, Ida Schomburg, Cornelia Munaretto, Michael Rother, Carola Söhngen, Michael Stelzer, Juliane Thiele, and Dietmar Schomburg. Brenda, the enzyme information system in 2011. *Nucleic acids research*, 39(suppl_1):D670–D676, 2010. 4.1.1, 5

[82] Jan Schellenberger, Richard Que, Ronan MT Fleming, Ines Thiele, Jeffrey D Orth, Adam M Feist, Daniel C Zielinski, Aarash Bordbar, Nathan E Lewis, Sorena Rahmanian, et al. Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox v2. 0. *Nature protocols*, 6(9):1290, 2011. 1.3, 3.2.1

[83] Wilhelmus HA Schilders, Henk A Van der Vorst, and Joost Rommes. *Model order reduction: theory, research aspects and applications*, volume 13. Springer, 2008. 4.2

[84] Thomas J Snowden, Piet H van der Graaf, and Marcus J Tindall. Methods of model reduction for large-scale biological systems: a survey of current methods and trends. *Bulletin of mathematical biology*, 79(7):1449–1486, 2017. 4.2

[85] RC Spear and GM Hornberger. Eutrophication in peel inletii. identification of critical uncertainties via generalized sensitivity analysis. *Water research*, 14(1):43–49, 1980. 5.2

[86] Warren E Stewart, Mike Caracotsios, and Jan P Sørensen. Parameter estimation from multiresponse data. *AIChE Journal*, 38(5):641–650, 1992. 5

[87] R Core Team et al. R: A language and environment for statistical computing, 2013. 3.2.2, 5.2.1

[88] Ines Thiele, Neil Swainston, Ronan MT Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K Aurich, Hulda Haraldsdottir, Monica L Mo, Ottar Rolfsson, Miranda D Stobbe, et al. A community-driven global reconstruction of human metabolism. *Nature biotechnology*, 31(5):419–425, 2013. 1.2

[89] Chris Tofallis. A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, 66(8):1352–1362, 2015. 5.1

[90] Valentin F Turchin. On the computation of multidimensional integrals by the monte-carlo method. *Theory of Probability & Its Applications*, 16(4):720–724, 1971. 3.2.1

[91] Karel Van den Meersche, Karline Soetaert, and Dick Van Oevelen. xsample (): an r function for sampling linear inverse problems. *Journal of Statistical Software*, 30(Code Snippet 1), 2009. 3.2.2, 3.2.2, 3.2.3

[92] Robert J Vanderbei. *Linear programming: foundations and extensions*, volume 285. Springer Nature, 2020. 3.1.1, 3.1.1

[93] Amit Varma, Brian W Boesch, and Bernhard O Palsson. Biochemical production capabilities of escherichia coli. *Biotechnology and bioengineering*, 42(1):59–73, 1993. 3.1.1

[94] Amit Varma, Brian W Boesch, and Bernhard O Palsson. Stoichiometric interpretation of escherichia coli glucose catabolism under various oxygenation rates. *Applied and environmental microbiology*, 59(8):2465–2473, 1993. 3.1.1

[95] Amit Varma and Bernhard O Palsson. Metabolic capabilities of escherichia coli: I. synthesis of biosynthetic precursors and cofactors. *Journal of theoretical biology*, 165(4):477–502, 1993. 3.1.1

[96] Amit Varma and Bernhard O Palsson. Metabolic capabilities of escherichia coli ii. optimal growth patterns. *Journal of Theoretical Biology*, 165(4):503–522, 1993. 3.1.1

[97] Amit Varma and Bernhard O Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type escherichia coli w3110. *Applied and environmental microbiology*, 60(10):3724–3731, 1994. 3.1.1

[98] Alejandro F Villaverde, Antonio Barreiro, and Antonis Papachristodoulou. Structural identifiability of dynamic systems biology models. *PLoS computational biology*, 12(10):e1005153, 2016. 5.3.1

[99] Alejandro F Villaverde, Nikolaos Tsiantis, and Julio R Banga. Full observability and estimation of unknown inputs, states and parameters of nonlinear biological models. *Journal of the Royal Society Interface*, 16(156):20190043, 2019. 5.3.1, 5.3.1, 5.3.1

[100] Hao Wang, Simonas Marcišauskas, Benjamín J Sánchez, Iván Domenzain, Daniel Hermansson, Rasmus Agren, Jens Nielsen, and Eduard J Kerkhoven. Raven 2.0: A versatile toolbox for metabolic network reconstruction and a case study on streptomyces coelicolor. *PLoS computational biology*, 14(10):e1006541, 2018. 1.3

[101] MR Watson. Metabolic maps for the apple ii, 1984. 3.1.1

[102] Sharon J Wiback, Iman Famili, Harvey J Greenberg, and Bernhard Ø Palsson. Monte carlo sampling can be used to determine the size and shape of the steady-state flux space. *Journal of theoretical biology*, 228(4):437–447, 2004. 3.2.1

[103] Ulrike Wittig, Renate Kania, Martin Golebiewski, Maja Rey, Lei Shi, Lenneke Jong, Enkhjargal Algaa, Andreas Weidemann, Heidrun Sauer-Danzwith, Saqib Mir, et al. Sabio-rkdatabase for biochemical reaction kinetics. *Nucleic acids research*, 40(D1):D790–D796, 2012. 4.1.1, 5

[104] Fekadu Yadetie, Eystein Oveland, Anne Døskeland, Frode Berven, Anders Goksøyr, and Odd André Karlsen. Quantitative proteomics analysis reveals perturbation of lipid metabolic pathways in the liver of atlantic cod (gadus morhua) treated with pcb 153. *Aquatic Toxicology*, 185:19–28, 2017. 1.1.1

[105] Fekadu Yadetie, Xiaokang Zhang, Eileen Marie Hanna, Libe Aranguren-Abadía, Marta Eide, Nello Blaser, Morten Brun, Inge Jonassen, Anders Goksøyr, and Odd André Karlsen. Rna-seq analysis of transcriptome responses in atlantic cod (gadus morhua) precision-cut liver slices exposed to benzo [a] pyrene and 17α-ethynylestradiol. *Aquatic toxicology*, 201:174–186, 2018. 1.1.1

[106] PC Young, GM Hornberger, and RC Spear. Modeling badly defined systems: some further thoughts. In *Proceedings SIMSIG Conference, Canberra*, pages 24–32, 1978. 5.2

[107] Yuting Zheng and Ganesh Sriram. Mathematical modeling: bridging the gap between concept and realization in synthetic biology. *Journal of Biomedicine and Biotechnology*, 2010, 2010. 5

[108] Zhike Zi. Sensitivity analysis approaches applied to systems biology models. *IET systems biology*, 5(6):336–346, 2011. 5.2

Part II

Included papers

Paper A

A comparison of Monte Carlo sampling methods for metabolic network models

Shirin Fallahi, Hans J. Skaug, Guttorm Alendal
PLOS ONE, **15**:7 (2020)

PLOS ONE

RESEARCH ARTICLE

A comparison of Monte Carlo sampling methods for metabolic network models

Shirin Fallahi*, Hans J. Skaug, Guttorm Alendal

Department of Mathematics, University of Bergen, Bergen, Norway

* shirin.fallahi@uib.no

OPEN ACCESS

Citation: Fallahi S, Skaug HJ, Alendal G (2020) A comparison of Monte Carlo sampling methods for metabolic network models. PLoS ONE 15(7): e0235393. https://doi.org/10.1371/journal.pone.0235393

Editor: Inés P. Mariño, Universidad Rey Juan Carlos, SPAIN

Received: October 2, 2019

Accepted: June 15, 2020

Published: July 1, 2020

Data Availability Statement: All data supporting the findings of this study are available within the paper and can also be accessed on figshare or generated using the publicly available source code via the following URLs: https://doi.org/10.6084/m9.figshare.12356258 and https://doi.org/10.6084/M9.figshare.12356318.

Funding: This research was supported by the Research Council of Norway through grant 248840, 704 dCod 1.0. The funding body did not play any role in the design of the study and in the 705 writing of the manuscript.

Abstract

Reaction rates (fluxes) in a metabolic network can be analyzed using constraint-based modeling which imposes a steady state assumption on the system. In a deterministic formulation of the problem the steady state assumption has to be fulfilled exactly, and the observed fluxes are included in the model without accounting for experimental noise. One can relax the steady state constraint, and also include experimental noise in the model, through a stochastic formulation of the problem. Uniform sampling of fluxes, feasible in both the deterministic and stochastic formulation, can provide us with statistical properties of the metabolic network, such as marginal flux probability distributions. In this study we give an overview of both the deterministic and stochastic formulation of the problem, and of available Monte Carlo sampling methods for sampling the corresponding solution space. We apply the ACHR, OPTGP, CHRR and Gibbs sampling algorithms to ten metabolic networks and evaluate their convergence, consistency and efficiency. The coordinate hit-and-run with rounding (CHRR) is found to perform best among the algorithms suitable for the deterministic formulation. A desirable property of CHRR is its guaranteed distributional convergence. Among the three other algorithms, ACHR has the largest consistency with CHRR for genome scale models. For the stochastic formulation, the Gibbs sampler is the only method appropriate for sampling at genome scale. However, our analysis ranks it as less efficient than the samplers used for the deterministic formulation.

Introduction

Cell metabolism involves many chemical reactions, catalyzed by thousands of enzymes, and is often represented as metabolic networks [1]. The dynamics of a metabolic network, consisting of m metabolites and n reactions, can be mathematically modelled by a system of Ordinary Differential Equations (ODEs) written in short form as

$$\frac{d\mathbf{x}}{dt} = S\mathbf{v}(\mathbf{x}(t), \boldsymbol{\alpha}, t). \tag{1}$$

Here, $\mathbf{x} \in \mathbb{R}^m$ is a vector containing of metabolite concentrations, $\boldsymbol{\alpha} \in \mathbb{R}^k$ is a vector of parameters, $S \in \mathbb{R}^{m \times n}$ is the stoichiometric matrix, i.e. a matrix representation of the network, and $\mathbf{v}(\mathbf{x}, \boldsymbol{\alpha}, t) \in \mathbb{R}^n$ are the flux rates in the n reactions [2].

Competing interests: The authors have declared that no competing interests exist.

The stoichiometric matrix S is constructed so that element S_{ij} is positive (negative) if metabolite i is created (consumed) by reaction j, represented by the flux rate v_j, and is assumed constant. A challenge is to establish models of the different flux rates, in general nonlinear in $\mathbf{x}$, and to estimate the k parameters in $\boldsymbol{\alpha}$ through in-vivo and in-vitro experiments. The non-linearity of the ODE system also makes the system susceptible to chaotic behavior, bifurcation and sensitivity to parameter values [3].

In Flux Balance Analysis (FBA) [4] the model system is assumed to be in a steady state

$$\frac{d\mathbf{x}}{dt} = S\,\mathbf{v} = \mathbf{0}, \tag{2}$$

i.e. the problem goes from being a set of differential equations in $\mathbf{x}$ to become an algebraic problem, with the flux rates $\mathbf{v}$ as unknowns. Often the flux rates are constrained with upper and lower bounds

$$\mathbf{v}^{lb} \leq \mathbf{v} \leq \mathbf{v}^{ub}. \tag{3}$$

However, since a typical metabolic network has fewer metabolites than reactions, i.e. $m < n$, the system in Eqs (2) and (3) is in general undetermined. The system might have many feasible solutions in a closed convex polytope, the n-dimensional analogue to the three dimensional polyhedron, formed by the intersection of the kernel of S and the linear inequalities in Eq (3) [5]. A unique solution might be found by introducing an objective function which aims to optimize some biological functionality, for example maximizing cell growth rate or ATP production of an organism [2]. A challenge in FBA is to choose the most appropriate objective function.

An alternative to FBA, which avoids the need to specify an objective function, is to sample (uniformly) from the flux polytope defined by Eqs (2) and (3). The solution space can then be characterized statistically from the set of sampled $\mathbf{v}$ vectors in terms of a probability density function (pdf), which we denote by $p(\mathbf{v})$ [6]. We will distinguish between a *deterministic* and *stochastic* formulation of given metabolic model and the associated flux measurements. The stochastic formulation is more flexible in that it can account for measurement error and allows relaxation of the steady state condition in Eq (2).

Fig 1 illustrates the key concepts used in this paper. The simple metabolic network consists of a single input flux, v_1, $m = 1$ metabolite, and a single output flux v_2, i.e. $n = 2$ fluxes in total. The resulting constrained steady state equation is given in Fig 1a). Panel b) shows the polytope representing the solution space, which in this case is a line segment in the v_1-v_2 plane. Panel b) also shows the uniform pdf's $p(v_1)$ and $p(v_2)$ indicating that all flux values in the feasible intervals are "equally probable".

Fig 1 illustrates how the deterministic and stochastic frameworks differ in the way they incorporate flux measurements. In the deterministic case (Panel c), fixing v_1 experimentally uniquely determines v_2. Both pdf's collapse to point masses, and all other a-priori feasible values have zero probability. In a stochastic framework (Panel d), on the other hand, the uncertainty in the measurement of v_1 can be taken into account. When this uncertainty is combined with constraints imposed by the polytope, the resulting pdf's $p(v_1)$ and $p(v_2)$ are non-degenerate as shown in Panel d), and displays the marginal likelihood of each feasible flux value.

Another limiting assumption of the deterministic formulation is the exact steady state assumption. This assumption is not always realistic and should be relaxed to have a model compatible with the stochastic nature of biological networks [7, 8]. In Fig 1e), we relax the steady state assumption (2), while still incorporating the uncertain measurement v_1. This leads to wider pdf's (Panel e versus d), and the solution polytope is not necessarily convex any more.

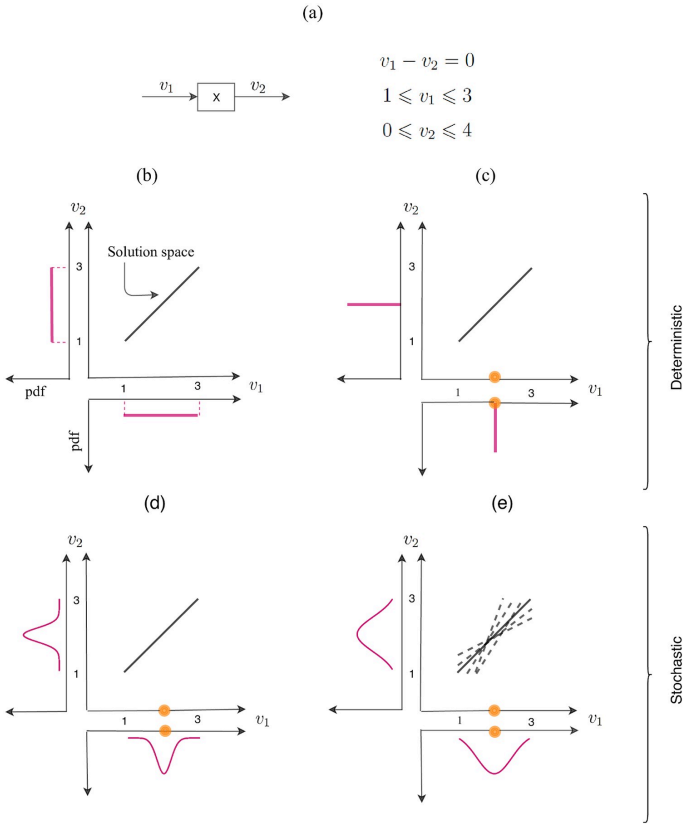


Fig 1. Solution space and sampling pdf $p(v)$ (pink curve) under different experimental setups. (a): Example metabolic network and corresponding mathematical model. (b): Deterministic formulation without measurements. (c) and (d): Flux measurement of v_1 (orange circle) available in deterministic setup (c) and stochastic setup (d). (e): Relaxed steady state assumption and flux measurement of v_1 in stochastic setup.

https://doi.org/10.1371/journal.pone.0235393.g001

For genome-scale metabolic models the dimension (n) of the polytope formed by Eqs (2) and (3) is typically high and deterministic sampling from such polytopes is challenging [9]. Hence, Monte Carlo (MC) approximations are often used [10]. In Wiback et al. [11] a MC rejecting sampling algorithm was used to sample low dimensional polytopes. However, this algorithm becomes inefficient when n is large, so a more commonly used algorithm is the hit-and-run (HR) [12, 13], which is a Markov Chain Monte Carlo (MCMC) method. Almaas et al. [14] originally applied the HR algorithm to the bacterium Escherichia coli metabolic network. The algorithm efficiently samples from the solution space as long as the polytope is isotropic in scales of the fluxes, i.e. being independent on direction in the high dimensional sample space.

High dimensional polytopes that are very narrow in some directions are difficult to sample properly. To cope with this anisotropy problem, the artificial centering hit-and-run (ACHR) algorithm has been developed [15]. The ACHR algorithm and an algorithm based on ACHR, known as optimized general parallel sampler (OPTGP) [16], are widely used to sample the solution space of metabolic models. However, both samplers suffer from convergence problems due to the non-Markovian nature of ACHR [17]. The ACHR algorithm is implemented in both the COnstrained Based Reconstruction and Analysis (COBRA) toolbox [18] (in Matlab) and COBRApy (in Python). The OPTGP is available only in COBRApy. Recently, rounding procedures have been proposed to remove the heterogeneity issue of the solution space, and then a modified version of HR is used [17, 19]. Coordinate hit-and-run with rounding (CHRR) [19] is also implemented in the COBRA toolbox. The algorithms mentioned so far are designed to sample the polytope formed by a deterministic formulation of the model (Fig 1b and 1c). The run time and convergence of the two ACHR based algorithms and CHRR are compared using three constraint-based models in the study by Herrmann et al. [6].

In the study by Van den Meersche et al. [20] a general framework to solve a linear inverse problem using a MCMC algorithm is presented. The suggested framework can be used to sample the solution space of a metabolic network model which is constructed to encode an exact steady state assumption, bounded fluxes and flux observations with related experimental noise (Fig 1d). A function is available in the *limSolve* R package [21] to perform the sampling in this framework.

Another option is to relax the steady state constraint in Eq (2) while including the flux data and corresponding noise (Fig 1e). Considering these assumptions, a statistical model using Bayesian framework has been introduced by Heinonen et al. [22], and a truncated multivariate normal (*TMVN*) posterior distribution for the fluxes has been presented. Efficient sampling from a truncated multivariate normal distribution is a challenging task, and often Gibbs sampling is applied [23]. The Bayesian metabolic flux analysis (BMFA) is implemented in the COBRA toolbox by Heinonen et al. [22].

To our knowledge this is the first time that both available deterministic and stochastic frameworks are reviewed and corresponding sampling algorithms are compared to each other. In this study we have evaluated ACHR, OPTGP and CHRR algorithms which are appropriate for the deterministic formulation. Even if we use different criteria than the ones used by Herrmann et al. [6] our results are in good agreement with their findings. In addition, we have evaluated sampling algorithms *xsample()* and Gibbs which are related to the stochastic formulation. These algorithms have not been discussed by Herrmann et al. [6].

First we give an overview of available MC sampling algorithms for the different cases presented in Fig 1, and discuss their pros and cons. Then, an assessment of algorithms in terms of convergence, consistency and efficiency is given. We conclude the paper with a discussion on which framework and sampling algorithm might be better to use considering restrictions in the model and level of uncertainty for available flux measurements.

Survey of sampling algorithms

Below follows a brief description each of the algorithms included in this study, cast in a common notation. For more details the reader is referred to the background papers.

Deterministic formulation

We begin by describing the standard hit-and-run (HR) algorithm to sample from a convex set. We then review HR related algorithms to approximate uniform sampling from a convex

polytope, which is a convex set of points, constructed by the exact steady state in Eq (2) and the capacity constraints in Eq (3) on metabolic fluxes.

The Hit-and-Run sampling algorithm (HR). The standard HR algorithm collects samples from a given N dimensional convex set P by choosing an arbitrary starting point $\mathbf{v}^{(0)} \in P$, setting $a = 0$ where a is the iteration number and going iteratively through three steps:

1. choosing an arbitrary direction $\boldsymbol{\theta}^{(a)}$ uniformly distributed on the boundary of the unit sphere in $\mathbb{R}^N$;
2. finding the minimum (maximum) value of $\lambda \in \mathbb{R}$ denoted by λ_{min} (λ_{max}) such that $\mathbf{v}^{(a)} + \lambda\boldsymbol{\theta}^{(a)} \in P$ and choose a random step size $\lambda^{(a)} \in [\lambda_{min}, \lambda_{max}]$;
3. generating a new sample $\mathbf{v}^{(a+1)} = \mathbf{v}^{(a)} + \lambda^{(a)}\,\boldsymbol{\theta}^{(a)}$ by taking a step of size $\lambda^{(a)}$ from the current sample $\mathbf{v}^{(a)}$ in the direction $\boldsymbol{\theta}^{(a)}$ and then set $a = a + 1$.

The HR technique is a MCMC approach since it generates a new sample by using only the current sample point, which is the definition of the Markov property. Convergence to the target distribution is guaranteed for a MCMC sampling approach, see for instance [24].

The simple HR algorithm performs effectively in a high dimensional space as long as the solution space is isotropic. A bottleneck of the standard HR is the diffusion in the presence of narrow corners in the solution space due to tightly constrained fluxes. In narrow regions HR has to take small steps and consequently the new sample is close to the previous one. This prevents the sampler to perform a full exploration of the solution space of an irregular shape in a finite time, and is known as *slow mixing*.

Artificial Centering Hit-and-Run (ACHR). The artificial centering hit-and-run (ACHR) was proposed by Kaufman et al. [15] to overcome the problem of slow mixing. In a highly heterogeneous solution space a uniform direction choice on the boundary of the unit sphere is a poor choice. The core idea of the ACHR is to use optimal direction choices in HR to allow for larger steps along the elongated directions. In each iteration the sampler tries to approximate a center for the space by computing the mean of all the samples generated so far for each coordinate. Then it chooses randomly a sample from all the samples generated and find a new direction by normalizing the difference between the selected sample and the current approximated center. Considering an arbitrary starting point $\mathbf{v}^{(0)} \in P$, a number of warm up samples $M_{warm} \geq N$, setting $a = 0$ and an initial center $\hat{\mathbf{c}} = \mathbf{v}^{(0)}$, ACHR generates samples iteratively by performing four steps:

1. generate a direction: if $a < M_{warm}$ (warm up phase), select a direction $\boldsymbol{\theta}^{(a)}$ as in the standard HR approach. Otherwise (main phase) choose a number i uniformly distributed on $\{0, 1, \ldots, a\}$ and compute a direction $\boldsymbol{\theta}^{(a)} = \frac{\mathbf{v}^{(i)} - \hat{\mathbf{c}}}{\|\mathbf{v}^{(a)} - \hat{\mathbf{c}}\|}$;
2. choose a random step size $\lambda^{(a)}$ as in the standard HR;
3. generate a new sample $\mathbf{v}^{(a+1)} = \mathbf{v}^{(a)} + \lambda^{(a)}\,\boldsymbol{\theta}^{(a)}$ and then set $a = a + 1$;
4. update the artificial center by setting $\hat{\mathbf{c}} = \frac{a\hat{\mathbf{c}} + \mathbf{v}^{(a)}}{a+1}$.

In each iteration of ACHR in the main phase, the direction is dependent on all previous iterates and directions and this makes the sampler a non-Markovian algorithm. Therefore it is not guaranteed that the sequence of iterates converges toward the target distribution.

For genome scale metabolic models, this algorithm might perform slow to sample the polytope formed by the solution space. To make the sampling process faster, an algorithm named the optimized general parallel sampler (OPTGP) was proposed by Megchelenbrink et al. [16]. In this algorithm the flux through each reaction is maximized and minimized to generate the

$2n$ warm-up points. From warm up points, this algorithm generates multiple short chains in parallel using the approximated center as in ACHR and it takes only the k^{th} point of the chain as a sample point [16]. In the study by Megchelenbrink et al. [16] it has been shown that the OPTGP performs more efficient than the ACHR by generating samples with higher randomness in a shorter time. Clearly, the ACHR is at the core of the OPTGP and this leads to a non-Markovian algorithm. Even though both algorithms are commonly used in the literature, both of them suffer from convergence problems [17].

Coordinate Hit-and-Run with Rounding (CHRR). As mentioned, the performance of the HR algorithm can be strongly affected by irregularity in the shape of the polytope P representing the solution space, known as ill-conditioning. Suppose R_b is the radius of the biggest ball that can be placed inside the polytope and R_s is the radius of the smallest ball inscribing the polytope. The time a sampling algorithm takes to converge to the target distribution is called the mixing time τ and in Lovász et al. [25] it has been shown that the mixing time of the HR algorithm scales by

$$\tau \simeq O(N^2 \frac{R_s^2}{R_b^2}), \tag{4}$$

where N is the dimension of the polytope. The degree of ill-conditioning for the sampling problem is measured by R_s/R_b, known as the sandwiching ratio of the body. This ratio depends on the orders of magnitude of the flux scales and in genome scale problems this number can reach 10^5 which indicates very high irregularity of the polytope to be sampled [17].

To reduce the sandwiching ratio and eliminate ill-conditioning, an approach is presented in Haraldsdottir et al. [19] that consists of two steps; rounding and sampling. In the rounding phase a maximum volume inscribed ellipsoid is built, based on the presented algorithm in Zhang et al. [26], to match closely the heterogeneous polytope. Then the polytope is rounded through transforming the inscribed ellipsoid to a unit ball. A variant of HR algorithm known as coordinate hit-and-run (CHR) [27] is used to sample from the rounded polytope. In the CHR algorithm the direction $\theta^{(a)}$ is selected randomly along the coordinate directions instead of picking randomly from the unit sphere in $\mathbb{R}^N$. Otherwise the CHR algorithm operates similar to the HR. After running the CHR algorithm the sampled points are transformed back to the original space through an inverse transformation. Since the CHRR uses CHR Markov chain for sampling purpose, its convergence to the target distribution is guaranteed in contrast to ACHR based algorithms [28].

Stochastic formulation

In this part we review the studies of Van den Meersche et al. [20] and Heinonen et al. [22] in which statistical frameworks have been proposed to analyze metabolic fluxes while integrating flux measurements with their noise in the formulation and relaxing the steady state assumption in Eq (2). To our knowledge, these two studies are the only studies presenting sampling algorithms applicable at genome scale.

Sampling linear inverse problems (*xsample()*). In the deterministic formulation represented by Eqs (2) and (3) if the experimental values for some of the fluxes are available, they are integrated in the formula by fixing the fluxes at the given values. However, we do not account for the uncertainty of the flux measurements in the equations if we fix the fluxes at their measured values and this might result in overconfidence in outcomes and conclusions. In Van den Meersche et al. [20] the uncertainties corresponding to the experimental values

were included in the Eqs (2) and (3) by adding a noise term to the algebraic equation

$$A\mathbf{v} = \mathbf{b} + \epsilon, \tag{5}$$

where the data vector is denoted by $\mathbf{b}$ and corresponding uncertainties are encoded by $\epsilon \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. The diagonal matrix $\boldsymbol{\Sigma} = diag(\sigma_1, \ldots, \sigma_n)$ represents the variances of flux data. The matrix A is a diagonal matrix where a_{ii} is one in the presence of data for v_i and otherwise zero. The model describes the exact steady state phase of the network considering the limited capacity of the fluxes and it also accounts for the available flux measurements with their experimental noise.

Van den Meersche et al. [20] provided a function named *xsample()* in R [29] to produce a set of samples of fluxes $\mathbf{v}$ in this framework. The function produces the samples by carrying out a two-staged process. First the equality constraint $S\,\mathbf{v} = \mathbf{0}$ is eliminated since all solutions $\mathbf{v}$ for this system of equations can be written as

$$\mathbf{v} = G\mathbf{u} \tag{6}$$

where $G \in \mathbb{R}^{n \times (n - r_s)}$ is an orthonormal matrix formed by the basis for the null space of S (r_s is the rank of S). The linearly dependent variables $\mathbf{v} \in \mathbb{R}^n$ are transformed to linearly independent variables $\mathbf{u} \in \mathbb{R}^{n - r_s}$. The constraints in terms of $\mathbf{u}$ are

$$AG\mathbf{u} = \mathbf{b} + \epsilon \tag{7}$$

$$\mathbf{v}^{\mathrm{lb}} \leq G\mathbf{u} \leq \mathbf{v}^{\mathrm{ub}}. \tag{8}$$

In the second stage the variables $\mathbf{u}$ are sampled from a proposed *TMVN* distribution with probability density function

$$p(\mathbf{u}) \propto \begin{cases} e^{-\frac{1}{2}(AG\mathbf{u}-\mathbf{b})^T \Sigma^{-1} (AG\mathbf{u}-\mathbf{b})} & \text{if } \mathbf{v}^{\mathrm{lb}} \leq G\mathbf{u} \leq \mathbf{v}^{\mathrm{ub}} \\ 0 & \text{otherwise} \end{cases}. \tag{9}$$

To sample from this distribution, the *xsample()* applies the Metropolis algorithm [30]. The *xsample()* function in R allows to examine three different jump (proposal) algorithms. However, here we discuss only one of them named the mirror algorithm which has been found to perform more efficient for high-dimension problems [20]. This algorithm uses the inequality constraints in Eq (8) as reflective planes. Assume $\mathbf{u}^{(a)}$ is a feasible sample and a new point will be drawn

$$\mathbf{u}_0^{(a+1)} \in N(\mathbf{u}^{(a)}, \boldsymbol{\Omega}) \tag{10}$$

where the normal distribution is in the unrestricted space with mean $\mathbf{u}^{(a)}$ and a set of fixed standard deviations collected in the diagonal matrix $\boldsymbol{\Omega} = diag(\omega_1, \ldots, \omega_{n-r_s})$. If the point $\mathbf{u}_0^{(a+1)}$ fulfills all inequalities in Eq (8), it is accepted as the point $\mathbf{u}^{(a+1)}$ to be evaluated by the acceptance ratio test in the Metropolis algorithm [40]. But if the point $\mathbf{u}_0^{(a+1)}$ violates some inequalities, it is mirrored consecutively in the hyperplane formed by violated inequalities [20]. Then the resulting point $\mathbf{u}^{(a+1)}$ satisfies all inequalities and will be evaluated through the acceptance ratio test to be accepted or rejected.

The diagonal elements of the matrix $\boldsymbol{\Omega}$ are the jump lengths of the Markov Chain. The jump lengths define the step lengths taken and they determine the distance covered within the solution space in one iteration and also the number of reflections in the solution space

boundaries. Due to this the jump lengths have a significant influence on the efficiency of this algorithm.

Bayesian Metabolic Flux Analysis (BMFA). So far we have considered frameworks in which the metabolic network is constrained to the exact steady state. In 2016, it was shown that metabolites can accumulate or deplete in a metabolic network [8] and recently MacGillivray et al. [7] studied metabolic networks under the relaxed steady state assumption through the so-called RAMP model. They have presented an argument that the exact steady state constraint (Eq (2)) on the fluxes should be relaxed to be in agreement with the stochastic nature of a cell. In 2019, a statistically relaxed steady state model was presented in Heinonen et al. [22]

$$S\mathbf{v} = \mathbf{0} + \boldsymbol{\beta}, \tag{11}$$

where $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Gamma})$ is a vector of disturbances around the steady state assumption $S\,\mathbf{v} = \mathbf{0}$. The allowed variances around the steady state are collected in the diagonal matrix $\Gamma = diag(\boldsymbol{\gamma}) = diag(\gamma_1, \ldots, \gamma_m)$. Note that by considering very small variances, $\boldsymbol{\gamma} \to 0$, the model will be compatible with the strict steady state case.

Heinonen et al. [22] implemented Eq (11) in a Bayesian framework in which multivariate Gaussian priors for fluxes were assumed. The prior mean for a flux was set to zero or to the closet value to zero considering the flux upper and lower bounds. The prior variances as a hyperparameter defines the a priori values a flux can take. A *TMVN* distribution $TMVN(\boldsymbol{\mu}, C, \mathbf{v}^{lb}, \mathbf{v}^{ub})$ was proposed as the target distribution from which fluxes $\mathbf{v}$ were sampled. For sampling purpose, Heinonen et al. [22] used the Gibbs algorithm [31], which is a MCMC algorithm suitable for Bayesian models. Detailed formulas for the mean vector $\boldsymbol{\mu}$ and the covariance matrix C can be found in [22].

In Heinonen et al. [22], the flux variables $\mathbf{v}$ were first transformed to uncorrelated variables $\tilde{\mathbf{v}} = L^{-1}(\mathbf{v} - \boldsymbol{\mu})$ using a Cholesky decomposition of the covariance matrix $C = LL^T$ to make the sampling process more efficient. Thereafter the problem was converted to sample $\tilde{\mathbf{v}}$ from the distribution $TMVN(\mathbf{0}, I, \tilde{\mathbf{v}}^{lb}, \tilde{\mathbf{v}}^{ub})$ where I is the identity matrix, $\tilde{\mathbf{v}}^{lb} = \mathbf{v}^{lb} - L\boldsymbol{\mu}$ and $\tilde{\mathbf{v}}^{ub} = \mathbf{v}^{ub} - L\boldsymbol{\mu}$. In the Gibbs algorithm an initial sample point $\tilde{v}_j^{(0)}$ is drawn from the Gaussian prior distribution for $j = 1 \ldots n$. Then, at each iteration the algorithm cyclically ($j = 1 \ldots n$) draws $\tilde{v}_j$ from the conditional posterior density $p(\tilde{v}_j \mid \tilde{\mathbf{v}}_{-j})$, where $\tilde{\mathbf{v}}_{-j}$ is a vector including all fluxes except the flux $\tilde{v}_j$. Using properties of the *TMVN* distribution, it can be shown that these conditional distributions again are within the *TMVN*, and Heinonen et al. [22] has provided closed form expressions for the upper and lower bounds $\tilde{\mathbf{v}}^{lb}$ and $\tilde{\mathbf{v}}^{ub}$.

A summary of the sampling algorithms and their main characteristics are presented in the Table 1.

Experimental setup and implementation

The four sampling algorithms (ACHR, OPTGP, CHRR and Gibbs) were applied to sample from ten metabolic models, which were obtained from the BiGG database [32]. The sampling

Table 1. A summary of sampling algorithms and their main characteristics.

Sampling algorithm	Programming language	Convergence guaranteed?	Relevant formulation
ACHR	Matlab/Python	No	Deterministic
OPTGP	Python	No	Deterministic
CHRR	Matlab	Yes	Deterministic
Gibbs/BMFA	Matlab	Yes	Stochastic
xsample()	R	Yes	Stochastic

https://doi.org/10.1371/journal.pone.0235393.t001

Table 2. Constraint-based metabolic models and run times (min) for different sampling algorithms. The m, n, n_{red} denote the number of metabolites, reactions of the full model and of the reduced model, respectively. The AFR is the Average Flux Range of the full model. The 20, 000 samples for each flux in each metabolic model were drawn on an Intel Core i7 at 2.5 GHz. In all sampling algorithms the thinning parameter was set to 1000.

Network					Run time			
Model	m	n	n_{red}	**AFR**	**ACHR (Deterministic)**	**OPTGP (Deterministic)**	**CHRR (Deterministic)**	**Gibbs/BMFA (Stochastic)**
E. coli core	72	95	87	1474	68.78 min	14.81 min	6.17 min	69.96 min
iAB_RBC_283	342	469	453	1080	99.67 min	18.53 min	9.46 min	1148.50 min
iLJ478	570	652	380	1292	91.08 min	19.83 min	7.64 min	1884.00 min
iSB619	655	743	450	1267800	96.09 min	20.83 min	9.55 min	2173.50 min
iHN637	698	785	522	1257	103.38 min	22.13 min	7.85 min	2483.50 min
iAT_PLT_636	738	1008	1008	1444	132.55 min	27.56 min	13.81 min	2244.80 min
iJN746	907	1054	652	1329200	116.40 min	24.80 min	10.67 min	3179.70 min
iSDY_1059	1888	2539	1502	1248	148.64 min	40.15 min	18.40 min	17393.00 min
iJO1366	1805	2583	1687	1242	192.22 min	38.43 min	21.93 min	18177.00 min
Recon1	2766	3741	2467	1414100	308.97 min	51.71 min	36.20 min	22268.00 min

https://doi.org/10.1371/journal.pone.0235393.t002

algorithms were applied on one core model (E. coli core) and nine genome scale metabolic models with the number of fluxes ranging from $n = 95$ to $n = 3741$. The $M = 20,000$ samples were generated for each flux in each model, with a thinning parameter of 1000 in each sampling algorithm where we kept every 1000 draw from the target distribution and discarded the rest.

The OPTGP and Gibbs algorithm sampled from the full models, while ACHR and CHRR sampled from reduced versions of the models, obtained as follow. The upper and lower bounds on the fluxes ($\mathbf{v}^{lb}$ and $\mathbf{v}^{ub}$) were changed to the minimum and maximum achievable flux values computed through flux variability analysis [33]. Then, the model was reduced by discarding the reactions which could not carry any flux (null reactions with maximum and minimum achievable values less than a threshold). Table 2 shows summary statistics for each metabolic model, including the number of reactions before (n) and after (n_{red}) reduction. Also shown are AFR values, i.e. Average Flux Range of the full models calculated by $AFR = 1/n \sum_{j=1}^{n} (v_j^{ub} - v_j^{lb})$.

In both ACHR and CHRR the number of initial iterations that have been discarded at the beginning of the sampling (warm up) was set to $M_{warm} = 20,000$. The design of the OPTGP algorithm is such that it always generates a fixed number ($2n$) of warm up points. For BMFA there is no warm up phase since its Gibbs algorithm starts out from the posteriori mode of a truncated normal distribution.

In the BMFA framework the variances γ_i around the relaxed steady state condition of Eq (11) were set to $\gamma_i = 0.0001$ ($i = 1...m$), as in Heinonen et al. [22]. Defining a nearly strict steady state condition by using such small variances (γ_i) should not have a large impact on generated samples by Gibbs algorithm. The average flux ranges (AFR) reported in Table 2 indicate that the models have different flux ranges and in the BMFA framework, the prior variances for fluxes should be adjusted according to the flux ranges. For all models except iLJ478 and iSB619, the prior variances for fluxes were set to $(min(0.5(\mathbf{v}^{ub} - \mathbf{v}^{lb}), 1000))^2$ to cover the flux ranges. To avoid numerical instabilities in the covariance matrix for the iLJ478 and iSB619, the prior variances were set to $(min(0.5(\mathbf{v}^{ub} - \mathbf{v}^{lb}), 500))^2$ and $(min(0.5(\mathbf{v}^{ub} - \mathbf{v}^{lb}), 100))^2$, respectively.

The implementations of the ACHR and CHRR algorithms available via the *sampleCbModel()* function from the COBRA toolbox (version 3.0) [18] of Matlab was used. The *bmfa()* from the COBRA toolbox was applied to generate the samples based on the Gibbs algorithm used in the BMFA. We have made a minor change in the script of the *bmfa()* function in order to allow the

user to adjust the prior variations for fluxes according to the flux ranges of a metabolic model. The samples from OPTGP algorithm were drawn using the *optGPSampler()* function from the COBRA toolbox in Python (COBRApy) [34].

Three of the algorithms (ACHR, OPTGP and CHRR) were run on a computer with an Intel Core i7 processor (2.5 GHz). The run time of the algorithms while sampling each of the ten models were measured using *tic/toc* function in Matlab and *time* function in Python which reports the elapsed "wall-clock" time (Table 2). Both OPTGP and CHRR were run in parallel on four threads, while both ACHR and Gibbs were run on a single thread, since their current implementation can not exploit parallelism. The more computationally demanding Gibbs algorithm was run on a server with 32 Cores (2.7 GHz). A pro-rata conversion was applied in order for its run time to be comparable to that of the three other algorithms. To this end 200 samples from the Gibbs sampler were generated on the Intel Core i7 processor, and the corresponding run time formed the basis of the conversion factor.

Convergence diagnostics

The $M = 20,000$ samples from each algorithm have been validated and compared in R [29]. A sample generated by a MCMC algorithm is guaranteed to be representative of the true flux distribution only if the sample chain has converged (in distribution). It is hence customary to apply one or more convergence diagnostics to avoid incorrect inference [6]. In the present study we investigated and compared four different convergence diagnostics. Distributional convergence may be assessed within a chain or across multiple chains run in parallel, started from different values inside the solution space. Not all the implementations of the algorithms used here allows the starting to be controlled, so we focused our comparison on single-chain diagnostics. The diagnostics were applied separately to each flux of a model, and we have presented the proportion of converged chains as a summary statistic.

When applying a MCMC method there are three constants that must be specified. First, the number of warm up samples, M_{warm}, determines how many samples must be discarded initially before distributional convergence is achieved. Then sampling continues for M iterations, which yields the sample $v^{(1)}, \ldots, v^{(M)}$ that is used for inference. The third constant is the so called "thinning" parameter, which in the current study was set to 1000 in all sampling algorithms. This means that only every 1000th sample from the underlying Markov chain was kept. The purpose is to reduce the autocorrelation. Note that autocorrelation in the chain per se does not invalidate the inference drawn, but it reduces the information content.

Below the four diagnostic tools are reviewed briefly. For more details the reader is referred to the background papers. We let $v^{(1)}, \ldots, v^{(M)}$ denote the sample chain for one specific flux.

Raftery and Lewis

Based on a single chain of flux samples (pilot chain), $v^{(1)}, \ldots, v^{(M)}$, the Raftery and Lewis diagnostic [35] provided an estimate of the number of iterations in the warm up phase, M_{warm}, and the required number of further iterations, N_{max}, to estimate the quantile q to within a precision of $\pm e$ with probability p. It further determined the minimum number of iterations, N_{min}, that should be run as a pilot chain assuming independent samples. Using these statistics, this test determined a dependence factor $I = (M_{warm} + N_{max})/N_{min}$ as a measure of dependency between consecutive samples (autocorrelation). Here we considered the chains with $I > 5$ as highly autocorrelated chains that were not run long enough. Here, all statistics in Raftery and Lewis diagnostic were calculated to estimate a quantile of 0.025 to within a precision of ±0.005 with probability 0.95 using the *raftery.diag()* function from the *CODA* R package [36].

Geweke

Geweke [37] proposed a single-chain convergence diagnostic which compares the average value of the first and last segments of the chain $v^{(1)}, \ldots, v^{(M)}$. Let B_1 denotes the first 10% of the samples, and B_2 denotes the last 50%. The test statistic for the Geweke diagnostic is the Z-score

$$Z = \frac{\bar{B}_1 - \bar{B}_2}{\sqrt{\sigma^2_{\bar{B}_1} + \sigma^2_{\bar{B}_2}}}, \tag{12}$$

where $\bar{B}_1$ and $\bar{B}_2$ are the averages of the two segments, and $\sigma^2_{\bar{B}_1}$ and $\sigma^2_{\bar{B}_2}$ are the associated standard errors. If the chain has converged in distribution, $\bar{B}_1$ and $\bar{B}_2$ have the same expected (mean) value. When M is sufficiently large, $\bar{B}_1$ and $\bar{B}_2$ will approximately be normally distributed, and Z will follow a standard normal distribution. Here, the Z-score was computed using the *geweke.diag()* function from the *CODA* package in R [36]. The convergence criterion for the Geweke diagnostic is $|Z| \leq 1.28$.

Interval Based Scale Reduction Factor (IPSRF)

Our third convergence diagnostic is based on the Gelman-Rubin diagnostic [38]. This is originally a multiple-chain diagnostic which compares the difference in across- and within-chain variances. The idea is that if all chains have converged the sample variances will be the same. The original Gelman-Rubin diagnostic assumes normality of the samples. As a typical flux distribution is not normal for a genome scale metabolic model [6], a modified version known as the Interval-based potential reduction factor (IPSRF) should instead be used [39].

To apply the IPSRF diagnostic to a single chain, the first and last third of the chain can be treated as two "parallel" chains. The resulting IPSRF value was estimated using the *ipsrf()* function in the MCMC diagnostics toolbox in Matlab. The test criterion is IPSRF < 0.9 or IPSRF > 1.1, in which case the single chain was considered to have not converged.

Hellinger distance

The Hellinger distance is a density based convergence diagnostic that can be used for a single chain or multiple chains [40]. The basic idea is to compare the flux density estimated from the first third segment of the chain, $p_1(v)$, with that of the last third segment, $p_3(v)$. The probability densities p_1 and p_3 are calculated using the *densityfun()* function of the *statip* package in R [41]. The Hellinger distance statistic is defined as

$$HD(p_1, p_3) = \sqrt{\frac{1}{2}\int_{-\infty}^{\infty}\left(\sqrt{p_1(v)} - \sqrt{p_3(v)}\right)^2 dv}. \tag{13}$$

It is a proper metric, symmetric in p_1 and p_3. Further, it is bounded by $0 \leq HD \leq 1$, where 0 indicates no divergence and 1 indicates no common support between the two distributions. As suggested by Boone et al. [40], if the Hellinger distance between the two probability density functions of two segments was less than 0.1 ($HD \leq 0.1$), then the chain has been considered to have converged else not. We wrote a script in R to calculate the Hellinger distance where we used the *integral()* from the *pracma* package [42].

Comparison of algorithms

Correlation coefficient

The two most important statistical summaries of a sample $v^{(1)}, \ldots, v^{(M)}$ are its mean and variance:

$$\bar{v} = \frac{1}{M}\sum_{l=1}^{M} v^{(l)} \text{ and } s^2 = \frac{1}{M-1}\sum_{l=1}^{M}(v^{(l)} - \bar{v})^2. \quad (14)$$

If two sampling algorithms yield the same flux distributions, they should give the same values of $\bar{v}$ (and similarly for s^2) for a given reaction. We compare algorithms in terms of their Pearson correlation across reactions for both of these quantities. In term of the sample average the Pearson correlation between Algorithm 1 and 2 is given as

$$r = \frac{\sum_{j=1}^{n}[(\bar{v}_{j,1} - \bar{\bar{v}}_1)(\bar{v}_{j,2} - \bar{\bar{v}}_2)}{\sqrt{\sum_{j=1}^{n}(\bar{v}_{j,1} - \bar{\bar{v}}_1)^2(\bar{v}_{j,2} - \bar{\bar{v}}_2)^2}} \quad (15)$$

where $\bar{v}_{j,1}$ is the sample average for the jth flux, and $\bar{\bar{v}}_1 = \frac{1}{n}\sum_{j=1}^{n}\bar{v}_{j,1}$ is the across-flux average, both for Algorithm 1 (and similar quantities for Algorithm 2). The Pearson correlation is well suited as a measure of association because the flux average $\bar{v}$ will be approximately normally distributed by the central limit theorem. Further, r varies between −1 and +1. A perfect positive (linear) association is indicated by a value of + 1, while 0 represents no association [43].

We used CHRR as a reference in the comparison with the three other algorithms. Outliers were determined in the following way, and subsequently omitted when calculating the Pearson correlation. In the case of CHRR versus ACHR, say, a reaction was considered an outlier if the difference $\bar{v}^{\text{CHRR}} - \bar{v}^{\text{ACHR}}$ exceeded 2 standard deviations (of this difference, across reactions). A similar outlier criterion, based on $s^{\text{CHRR}} - s^{\text{ACHR}}$, was applied on the sample standard deviations s. The set of omitted reactions includes the outliers in both the means and the standard deviations of the flux values. The value of the Pearson correlation, r, is calculated using the *cor()* function from the *stats* package in R [29].

Kullback-Leibler divergence

We also compared the distributional shape resulting from different algorithms, using the Kullback-Leibler divergence (KLD) as a measure of dissimilarity. Let $p_1(v)$ and $p_2(v)$ denote flux densities resulting from two algorithms, and define

$$KLD(p_2|p_1) = \int_{-\infty}^{\infty} \ln\left(\frac{p_1(v)}{p_2(v)}\right) p_1(v)dv. \quad (16)$$

It may be shown that $KLD(p_2|p_1) \geq 0$, and that it is zero only if p_1 and p_2 are identical functions [44]. Note that $KLD(p_2|p_1)$ is not symmetric in p_1 and p_2, we will refer to p_1 as the *reference*. The CHRR will be used as the reference against the three other methods. A script has been written in R to calculate the KLD in Eq (16). The probability densities p_1 and p_3 are calculated using the *density()* function of the *stats* package in R [29].

We classified the accuracy of the approximation as *good* agreement $KLD < 0.05$, *medium* agreement $0.05 \leq KLD \leq 0.5$ and *poor* match $KLD > 0.5$. This classification was adopted from De Martino et al. [17].

Effective sample size

The effective sample size (ESS) of an autocorrelated MCMC sample of size M is the equivalent number of independent draws from the target distribution. Gelman et al. [45] defines the effective sample size (for sample mean) as

$$ESS = \frac{M}{1 + 2\sum_{k=1}^{\infty} \rho_k}, \tag{17}$$

where ρ_k is the autocorrelation at lag k. From a given sample $v^{(1)}, \ldots, v^{(M)}$ the estimate of ρ_k is given as

$$\rho_k = \frac{\frac{1}{M-k}\sum_{l=1}^{M-k}(v^{(l)} - \bar{v})(v^{(l+k)} - \bar{v})}{\frac{1}{M}\sum_{l=1}^{M}(v^{(l)} - \bar{v})^2}, \tag{18}$$

where $\bar{v}$ is the mean of the samples [46]. Due to the random walk like behaviour of MCMC algorithms, one typically has $0 \leq \rho_k \leq 1$ which implies $ESS \leq M$. A low value of ESS/M indicates that the algorithm generates highly autocorrelated samples (large ρ_k). The higher the autocorrelation is, the less information about the target distribution is contained in a sample of fixed size. Increasing the value of the thinning parameter will reduce the autocorrelation, but this gain comes at a computational cost.

In order to compare the efficiency of two algorithms in terms of ESS, the computation time must be taken into account since one algorithm may generate a larger number of independent samples slowly, while another may generate highly autocorrelated samples fast. The efficiency of each algorithm in generating independent samples per time unit for each individual flux was measured by

$$E = \frac{ESS}{Run\ time}, \tag{19}$$

where the ESS value has been calculated with the *effectiveSize()* function from the *CODA* R package [36] and the run time is reported in Table 2 for each algorithm across the ten models.

Results

The sampling algorithms have been compared on the ten metabolic models using the criteria described earlier. First, the degree of convergence was investigated. Secondly, the flux densities generated by the different algorithms were compared. Finally, the computational efficiency of the algorithms was assessed.

We were only able to successfully apply the *xsample()* algorithm in one (E. coli core) out of the ten models (details given below). Hence, the comparison of algorithms was performed only between CHRR, ACHR, OPTGP and the Gibbs sampler.

Convergence of algorithms

For all ten models, the convergence of the generated samples was assessed (by reaction) via the four single-chain convergence diagnostics. Fig 2 shows the percentage of reactions that failed for each of the Raftery and Lewis diagnostic ($I > 5$), Geweke test ($|Z|>1.28$), IPSRF test (IPSRF<0.9 or IPSRF>1.1) and Hellinger distance test ($HD > 0.1$). In the majority of the models, CHRR was the algorithm with the least convergence problems. All four diagnostics agree on this, but when it comes to the ranking of ACHR, OPTGP and Gibbs sampler, the diagnostics tell less coherent stories, so it is difficult draw general conclusions. ACHR did

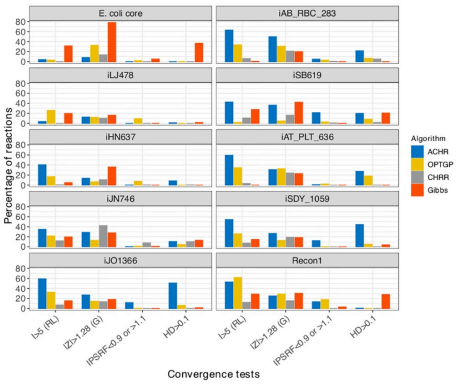


Fig 2. Four convergence diagnostics across four algorithms and ten models. The vertical axis shows the proportions of reactions in each model rejected by the different convergence tests: Raftery and Lewis (RL), Geweke (G), IPSRF and Hellinger distance (HD) on the horizontal axis.

https://doi.org/10.1371/journal.pone.0235393.g002

however seem to have convergence problems for many models, and the Gibbs sampler had problems for E. coli core in particular.

We only show summaries statistics for the diagnostics. It was also possible to inspect convergence for individual reactions, and when doing so we found that it is not necessarily the same reactions that failed to converge according to the different diagnostics. Therefore a combination of convergence diagnostics should be used to make a certain decision about sampling convergence. Apparently, the IPSRF test is more liberal in accepting convergence, but it should be noted that this conclusion is specific to our chosen settings (the default) for that diagnostic.

Comparison means and standard deviations

Fig 3 compares CHRR against each of three other algorithms in terms of sample means ($\bar{v}$) and standard deviations (s) as given by Eq (14). The figure only shows four models (E. coli core, iHN637, iAT_PLT_636 and Recon1), but plots for the remaining six models are provided in the online Supplementary (S1 Fig).

In general, the four algorithms returned very similar sample means $\bar{v}$, as can be seen from the fact that the points in the plot lie along the identity line. This is also reflected in a Pearson correlation close to $r = 1$. The exception is the Gibbs sampler (versus CHRR), especially for the Recon1 model. For this model the range of $\bar{v}$ values was much smaller for the Gibbs sampler than for CHRR. Note, however, that the Pearson correlation is substantial ($r = 0.50$), which implies that there is still a strong linear relationship, although with slope different from 1. The same effect, but to a much smaller degree, is also observed for the iAT_PLT_636 model. The effect is known as "shrinkage-toward-zero", and is caused by the prior distribution applied to fluxes in the Gibbs algorithm. Ideally, such priors should be made "non-informative" by choosing the prior variance sufficiently large, but in the case of Recon1 it was not possible to make the prior cover the full flux range (AFR in Table 2) without encountering numerical problems in the Gibbs sampler.

Fig 3 includes also the reactions for which the algorithms did not converge, but reactions for which at least one of the two algorithms in a comparison failed the Geweke test are marked

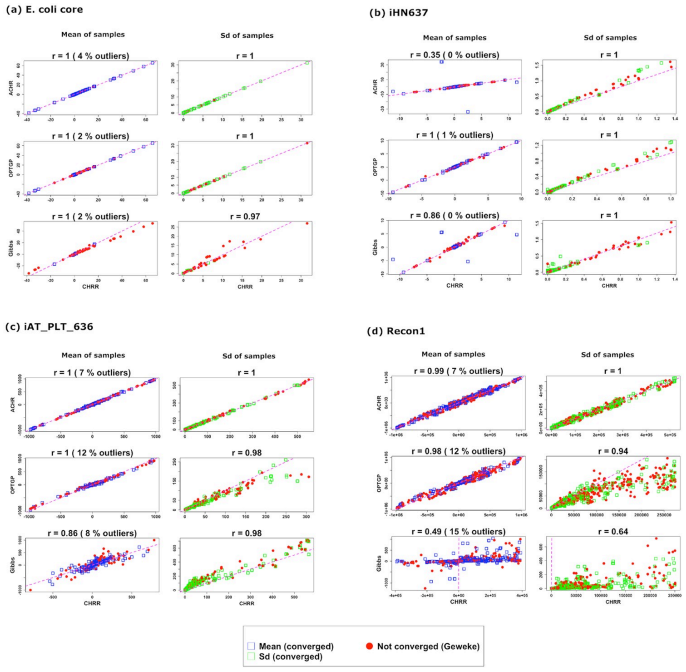


Fig 3. Scatter plot of sample means (blue) and standard deviations *s* (green) for ACHR, OPTGP and the Gibbs sampler (vertical axis) versus CHRR (horizontal axis) for four models. Sample means ($\bar{v}$) and standard deviations (s) are calculated according to Eq (14). The Pearson correlation r is shown on top of each scatter plot, and the proportion of outliers removed is given in parenthesis. The sample means and standard deviations marked in red correspond to the reactions for which at least one of the two algorithms in a comparison failed the Geweke test. The identity line (pink dashed) is included to ease comparison.

https://doi.org/10.1371/journal.pone.0235393.g003

in red. For E. coli core there is a tendency that the largest fluxes (negative or positive) face convergence problems for the Gibbs sampler, while for the other algorithms and models there is no such clear pattern. Recall that Fig 2 summarized convergence for each algorithm separately.

The standard deviations from ACHR, OPTGP and CHRR agree well in general, i.e their green points lie close to the identity line. For the Recon1 model, OPTGP has lower variance than CHRR, and there is more spread ($r = 0.94$). The Gibbs sampler is in fairly good agreement with CHRR, but for Recon1 its standard deviations are much smaller than those from the Gibbs sampler. This reflects the shrinkage-toward-zero effect caused by the narrow Bayesian priors applied in the Gibbs sampler, as discussed above. For iAT_PLT_636 the standard deviations from the Gibbs sampler exceed those of CHRR, indicating that the Gibbs sampler is better (than CHRR) able to explore the flux space for this model.

The % outliers shown on top of each plot indicates the percentage of reactions for which large differences have been observed between the sample means or standard deviations from two algorithms. Note that in the plots of standard deviations the reactions with the standard deviations smaller than 99% quantile have been included.

Comparison of marginal distributions

While Fig 3 compares algorithms only in terms of sample mean and standard deviation, Fig 4 compares the full distributional shape of the flux densities. The figure shows cumulative distribution function of KLD (Kullback-Leibler divergence) across reactions, where CHRR is used as the reference for each of ACHR, OPTGP and the Gibbs sampler. Only reactions for which both algorithms in a comparison, ACHR and CHRR say, converged according to the Geweke diagnostic are included in the figure. The KLD is affected by discrepancies in means and standard deviations, so any off-diagonal reactions in Fig 3 will result in a large KLD value. In addition, Fig 4 shows differences caused by different degree of skewness in the densities.

Before discussing the results in Fig 4, we recall the qualitative (*good–medium–poor*) scale of the KL divergence (KLD). To get a visual impression of what this amounts to in a density plot, Fig 5 shows flux densities and KLD values for the Fumarase mitochondrial reaction (v_{553}) of the iAT_PLT_636 model. According to this KLD scale ACHR has a good similarity to CHRR ($KLD = 0.01 < 0.05$), and OPTGP has a medium similarity to CHRR ($0.05 < KLD = 0.43 < 0.5$) while the Gibbs algorithm has a poor similarity to CHRR ($KLD = 0.82 > 0.5$). Returning to Fig 4, it is seen that almost all of the reactions of the iHN637 model are in the *good* category for all three algorithms. The E. coli core model is the only model for which both the ACHR, OPTGP and the Gibbs algorithm present good consistency with CHRR for all reactions ($KLD < 0.05$). For the other eight model, however, a large proportion of the reactions are in the *poor* category. Taking iAB_RBC_283 as an example, for the Gibbs sampler approximately 50% of the reactions have $KLD > 0.5$. For ACHR and OPTGP the proportion with $KLD > 0.5$ is somewhat lower (10-15%). In Recon1 a large proportion of the reactions are outside the range of the horizontal axis for the Gibbs sampler, and hence do not show in the plot. These reactions are affected by the shrinkage-towards-zero effect displayed in Fig 3.

The general message from Fig 4 is that ACHR is producing flux distributions most similar to CHRR. This conclusion is based on the fact that its cumulative distribution curve (cyan) lies above the two others. The latter does not preclude ACHR having a lower *KLD* value than OPTGP, say, for individual reactions, but it is a statement that is valid as a summary across all reactions. For the majority of the ten models, OPTGP was much closer to ACHR in

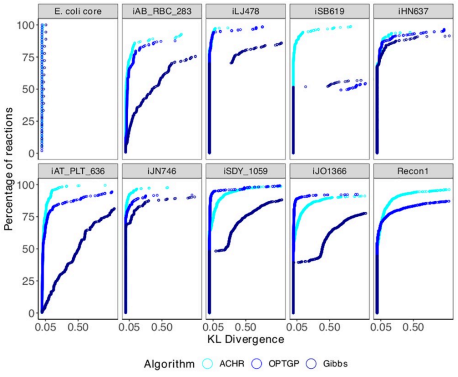


Fig 4. Comparison of flux densities between algorithms by model in terms of the KL divergence. Each plot shows the cumulative distribution functions of KLD across reactions, as defined in Eq (16) with CHRR as the reference.

https://doi.org/10.1371/journal.pone.0235393.g004

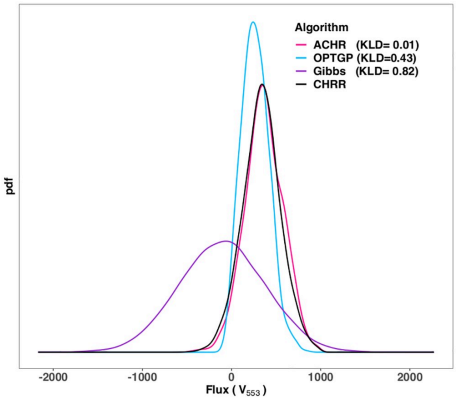


Fig 5. Flux densities resulting from different algorithms and corresponding KLD values (relative to CHRR). The reaction shown is the Fumarase mitochondrial reaction (v_{553}) of the iAT_PLT_636 model.

https://doi.org/10.1371/journal.pone.0235393.g005

comparison to the Gibbs sampler. The only exception to this was the iSB619 model for which the cumulative distribution function for OPTGP lies below that of Gibbs sampler. In conclusion, ACHR has the highest consistency with CHRR, followed by OPTGP. The Gibbs sampler is ranked as the least consistent method with CHRR. The latter is most likely due to the shrinkage-towards-zero effect caused by the use of informative priors in the Gibbs sampler.

Sampling efficiency

Fig 6 compares the cumulative distribution functions for the efficiency measure E, given by Eq (19), of the different metabolic models, separately for each sampler. Recall that for two curves that never cross each other, such as the yellow (E. coli core) and any of the blue curves in Panel a), the distribution of E for one model (blue) is stochastically larger than the other (yellow).

The models have been categorized in four groups based on the number of reactions: $n < 100$, $100 < n < 1000$, $1000 < n < 3000$ and $n > 3000$. The yellow curve (E. coli core) has the highest effective sample size per time unit for all four algorithms. This was expected as E. coli core is the smallest model ($n = 95$ reactions). If it can be assumed that the number of metabolites (m) is proportional to n, the computation time for the matrix vector product $S\,\mathbf{v}$ in Eq (2) grows as n^2 (ignoring that S is a sparse matrix). Assuming that the product $S\,\mathbf{v}$ constitutes the main computational task of any of the sampling algorithm, we expect E will decrease proportionally to n^{-2} as n increases. This theoretical expectation is confirmed, at least qualitatively, in Fig 6 for all four sampling algorithms. The largest model, Recon1, has very low sampling efficiency.

When comparing the four algorithms, we first note that the scales on the horizontal axes differ across panels in Fig 6. The CHRR has the highest sampling efficiency, followed by the ACHR, then by the OPTGP, and finally by the Gibbs sampler. Note that ACHR and CHRR sample the reduced models (of size n_{red}), while OPTGP and Gibbs sample the full models (of size n). We see from Table 2 that n/n_{red} is never larger than 2, and attempting to account for model size by multiplying the efficiency of the Gibbs sampler by 4, it is observed that the Gibbs algorithm is still the algorithm with least efficiency.

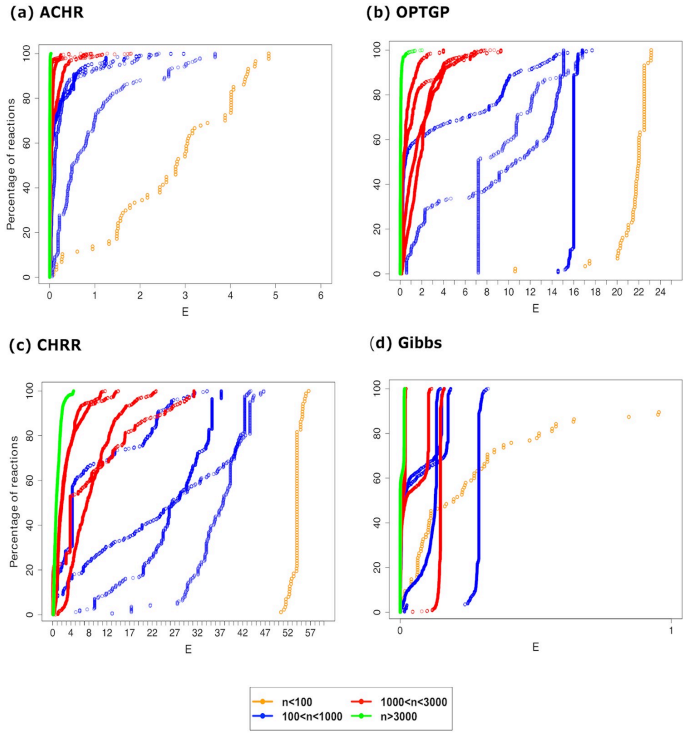


Fig 6. Comparison of sampling efficiency across four algorithms and ten models. The vertical axis shows the proportions of reactions being less than a given value of the efficiency measure E on the horizontal axis. The ten different curves correspond to the ten models which are classified in four groups according to their number of reactions (see legend).

https://doi.org/10.1371/journal.pone.0235393.g006

To further illustrate how sampling efficiency depends on model size we computed the time it takes to generate 100 independent (uncorrelated) samples. This was computed as $100(mean(ESS))^{-1} * (\text{Run time}) = 100(mean(E))^{-1}$ where run time is provided in Table 2 and E is given by (19), and the results are shown in Fig 7. As expected, the computation time in general increases with model size, but there are exceptions to this (values of n in the rage 1000 to 2500 for OPTGP). These exceptions show that there are other aspects of a model than n that determines sampling efficiency. For most of the models, ACHR and the Gibbs sampler (right vertical axis) are slower than OPTGP and CHRR (left axis). We observe that ACHR is the slowest algorithm to generate 100 independent samples, closely followed by the Gibbs sampler which we recall performs sampling on the full models.

To shed further light on differences in sampling efficiency between algorithms, we inspected the autocorrelation functions ρ_k, given by Eq (18), for two individual reactions (Fig 8). Also shown in the figure is the corresponding measure of effective sample size (ESS) defined in

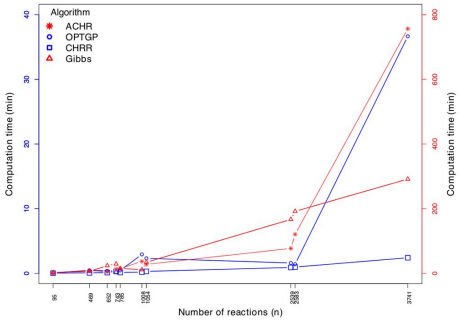


Fig 7. Computation time needed to generate 100 uncorrelated samples by model size (*n*) and algorithm. Each value of *n* shown on the horizontal axis correspond to one of the ten metabolic models, and is taken from Table 2. The left vertical axis is used for OPTGP and CHRR, while the right vertical axis belongs to ACHR and the Gibbs sampler.

https://doi.org/10.1371/journal.pone.0235393.g007

Eq (17). The algorithms differ widely in how fast ρ_k decayed as a function of k, and consequently, in the value of ESS. We note, however, that the numerical values shown in Fig 8 are specific to the value of the thinning parameter (1000) used, so absolute values are not relevant. The ACHR was the algorithm with the lowest ESS, followed by OPTGP. For the iAT_PLT_636

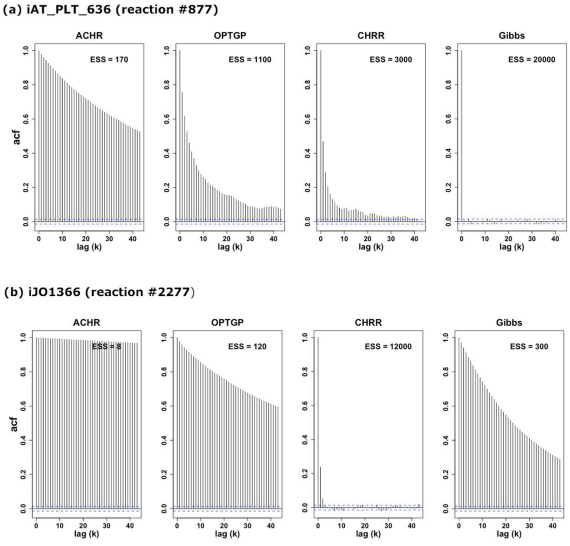


Fig 8. Autocorrelation ρ_k (acf) by lag for the flux v_{877} in the iAT_PLT_636 model (Panel a) and the flux v_{2277} in the iJO1366 model (Panel b) for each sampling algorithm. These fluxes, v_{877} and v_{2277}, correspond to 1D-myo-Inositol 4-phosphate phosphohydrolase and Ribose-1,5-bisphosphokinase reactions in the iAT_PLT_636 and iJO1366, respectively. The dotted blue lines indicate lag-wise 95% confidence intervals (CIs).

https://doi.org/10.1371/journal.pone.0235393.g008

model (Panel a), the Gibbs sampler yields an almost uncorrelated chain, meaning that the thinning parameter could have been set to a lower number than 1000, as we currently are discarding some useful information about the flux distributions. For iJO1366 (Panel b), CHRR had almost no autocorrelation, while the Gibbs sampler had a substantial autocorrelation. This shows that the details of the model plays an important role in determining which algorithms is the most efficient in terms of generating independent samples.

Performance of *xsample()*

The *xsample()* function in R [29] was attempted on the reduced versions of the ten metabolic models, but we were only able to successfully run it for the E. coli core model. The reason for the problem may be the large variation in flux ranges for the nine other models. For instance, the minimum and maximum of the flux ranges were of orders 10^{-6} and 10^{3}, respectively, in the reduced version of iAB_RBC_283. The jump length is a compromise to sample over these 9 orders of magnitude in which a small jump length is needed for the fluxes with small range and a large jump length is needed for the fluxes with large range. In the *xsample()* function, the jump lengths which are the diagonal elements of the matrix $\mathbf{\Omega}$ in Eq (10) were set to $0.5(\mathbf{v}_{red}^{ub} - \mathbf{v}_{red}^{lb})$ in order to scale them to the range of the fluxes. However having large step lengths made the sampling algorithm very inefficient since a lot of mirroring steps were required and the algorithm rejected many draws in each iteration.

We also tried $0.01(\mathbf{v}_{red}^{ub} - \mathbf{v}_{red}^{lb})$ for the jump lengths, and the algorithm was able to sample all the models, albeit very slow. Checking the generated samples, we observed that since the jump lengths were small the sampler moved barely from the initial flux vector. Due to this the generated samples were highly autocorrelated and we have not included them in the further analysis. So the best choice of jump lengths as a hyper parameter in the *xsample()* was not trivial and one has to use a cluster with simply a lot of brute computing power to deal with this.

For the 20, 000 samples that were successfully obtained from the E. coli core model, applying the jump lengths $0.5(\mathbf{v}_{red}^{ub} - \mathbf{v}_{red}^{lb})$, a statistical analysis was performed similar to that above for the other algorithms. The rates of non-convergence according to the four diagnostic tests were: 0% (Raftery and Lewis), 18.9% (Geweke), 0% IPSRF, and 0% (Hellinger distance). These are lower than for the other algorithms, except for the Geweke test, but still considerably lower than the Gibbs sampler (Fig 1). However, the run time of the *xsample()* to generate the samples for the E. coli core model was considerably larger than the Gibbs sampler. The scatterplots of sample means and standard deviations against CHRR look qualitatively similar to those for the Gibbs sampler in Fig 3.

Discussion and conclusion

In this study we have reviewed and compared five MC sampling algorithms for constraint-based modeling of metabolic networks (Table 1). The algorithms have been classified as allowing either a deterministic and stochastic formulation of the metabolic model (Fig 1). In the stochastic formulation, which is the most general, the steady state assumption can be relaxed and noisy flux observations can be incorporated in the model. However, to ensure a fair comparison of algorithms, all experiments were done considering no flux measurements.

We have reviewed and compared four standard convergence diagnostics that can be used to check if the algorithms have been run for sufficiently many iterations that the samples come from the target flux distribution. Finally, important metrics for comparing the algorithms have been similarity of flux distributions and computational efficiency.

The algorithms have been applied to ten genome scale metabolic networks (Table 2). However, in case of the *xsample()* algorithm we were only able to successfully apply it to a single

model (E.coli), so our comparison was done with only four algorithms (ACHR, OPTGP, CHRR and the Gibbs sampler). An efficient sampling algorithm which allows the stochastic framework of Van den Meersche et al. [20] to be applied at genome scale is thus lacking.

Comparing the algorithms in terms of convergence, the CHRR has the least convergence problems. This result is in agreement with the findings in Herrmann et al. [6] in which three algorithms ACHR, OPTGP and CHRR are compared in terms of convergence and run time. We have found that the set of reactions which fail the convergence criterion is not necessarily the same across different diagnostic tests. Also, the proportion of reactions for which a test fails can be substantial (Fig 2). Hence, from a practical perspective it does not seem feasible to require that all reactions have converged before the output from an algorithm can be trusted. Instead, the focus should be on reactions of special interest, and for those reactions one can follow the recommendation of Herrmann et al. [6] that the whole suite of diagnostics should be satisfied. Further, in our comparison of algorithms in Fig 3, there seems to be good agreement between algorithms also for reactions that have not converged.

Convergence to the target distribution is not guaranteed for ACHR and OPTGP, while for CHRR convergence is guaranteed due to its Markovian nature. For this the other algorithms were compared against CHRR. We found that ACHR generates the most similar (marginal) flux distributions to that of CHRR, followed by OPTGP. The Gibbs sampler deviated most from CHRR, which probably is due to the informative prior distribution imposed on some of the models.

When comparing the algorithms in terms of computational efficiency, we found that the CHRR method outperforms the three other algorithms by generating the highest number of independent samples per time unit for each flux. The main parameter that characterize a model is the number of reactions (n), but we have also observed that there are other aspects of a model that affect the performance of an algorithm.

Hamiltonian Monte Carlo (HMC) [47] is another sampling technique for exploring the posterior distribution in the Bayesian framework. In Heinonen et al. [22], the HMC was reported to be inefficient compared to Gibbs sampler in the genome scales metabolic models. We tried to apply HMC via the Template Model Builder (*TMB*) package [48] which is a statistical software platform in R [29]. Using the interval based scale reduction factor (IPSRF) [39] as the convergence criterion, we did not get reliable convergence. Most likely, the feasible truncated density region for high dimension models ($n > 1000$) was extremely narrow causing the HMC constantly to hit the boundaries of the polytope.

Our study ranks the CHRR as the best sampling algorithm for cases such as Fig 1b and 1c in which the steady state assumption has to be satisfied strictly and uncertainties in the observed flux values (if there are any) are negligible. The CHRR is currently available in Matlab. If an open-source programming language is preferred, a good alternative is the OPTGP, which is available in Python. For the stochastic formulation, such as Fig 1e, in which the flux observation and their uncertainty are encoded in a model compatible with relaxed steady state assumption, the only sampling algorithm applicable at the genome scale is the Gibbs sampler which is currently available in Matlab. However, this algorithm performs poorly in terms of sampling efficiency.

Supporting information

S1 Fig. Scatter plot of sample means and standard deviations. The plots are for ACHR, OPTGP and the Gibbs sampler (vertical axis) versus CHRR (horizontal axis) for six models. Sample means ($\bar{v}$) (blue) and standard deviations (s) (green) are calculated according to the formulas in the manuscript. The Pearson correlation r is shown on top of each scatter plot,

and the proportion of outliers removed is given in parenthesis. The sample means and standard deviations marked in red correspond to the reactions for which at least one of the two algorithms in a comparison failed the Geweke test. The identity line (pink dashed) is included to ease comparison.
(PDF)

Author Contributions

Data curation: Shirin Fallahi.

Methodology: Shirin Fallahi, Hans J. Skaug, Guttorm Alendal.

Supervision: Hans J. Skaug, Guttorm Alendal.

Visualization: Shirin Fallahi.

Writing – original draft: Shirin Fallahi.

Writing – review & editing: Shirin Fallahi, Hans J. Skaug, Guttorm Alendal.

References

1. Nelson DL, Cox MM. Lehninger principles of biochemistry. New York: WH Freeman, 2009.
2. Palsson BØ. Systems biology. Cambridge University Press, 2015.
3. Strogatz SH. Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering. CRC Press, 2018.
4. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis?. Nature biotechnology, 2010, 28(3):245–248. https://doi.org/10.1038/nbt.1614
5. Grünbaum B, Shephard GC. Convex polytopes. Bulletin of the London Mathematical Society, 1969, 1 (3):257–300
6. Herrmann HA, Dyson BC, Vass L, Johnson GN, Schwartz JM. Flux sampling is a powerful tool to study metabolism under changing environmental conditions. NPJ systems biology and applications, 5(1):1–8.
7. MacGillivray M, Ko A, Gruber E, Sawyer M, Almaas E, Holder A. Robust analysis of fluxes in genome–scale metabolic pathways. Nature Publishing Group, Scientific reports, 2017, 7(1):1–20.
8. Pakula TM, Nygren H, Barth D, Heinonen M, Castillo S, Penttilä M, Arvas M. Genome wide analysis of protein production load in Trichoderma reesei. Biotechnology for biofuels, 2016, 9(1):132. https://doi.org/10.1186/s13068-016-0547-5
9. Dyer ME, Frieze AM. On the complexity of computing the volume of a polyhedron. SIAM Journal on Computing, 1988, 17(5):967–974. https://doi.org/10.1137/0217060
10. Krauth W. Introduction to monte carlo algorithms. In Advances in Computer Simulation, Springer, 1989, 1–35
11. Wiback SJ, Famili I, Greenberg HJ, Palsson BØ. Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. Journal of theoretical biology, 2004, 228(4):437–447. https://doi.org/10.1016/j.jtbi.2004.02.006
12. Bélisle CJ, Romeijn HE, Smith RL. Hit-and-run algorithms for generating multivariate distributions. Mathematics of Operations Research, 1993, (2):255–266.
13. Turchin VF. On the computation of multidimensional integrals by the Monte–Carlo method. Theory of Probability and Its Applications, 1971, 16(4):720–724. https://doi.org/10.1137/1116083
14. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabási AL. Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature, 2004, 427(6977):839. https://doi.org/10.1038/nature02289
15. Kaufman DE, Smith RL. Direction choice for accelerated convergence in hit-and-run sampling. Operations Research, 1998, 46(1):84–95. https://doi.org/10.1287/opre.46.1.84
16. Megchelenbrink W, Huynen M, Marchiori E. optGpSampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. PloS ONE, 2014, 9(2). https://doi.org/10.1371/journal.pone.0086587 PMID: 24551039
17. De Martino D, Mori M, Parisi V. Uniform sampling of steady states in metabolic networks: heterogeneous scales and rounding. PloS one, 2015, 10(4).

18. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nature protocols, 2007, 2(3):727. https://doi.org/10.1038/nprot.2007.99

19. Haraldsdóttir HS, Cousins B, Thiele I, Fleming RM, Vempala S. CHRR: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. Bioinformatics, 2017, 33(11):1741–1743. https://doi.org/10.1093/bioinformatics/btx052

20. Van den Meersche K, Soetaert K, Van Oevelen D. xsample (): an R function for sampling linear inverse problems. Journal of Statistical Software, 2009, 30 (Code Snippet 1).

21. Soetaert K, Van den Meersche K, van Oevelen D. limSolve: Solving linear inverse models. Journal of Statistical Software, Code Snippets, 2009, 30(1).

22. Heinonen M, Osmala M, Mannerström H, Wallenius J, Kaski S, Rousu J, et al. Bayesian metabolic flux analysis reveals intracellular flux couplings. Bioinformatics, 2019, 35(14), pp. i548–i557. https://doi.org/10.1093/bioinformatics/btz315 PMID: 31510676

23. Altmann Y, McLaughlin S, Dobigeon N. Sampling from a multivariate Gaussian distribution truncated on a simplex: a review. IEEE, 2014, Workshop on Statistical Signal Processing (SSP),113–116.

24. Gilks WR, Richardson S, Spiegelhalter DJ. Introducing markov chain monte carlo. Markov chain Monte Carlo in practice, 1996, 1:19.

25. Lovász L. Hit-and-run mixes fast. Mathematical Programming, Springer,1999, 86(3):443–461. https://doi.org/10.1007/s101070050099

26. Zhang Y, Gao L. On numerical solution of the maximum volume ellipsoid problem. SIAM Journal on Optimization, 2003, 14(1):53–76. https://doi.org/10.1137/S1052623401397230

27. Telgen J. Private communication with A. Boneh. 1980.

28. Berbee HC, Boender CG, Ran AR, Scheffer CL, Smith RL, Telgen J. Hit-and-run algorithms for the identification of nonredundant linear inequalities. Mathematical Programming, 1987, 37(2):184–207. https://doi.org/10.1007/BF02591694

29. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013. http://www.R-project.org/.

30. Roberts GO. Markov chain concepts related to sampling algorithms. Markov chain Monte Carlo in practice, 1996, 57:45–58.

31. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence, 1984, (6):721–741.

32. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, Ebrahim A, Palsson BØ, Lewis NE. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic acids research, 2015, 44(D1):D515–D522.

33. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Elsevier, Metabolic engineering, 2003, 5(4):264–276. https://doi.org/10.1016/j.ymben.2003.09.002

34. Ebrahim A, Lerman JA, Palsson BØ, Hyduke DR. COBRApy: constraints-based reconstruction and analysis for python. BMC systems biology, 2013, 7(1):74. https://doi.org/10.1186/1752-0509-7-74

35. Raftery AE, Lewis S. How many iterations in the Gibbs sampler? Bernardo J. M., Berger J., Dawid A. P., Smith A. F. M. 4th edn, Oxford: Bayesian Statistics, 1992.

36. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. R news, 2006; 6(1):7–11.

37. Geweke J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Oxford: J. O. Berger, A. P. Dawid, Smith A. F. M. (ed. 4) Bayesian Statistics: Clarendon Press, 1991.

38. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Institute of Mathematical Statistics, Statistical science, 1992, 7(4):457–472.

39. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics, 1998, 7(4):434–55. https://doi.org/10.1080/10618600.1998.10474787

40. Boone EL, Merrick JR, Krachey MJ. A Hellinger distance approach to MCMC diagnostics. Journal of Statistical Computation and Simulation, 2014, 84(4):833–49. https://doi.org/10.1080/00949655.2012.729588

41. Poncet P., R Core Team. statip: Statistical Functions for Probability Distributions and Regression. R package version 0.2.3., 2019.

42. Borchers HW. Package 'pracma': Practical numerical math functions. R package version 2.2.9, 2019; 2(1).

43. Devore JL, Berk KN. Modern mathematical statistics with applications. Springer, 2012, 249–252.
44. MacKay DJ, Mac Kay DJ. Information theory, inference and learning algorithms. Cambridge university press, 2003, 34–35.
45. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. Chapman and Hall/CRC, 2013, ed. 3, 286–287.
46. Venables WN, Ripley BD. Modern applied statistics with S-PLUS. Springer Science and Business Media, 2013, 352–355.
47. Neal RM. MCMC using Hamiltonian dynamics. Handbook of markov chain monte carlo, 2011, 2(11):2.
48. Kristensen K, Nielsen A, Berg CW, Skaug H, Bell B. Template model builder TMB. J. Stat. Softw, 2015, 70:1–21

Appendix to paper A.

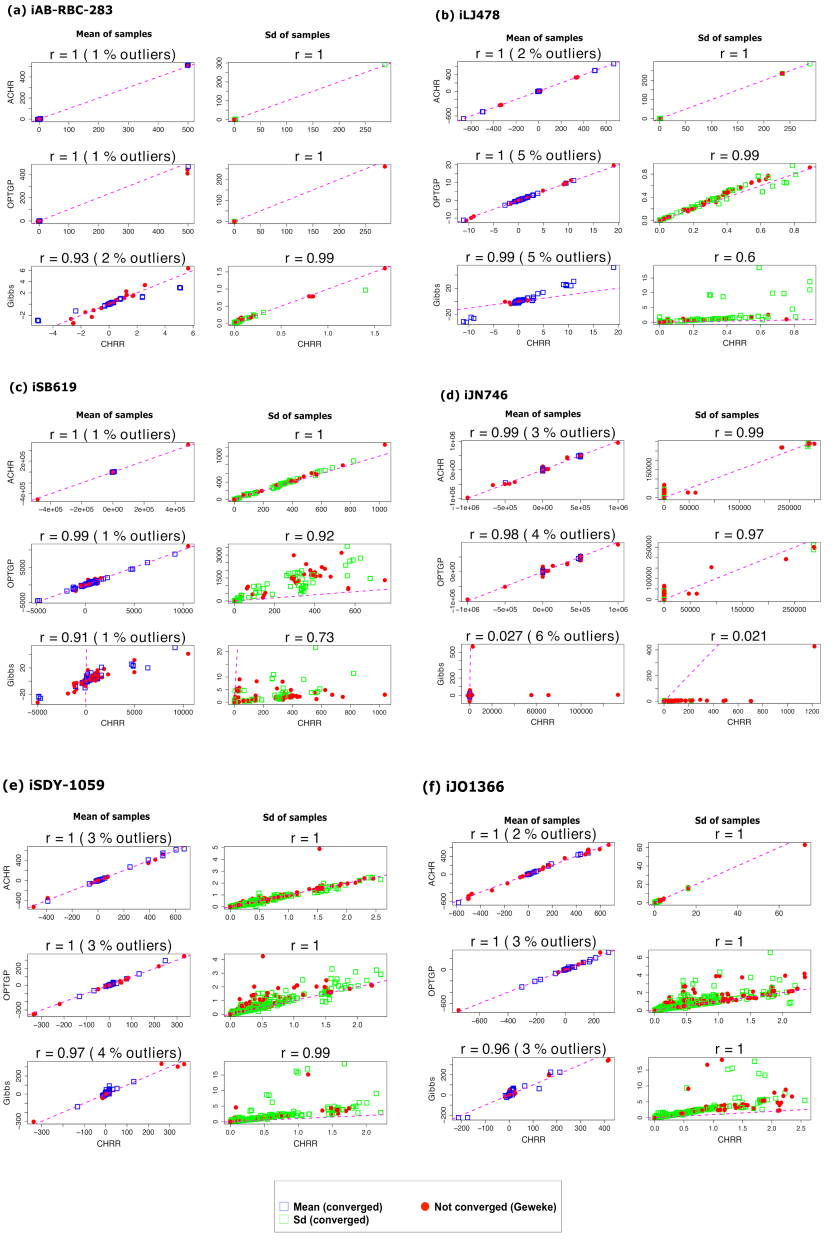


Fig S1. Scatter plot of sample means and standard deviations. The plots are for ACHR, OPTGP and the Gibbs sampler (vertical axis) versus CHRR (horizontal axis) for six models. Sample means ($\bar{v}$) (blue) and standard deviations (s) (green) are calculated according to the formulas in the manuscript. The Pearson correlation r is shown on top of each scatter plot, and the proportion of outliers removed is given in parenthesis. The sample means and standard deviations marked in red correspond to the reactions for which at least one of the two algorithms in a comparison failed the Geweke test. The identity line (pink dashed) is included to ease comparison.

Paper B

Evaluating model reduction under parameter uncertainty

Håvard G. Frøysa, **Shirin Fallahi** and Nello Blaser
BMC Systems Biology, **12**:79 (2018)

Frøysa *et al. BMC Systems Biology* (2018) 12:79
https://doi.org/10.1186/s12918-018-0602-x

BMC Systems Biology

METHODOLOGY ARTICLE Open Access

Evaluating model reduction under parameter uncertainty

Håvard G. Frøysa*, Shirin Fallahi and Nello Blaser

Abstract

Background: The dynamics of biochemical networks can be modelled by systems of ordinary differential equations. However, these networks are typically large and contain many parameters. Therefore model reduction procedures, such as lumping, sensitivity analysis and time-scale separation, are used to simplify models. Although there are many different model reduction procedures, the evaluation of reduced models is difficult and depends on the parameter values of the full model. There is a lack of a criteria for evaluating reduced models when the model parameters are uncertain.

Results: We developed a method to compare reduced models and select the model that results in similar dynamics and uncertainty as the original model. We simulated different parameter sets from the assumed parameter distributions. Then, we compared all reduced models for all parameter sets using cluster analysis. The clusters revealed which of the reduced models that were similar to the original model in dynamics and variability. This allowed us to select the smallest reduced model that best approximated the full model. Through examples we showed that when parameter uncertainty was large, the model should be reduced further and when parameter uncertainty was small, models should not be reduced much.

Conclusions: A method to compare different models under parameter uncertainty is developed. It can be applied to any model reduction method. We also showed that the amount of parameter uncertainty influences the choice of reduced models.

Keywords: Model reduction, Parameter uncertainty, Clustering, Systems biology

Background

Modelling of biochemical networks

Biochemical networks consist of chemical reactions between compounds, such as enzymes and metabolites. Through these reactions, the various compounds are consumed and produced. Each of these reactions has a reaction rate (flux) that typically depends on the compound concentrations, giving a dynamical behaviour of the system. The compound concentrations can thus be modelled by systems of ordinary differential equations (ODEs) and such dynamical models of biochemical networks may give biological insight that could not be obtained by modelling the compounds individually. However, the network dynamics may be complex and difficult to model accurately. The chemical reactions could possess advanced kinetics such as activation and inhibition. In addition, the dimensions of the network may be large, for example the central energy metabolism in E. *coli* consists of more than 50 metabolites and 100 reactions [1].

Model reduction

The potential high complexity of the ODEs in the model represents a major challenge in analysing the dynamics of the system. Model reduction is a method for studying biochemical networks as it aims to identify the main components governing the dynamics of the system. The reduced model should be simpler to analyse, but retain the dynamical behaviour of the original model. There are different approaches to reduce the complexity of biochemical reaction networks, with the most common ones being lumping, sensitivity analysis and time-scale analysis [2–4]. Lumping combines compounds with

*Correspondence: havard.froysa@uib.no
Department of Mathematics, University of Bergen, Mailbox 7803, 5020, Bergen, Norway

similar behaviour into pseudo-compounds and considers differential equations involving these lumped pseudo-compounds [5, 6]. By performing parameter sensitivity analysis, the parameters with the least effect on the system output are neglected [7, 8]. In time-scale separation, biological processes are split into fast and slow processes and then the focus is put on the relevant time scale [9–15].

Another challenge in the analysis of complex networks is the lack of information on the kinetic properties of the reactions and parameter values. Reduction approaches that are not influenced by parameter uncertainty or incompleteness are called parameter independent reduction methods. For example, some reduction techniques based on exact lumping methods [5, 6] or qualitative reduction methods [16, 17] are parameter independent. Such reduction methods have been used extensively for signalling networks. For most reduction techniques, including methods based on time-scale separation or sensitivity analysis, the full parametrization of the model is required. In parameter dependent reduction, model parameters can play a significant role in selecting the elements for reduction. For some biochemical networks, the accuracy and validity of the reduced model can be influenced by changing the range of parameters so that the reduced model is only valid locally [3]. For reaction networks with well separated parameter values, reduced models capture the dynamical behaviour of the original model with an acceptable level of accuracy for an extensive range of parameter values [11, 18]. This, however, is not the case for general networks.

While there is a large literature on model reduction techniques, there is a lack of methods for evaluating model reductions. Some ad-hoc methods are the difference or scaled difference between the full and reduced model [5, 9], an error integral [14] and a criterion based on the initial values [10]. We are not aware of any criteria for evaluation of model reductions that takes parameter uncertainty into account. We present a new way to evaluate model reductions that takes parameter uncertainty into account and show the benefit of this method on two example networks.

Methods

Mathematical framework

The state variables of the dynamical model are the concentrations of the compounds. These compounds occur in different combinations on the left and right hand side of the chemical reactions of the network, where such a combination is called a complex [14]. For example, the chemical reaction $X_1 + X_2 \rightarrow X_3$ consists of the compounds X_1, X_2 and X_3, and the complexes $X_1 + X_2$ and X_3. The complex on the left hand side of an equation being consumed is called the substrate complex of the reaction and the complex on the right hand side of the reaction being produced is called the product complex. All this information can be represented mathematically by a stoichiometric matrix [1] which gives the structure of the network.

In the notation of Rao et al. [14] the complexes are given by a matrix Z where the columns are the non-negative integer stoichiometric coefficients of the different complexes. The internal reactions are given by the linkage matrix B where each column corresponds to a reaction. This column is zero except in the rows corresponding to the substrate and product complex where it is -1 and 1, respectively. Let $x_i(t)$ be the concentration of compound i at time t and $\mathbf{x}(t)$ the corresponding vector quantity. The dynamics of any biochemical network is given by the system

$$\dot{\mathbf{x}} = ZB\mathbf{v} + Z\mathbf{v}_b \tag{1}$$

of ODEs where Z and B give the network structure as described above. The vector $\mathbf{v}$ provides the internal fluxes of the network and $\mathbf{v}_b$ the boundary fluxes, i.e. the fluxes entering or leaving the network. As the fluxes typically are functions of $\mathbf{x}$, we restrict the internal fluxes $\mathbf{v}$ to the form

$$v_j(\mathbf{x}) = k_j d_j(\mathbf{x}) \exp\left(Z_{\mathcal{S}_j}^T \mathrm{Ln}(\mathbf{x})\right) \tag{2}$$

considered in [14] where k_j is a kinetic proportionality constant of reaction j, $d_j(\mathbf{x})$ is any function of $\mathbf{x}$, $Z_{\mathcal{S}_j}$ is the column of Z corresponding to the substrate complex of reaction j and $\mathrm{Ln}(\mathbf{x})$ is the mapping defined by $(\mathrm{Ln}(\mathbf{x}))_i = \ln(x_i)$. Further, let $Z_{\mathcal{S}}$ be the matrix where column j is $Z_{\mathcal{S}_j}$, i.e. the substrate complex of the reaction.

The dynamical model (1) now has the parameters k_j in addition to potential parameters in $\mathbf{v}_b(\mathbf{x})$ and the functions $d_j(\mathbf{x})$. A given set of values for such a parametrization will be called a parameter set. The unreduced model described by (1) will be referred to as the full or original model.

Reduction

We use the reduction procedure of Rao et al. [14] to reduce the model for a given parameter set. The first step in this procedure is to specify a set $\mathcal{M}_1$ of compounds considered to be important in the view of experimental design, e.g. the ones that are possible to measure. Note that the choice of $\mathcal{M}_1$ is subjective, but plays a major role in the reduction as the dynamics of the compounds in $\mathcal{M}_1$ are the ones used to compare the different reduced models. Then, the complexes of the network are divided into two categories. The first category is the complexes containing at least one of the compounds in $\mathcal{M}_1$. These complexes will not be considered for reduction. The other category is the complexes not containing any of the compounds in $\mathcal{M}_1$, and these will be the complexes considered for reduction. The reduction is then based on the assumption that the model approaches some steady state that can be found by integrating the system for a long enough time and that

the model is asymptotically stable around the steady state. A complex is reduced by setting its concentration constant equal to the corresponding steady state value of the full model. This can be done simultaneously for any number of complexes.

Having the possibility to reduce any given set of complexes, an iterative method to choose the complexes to be reduced is presented in Rao et al. [14]. It is a greedy method that reduces one complex at the time, always choosing the one yielding the smallest error as defined below. Finally, it stops when an error threshold is reached.

However, since the reduced models are independent of the order of reduction, we consider all possible simultaneous reductions of complexes. Assume now that there are c complexes eligible for reduction. It is then possible to reduce anywhere from 0 to c complexes, where reducing 0 gives the full model. In total there are 2^c possible reduced models for a given original model and parameter set. For each of these models, the concentrations of the compounds in $\mathcal{M}_\mathrm{I}$ are then used to compare the models. When having n different parameter sets for the same original model, we perform the described reduction procedure for all the parameter sets. This yields 2^c possible reduced models for each parameter set and a total of $n \cdot 2^c$ different reduced models.

Comparing models

We need to be able to compare the dynamics of the different reduced models. In Rao et al. [14] the difference between the original model and a given reduced model is measured by an error integral. Let the concentration at time t of compound number i be $x_{ir}(t)$ and $x_{if}(t)$ for the reduced and the full model, respectively. Further, let $\mathbf{x}_r$ and $\mathbf{x}_f$ be the corresponding vector quantities for all the compounds. Finally, let $n\left(\mathcal{M}_\mathrm{I}\right)$ be the number of compounds in $\mathcal{M}_\mathrm{I}$ and $[0, T]$ the time interval that we evaluate the dynamics over. The error integral is then given by

$$I_T\left(\mathbf{x}_r, \mathbf{x}_f\right) = \sum_{i \in \mathcal{M}_\mathrm{I}} \frac{1}{Tn\left(\mathcal{M}_\mathrm{I}\right)} \int_0^T \left| 1 - \frac{x_{ir}(t)}{x_{if}(t)} \right| dt \qquad (3)$$

which gives the average relative difference between the full and reduced model for all the compounds in $\mathcal{M}_\mathrm{I}$ over the given time interval. Note that the error integral is non-symmetric in its arguments. However, we need to compare any two (reduced) models without favouring one of them. For this reason we introduce the symmetric error measure

$$E_T\left(\mathbf{x}_1, \mathbf{x}_2\right) = \frac{1}{2}\left(I_T\left(\mathbf{x}_1, \mathbf{x}_2\right) + I_T\left(\mathbf{x}_2, \mathbf{x}_1\right)\right) \qquad (4)$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the compound concentrations of any two (reduced) models. Note that this error measure can be calculated also for two models having different parameters as long as they have the same set $\mathcal{M}_\mathrm{I}$.

Clustering

We use single linkage clustering [19] with the symmetric error as dissimilarity measure to cluster all the $n \cdot 2^c$ models with different parameter sets and reductions. Single linkage clustering is an agglomerative clustering method, which means that initially every model is in its own cluster. The dissimilarity $d(C_1, C_2)$ between two clusters C_1 and C_2 is calculated as the minimal symmetric error $\min_{x \in C_1, y \in C_2} E_T(x, y)$. The two clusters with the lowest dissimilarity are combined into one cluster at a hight given by their dissimilarity. Clusters are iteratively combined until only one cluster remains. This stepwise process can be visualized in a dendrogram [20]. A dendrogram provides a complete description of the single linkage clustering. From such dendrograms it is apparent which models are most similar and which models are more different.

We then color the dendrogram according to the used reduction. Each reduction is mapped to a color and each leaf of the dendrogram receives the color associated to its reduction. Model reductions that cluster together with the original model do not change the model behaviour, while model reductions that are separated from the original model changed the model behaviour. So if the dendrogram separates colors, we consider the model reduction that causes the separation to change the model behaviour. The reduced models that are distributed in a similar way as the original model in the dendrogram are considered to be consistent for the given parameter uncertainty.

In order to analytically compare the distributions of different models in the dendrogram, we calculate the positions in the dendrogram for each model. We then use the test statistics of a Kolmogorov-Smirnov test [21] between a given model and the full model as score for the model. For a given threshold α, we say that models with a score lower than the threshold are consistent with the full model at threshold α. Finally, the best reduced model is then chosen to be the consistent model that uses the most reductions. In the case of several consistent models having the same number of reductions, the best model is the one with the lowest score. For the remainder of this article we use a threshold of $\alpha = 0.2$.

Simple example

To illustrate the method, we created a small example network consisting of four compounds as shown in Fig. 1. Each compound occurs only one place in the network and never in combination with other compounds, implying that the complexes are just the compounds. The set $\mathcal{M}_\mathrm{I}$ of important compounds is chosen to be number 1 and 4 such that the intermediate compounds 2 and 3 are considered for reduction.

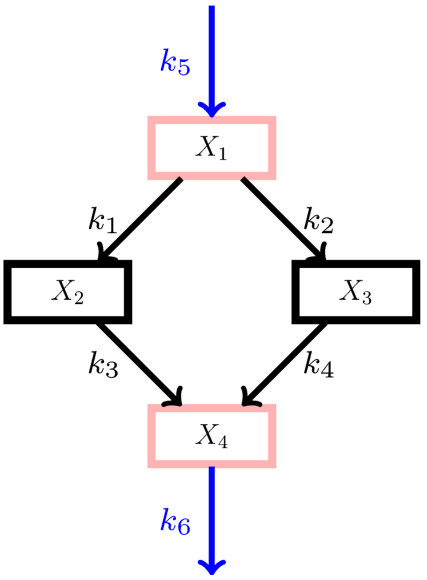


Fig. 1 Example network. Each node is a compound and each arrow a reaction. The k_j's are the kinetic parameters of the reactions. Important compounds $\mathcal{M}_I$ and candidate compounds for reduction are specified by pink and black rectangles, respectively. External fluxes are indicated by blue arrows

We apply mass action kinetics. Then k_j is the only kinetic parameter of reaction j. In the notation of [14] introduced earlier in the article, we have the matrices

$$Z = \begin{bmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1 \end{bmatrix} \quad B = \begin{bmatrix} -1&-1&0&0\\ 1&0&-1&0\\ 0&1&0&-1\\ 0&0&1&1 \end{bmatrix} \quad Z_{\mathcal{S}} = \begin{bmatrix} 1&1&0&0\\ 0&0&1&0\\ 0&0&0&1\\ 0&0&0&0 \end{bmatrix} \tag{5}$$

for the network. Using mass action we have $d_j(\mathbf{x}) = 1$ such that (2) becomes

$$v_j(\mathbf{x}) = k_j \exp\left(Z_{\mathcal{S}_j}^T \mathrm{Ln}(\mathbf{x})\right), \ j \in \{1,2,3,4\} \tag{6}$$

for the internal fluxes of $\mathbf{v}$. The boundary fluxes are given by

$$\mathbf{v}_b = \begin{bmatrix} k_5 & 0 & 0 & -k_6 x_4 \end{bmatrix}^T \tag{7}$$

where the last entry is negative since the flux is leaving the network.

The dynamics are now given by (1) and we have six kinetic parameters k_j associated with one of the six fluxes each. We sampled several parameter sets, which as expected lead to different reduction results. The parameter set that was chosen as reference because it gives particularly interesting reduction results is shown in Table 1. Then, 100 new parameter sets were sampled using this reference set by assuming the parameters to be independently log-normally distributed with the logarithm of the reference values as mean on the log scale and 0.1 as log standard deviation. We applied the reference initial values for all of the parameter sets, and the models were then reduced and clustered as described above.

Yeast glycolysis example

We also tested our method on a kinetic model of yeast glycolysis [22] shown in Fig. 2. This model was used in Rao et al. [14] to demonstrate the model reduction method which ignores parameter uncertainty. The model is asymptotically stable around the steady state and the governing equations of the system can be represented in the form of Eqs. 1 and (2) such that the reduction procedure can be applied. The important compounds to form $\mathcal{M}_I$ are Glci, TRIO, BPG, PYR, AcAld and NADH. Accordingly, the six candidates for reduction are F6P, G6P, P2G, P3G, PEP and F16BP, which leads to a total of $2^6 = 64$ possible reductions for a given parameter set including the full model.

The model has 89 parameters for the different reactions of the network. Each of these parameters should be non-negative, and have a reference value used in [14]. To study the effect of parameter uncertainty on the reduction we sampled parameter sets using these reference values. We assumed the parameters to be independently log-normally distributed with mean equal to the reference value and standard deviation equal to the reference value divided by a scaling parameter. The parameters with reference value zero were set to zero in the sampling. We sampled 100 parameter sets for each of the values 3, 5, 10, 20, 50 and 100 of the scaling parameter. For each of the parameter sets we performed model reduction and clustered all the $100 \cdot 64 = 6,400$ resulting models for each scaling parameter as described above. We ended up with six dendrograms containing 6400 models each.

Table 1 Initial values and reference kinetic parameter values for the example network of Fig. 1

Parameter	Value	Initial value	Value
k_1	0.44	$x_1(0)$	0.4
k_2	0.03	$x_2(0)$	0.0
k_3	0.55	$x_3(0)$	0.5
k_4	0.44	$x_4(0)$	0.4
k_5	0.42		
k_6	0.33		

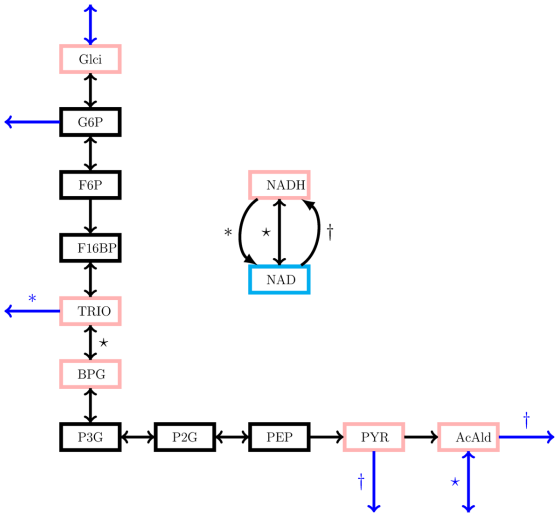


Fig. 2 Yeast glycolysis network. Each node is a compound and each arrow a reaction. Important compounds $\mathcal{M}_I$ and candidate compounds for reduction are specified by pink and black rectangles, respectively. NAD indicated by a cyan rectangle is not explicitly included in the model as the total amount of NAD and NADH is conserved. External fluxes are indicated by blue arrows. $*$, $\star$ and † show an irreversible reaction from NADH to NAD, a reversible reaction between NADH and NAD and an irreversible reaction from NAD to NADH, respectively. As indicated in the network, different types of these reactions bind to some of the fluxes

In order to check the sensitivity of the method to the number of parameter sets sampled, we also sampled 1000 parameter sets for the model with scaling parameter 50. For each parameter set we considered all model reductions with a Kolmogorov-Smirnov test score below a threshold of 0.5 for the 100 previous parameter sets. We performed model reduction and clustering as above.

All analyses were performed in MATLAB [23]. All code used to generate the results is available in the online supplementary material.

Results

Simple example

For the used parameter values, the model with both compounds number 2 and 3 reduced clustered together with the original model and had a Kolmogorov-Smirnov score of 0.17. Both the model with only compound 2 removed and the model with only compound 3 removed had a Kolmogorov-Smirnov score of 1.00. The models with only compound 3 reduced were the furthest from the cluster including the original model. Figure 3 shows the single linkage cluster dendrogram. The behaviour changes substantially for different parameter values and parameter uncertainties.

Yeast glycolysis example

The trajectories of the full model and all reduced models using the parameter set from [14] show no effect for Glci, two groups for TRIO, PYR and NADH, but no clear picture for BPG and ACALD (Fig. 4). For the reference parameter set, we found two big clusters. The first cluster contained the full models as well as all the models with compound F16BP not reduced, and the second cluster contained all models with F16BP reduced.

The clusterings for a distribution of parameters depended on the parameter distribution. When the standard deviation was high, there were no clear clusters and the full models were evenly distributed between the reduced models (Fig. 5, top left). This means that the uncertainty in the parameters had more effect than the model uncertainty due to reduction. The more certain the parameters were, the more we saw a clear picture emerge, with all models that had compound F16BP reduced clustering together and all other models forming a separate cluster (Fig. 5, top right, bottom left). When decreasing parameter uncertainty even further, the original models started forming a cluster of models where both compounds PEP and F16BP were not reduced (Fig. 5, bottom right).

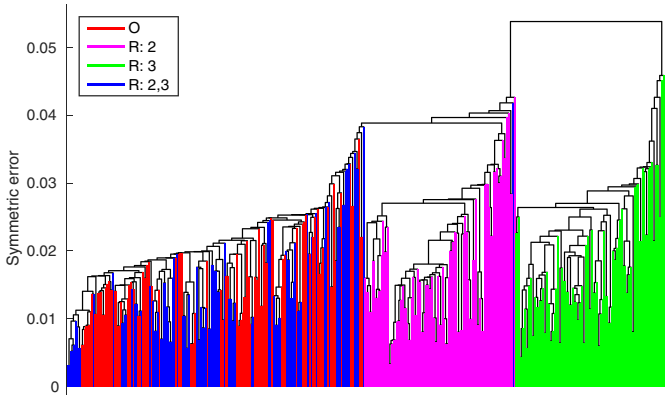


Fig. 3 Model clustering. Dendrogram from single linkage clustering of all model reductions with 100 parameters sets. Parameters were sampled from a log-normal distribution with log standard deviation 0.1

In addition to finding clusters that are inconsistent with the model uncertainty, we studied the distribution of the reduced models in the dendrogram. In the case of large parameter uncertainty (scaling parameters 3, 5, 10) the distribution of the fully reduced model in the dendrogram was similar to the distribution of the original model (Kolmogorov-Smirnov 0.11 or smaller). In the case of relatively large uncertainty (scaling parameter 20), all the models that did not reduce F16BP were distributed similarly to the original model (Kolmogorov-Smirnov 0.01).

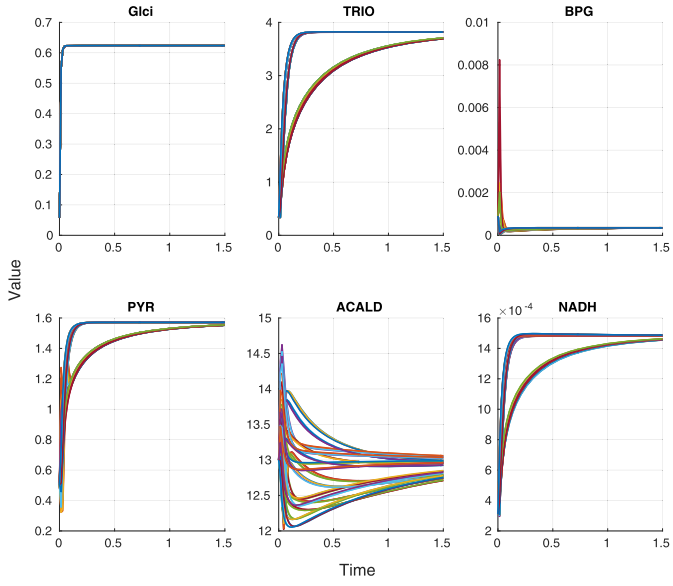


Fig. 4 Yeast trajectories. The $\mathcal{M}_1$ states are shown for the reference parameters. Each color corresponds to a different model reduction

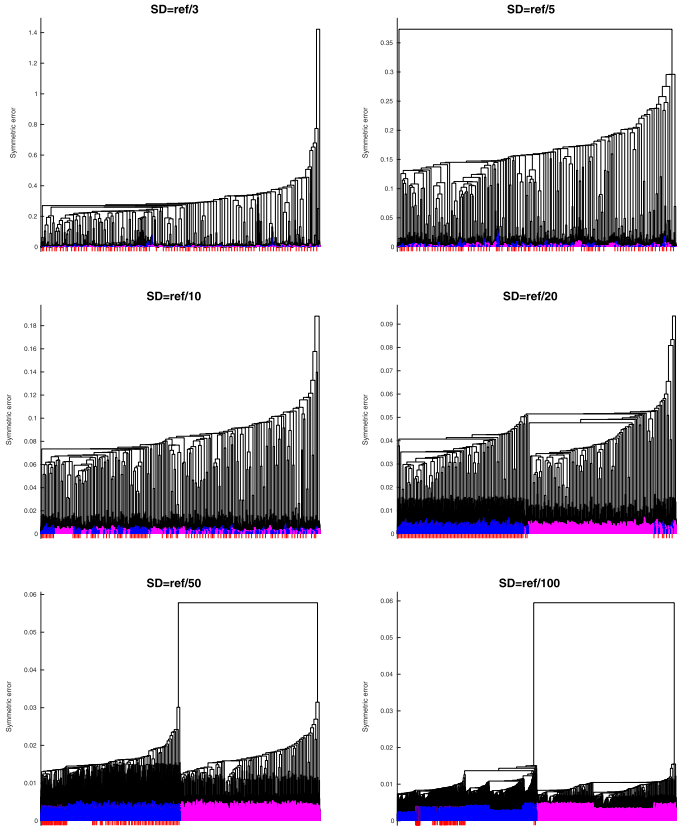


Fig. 5 Model clustering of F16BP reduced models. Dendrogram from single linkage clustering of all the model reductions using 100 parameters sets. Parameters were sampled from a log-normal distribution with standard deviation as reference value divided by 3 (top left), 5 (top right), 10 (center left), 20 (center right), 50 (bottom left) and 100 (bottom right). The original models are shown in red, models where F16BP was reduced are purple and all other models are blue

When the uncertainty was relatively low (scaling parameter 50), all models with F16BP and PEP not reduced clustered together with the full model (Kolmogorov-Smirnov 0.01 or 0.02). However, in the case of very low uncertainty (scaling parameter 100) the only model whose distribution in the dendrogram was similar to the distribution of the original model was the one where only F6P was reduced (Kolmogorov-Smirnov 0.01). The sensitivity analysis showed that whether or not a reduction was consistent for a given uncertainty did not dependent on the number of parameter sets (Fig. 6).

Discussion

We developed a new method to evaluate model reductions under parameter uncertainty based on the symmetric error measure in (4). In the yeast glycolysis example we showed that the amount of parameter uncertainty influences the model reduction. In particular, model uncertainty and parameter uncertainty are positively related. When the model parameters are uncertain, the model can be reduced further without increasing uncertainty in the model dynamics. We have also demonstrated empirically that if a model can be reduced to a certain degree for a

Reduction						Scaling						
G6P	F6P	F16P	P3G	P2G	PEP	3	5	10	20	50	50L	100
						0.01	0.01	0.01	0.01	0.02	0.06	0.50
						0.01	0.01	0.01	0.01	0.01	< 0.01	**0.01**
						0.10	0.08	0.11	0.92	1.00	–	1.00
						0.01	0.01	0.01	0.01	0.02	0.03	0.46
						0.01	0.01	0.01	0.01	0.02	0.03	0.46
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.01	0.01	0.01	0.01	0.02	0.06	0.50
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.01	0.01	0.01	0.01	0.02	0.07	0.54
						0.01	0.01	0.01	0.01	0.02	0.07	0.54
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.01	0.01	0.01	0.01	0.02	0.03	0.46
						0.01	0.01	0.01	0.01	0.02	0.03	0.46
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.10	0.08	0.11	0.92	1.00	–	1.00
						0.10	0.08	0.11	0.92	1.00	–	1.00
						0.10	0.08	0.11	0.92	1.00	–	1.00
						0.01	0.01	0.01	0.01	0.02	0.03	0.46
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.01	0.01	0.01	0.01	0.02	0.07	0.54
						0.01	0.01	0.01	0.01	0.02	0.07	0.54
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.01	0.01	0.01	0.01	0.02	0.07	0.54
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.01	0.01	0.01	0.01	0.02	0.03	0.46
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.10	0.08	0.11	0.92	1.00	–	1.00
						0.10	0.08	0.11	0.92	1.00	–	1.00
						0.10	0.08	0.11	0.92	1.00	–	1.00
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.01	0.01	0.01	**0.02**	**0.07**	0.54
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.01	0.01	0.01	0.01	0.38	0.86	0.98
						0.10	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.01	0.01	0.01	**0.01**	0.38	0.86	0.98
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	0.08	0.11	0.92	1.00	–	1.00
						0.09	**0.08**	**0.11**	0.92	1.00	–	1.00

Fig. 6 Kolmogorov-Smirnov test scores. Kolmogorov-Smirnov test scores for all the model reductions using 100 parameters sets as well as the sensitivity analysis with 1000 parameter sets (50L). The compounds in gray are reduced in the model of the corresponding row. Parameters were sampled from a log-normal distribution with standard deviation as reference value divided by the scaling factor. Models that are consistent with the original model are shown in light green and the best reduced model for each case is shown in dark green

given amount of uncertainty, then it can be reduced to at least the same degree if the uncertainty increases. If a model is used to analyse different scenarios, the parameters for all the scenarios should be considered when reducing a model. A full model should only be reduced to a model that is consistent for all considered scenarios. In addition to parameter values, uncertainty in initial values should also be considered. Our analysis shows that the

reduction of Rao et al. [14] for the yeast model agrees with our best reduction for a relatively high amount of uncertainty, but becomes inappropriate for low or very large uncertainty.

In the simple example we demonstrated that it is sometimes better to reduce two complexes than just one. This also shows that even without parameter uncertainty the iterative approach used in [14] may not find the best reduction. Whether or not the best reduction is found depends on the symmetric error cut-off value. In the example, the reduced model would be found with symmetric error cut-off value at least 0.04, even though the symmetric error is only 0.02. The reference values in Table 1 for the parameters were chosen to illustrate this behaviour.

The novelty of our approach is a new way to evaluate model reduction. This model reduction evaluation criterion can be applied together with any model reduction method. Our criterion does not assume that the full model with a given parameter set is optimal. Instead it compares the full model with a wide range of parameter values to reduced models with the same range of parameter values to find a reduced model with the same properties, including model uncertainty. A reduced model with lower uncertainty in the trajectories could lead to overconfidence in the results.

A limitation of our method is that we need to choose a set $\mathcal{M}_{\mathrm{I}}$ of important compounds. This choice is subjective and affects the resulting reduced model. However, there are some natural choices for the set $\mathcal{M}_{\mathrm{I}}$, which depend on the model purpose. Of course $\mathcal{M}_{\mathrm{I}}$ should contain all the compounds the study is investigating. It should also contain all the compounds whose concentrations are measured experimentally. Another limitation of our approach is that we have to choose the length T of the time series. It is important that at time T the trajectories are close to the steady state, because otherwise the error integral does not cover the entire model dynamics. On the other hand T should not be too large because otherwise the error integral reduces to the difference in steady states. If the model does not approach a steady state the dissimilarity measure we use may not be appropriate. There may also be some scaling issues with our proposed approach. Already in the case where we have to evaluate 64 models, we have to calculate a 6400×6400 matrix of dissimilarity measures using 100 parameter sets. For most practical examples, however, it is possible to reduce the sample space of reductions to a manageable size. In our sensitivity analysis with 1000 parameter sets, we have solved the issue by using the first 100 parameter sets to exclude some model reductions, which lead to a $32,000 \times 32,000$ dissimilarity matrix. The calculation of this matrix is the computational bottleneck of the method, but parallel computing can be applied. Moreover, it is possible to iteratively compare only a few models at a time. We suggest that investigators adapt their strategies for model reduction based on model size, complexity and choice of the set $\mathcal{M}_{\mathrm{I}}$. The Kolmogorov-Smirnov score leads to an automatic way of choosing the best reduced model. However, we believe that it is important to look at the dendrograms and not choose the model reduction only based on the Kolmogorov-Smirnov scores.

Conclusions

We presented a new method for evaluating models under parameter uncertainty and applied it for comparing full models to reduced models. We showed that multiple reductions can result in better models than individual reductions and that the amount of parameter uncertainty influences the choice of reduced models.

Additional file

Additional file 1: Scripts. Archive file containing all the scripts needed to produce and analyse the data of this paper. These scripts also produce Figs. 3, 4, and 5. (ZIP 23.3 kb)

Abbreviations

ODE: Ordinary differential equation

Acknowledgements

This work is a contribution to the dCod 1.0 project. https://digitallifenorway.org/gb/projects/dcod-1-0

We thank professors Guttorm Alendal, Jarle Berntsen and Hans J. Skaug for reading through the manuscript and providing valuable comments.

Funding

This research was supported by the Research Council of Norway through grant 248840, dCod 1.0. The funding body did not play any role in the design of the study and in the writing of the manuscript.

Availability of data and materials

All data used in this research can be obtained by running the scripts provided in Additional files. These scripts also produce Figs. 3, 4 and 5.

Authors' contributions

All authors designed the method, wrote and revised the manuscript, and contributed to the interpretation of the results and the final version of the manuscript. HGF and SF performed the model simulations and model reductions. NB and HGF performed the cluster analysis. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 13 September 2017 Accepted: 9 July 2018
Published online: 27 July 2018

References

1. Palsson BØ. Systems Biology: Constraint-based Reconstruction and Analysis. Cambridge, United Kingdom: Cambridge University Press; 2015.
2. Okino MS, Mavrovouniotis ML. Simplification of mathematical models of chemical reaction systems. Chem Rev. 1998;98(2):391–408.
3. Radulescu O, Gorban A, Zinovyev A, Noel V. Reduction of dynamical biochemical reactions networks in computational biology. Front Genet. 2012;3:131.
4. Snowden TJ, van der Graaf PH, Tindall MJ. Methods of model reduction for large-scale biological systems: A survey of current methods and trends. Bull Math Biol. 2017;79:1–38.
5. Borisov NM, Markevich NI, Hoek JB, Kholodenko BN. Signaling through receptors and scaffolds: independent interactions reduce combinatorial complexity. Biophys J. 2005;89(2):951–66.
6. Conzelmann H, Saez-Rodriguez J, Sauter T, Kholodenko BN, Gilles ED. A domain-oriented approach to the reduction of combinatorial complexity in signal transduction networks. BMC Bioinformatics. 2006;7(1):34.
7. Liu G, Swihart MT, Neelamegham S. Sensitivity, principal component and flux analysis applied to signal transduction: the case of epidermal growth factor mediated signaling. Bioinformatics. 2004;21(7):1194–202.
8. Maurya M, Bornheimer S, Venkatasubramanian V, Subramaniam S. Reduced-order modelling of biochemical networks: application to the gtpase-cycle signalling module. IEE Proc Syst Biol. 2005;152(4):229–42.
9. Maas U, Pope SB. Simplifying chemical kinetics: intrinsic low-dimensional manifolds in composition space. Combust Flame. 1992;88(3):239–64.
10. Surovtsova I, Sahle S, Pahle J, Kummer U. Approaches to complexity reduction in a systems biology research environment (sycamore). In: Proceedings of the 38th Conference on Winter Simulation. Monterey: Winter Simulation Conference; 2006. p. 1683–1689.
11. Radulescu O, Gorban A, Zinovyev A, Lilienbaum A. Robust simplifications of multiscale biochemical networks. BMC Syst Biol. 2008;2(1):86.
12. Hardin HM. Handling biological complexity: as simple as possible but not simpler. 2010.
13. Kourdis PD, Palasantza AG, Goussis DA. Algorithmic asymptotic analysis of the nf-κb signaling system. Comput Math Appl. 2013;65(10):1516–34.
14. Rao S, Van der Schaft A, Van Eunen K, Bakker BM, Jayawardhana B. A model reduction method for biochemical reaction networks. BMC Syst Biol. 2014;8(1):52.
15. West S, Bridge LJ, White MR, Paszek P, Biktashev VN. A method of 'speed coefficients' for biochemical model reduction applied to the NF-κB system. J Math Biol. 2015;70(3):591–620.
16. Naldi A, Remy E, Thieffry D, Chaouiya C. A reduction of logical regulatory graphs preserving essential dynamical properties. In: Computational Methods in Systems Biology. Bologna: Springer; 2009. p. 266–280.
17. Naldi A, Remy E, Thieffry D, Chaouiya C. Dynamically consistent reduction of logical regulatory graphs. Theor Comput Sci. 2011;412(21):2207–18.
18. Gorban A, Radulescu O. Dynamic and static limitation in multiscale reaction networks, revisited. Adv Chem Eng. 2008;34:103–73.
19. Florek K, Łukaszewicz J, Perkal J, Steinhaus H, Zubrzycki S. Sur la liaison et la division des points d'un ensemble fini. Colloq Math. 1951;2: 282–2851952.
20. Houlahan P, Scalo J. Recognition and characterization of hierarchical interstellar structure. ii - structure tree statistics. Astrophys J. 1992;393: 172–87.
21. Conover WJ. Practical Nonparametric Statistics. New York: Wiley; 1971.
22. van Eunen K, Kiewiet JA, Westerhoff HV, Bakker BM. Testing biochemistry revisited: how in vivo metabolism can be understood from in vitro enzyme kinetics. PLoS Comput Biol. 2012;8(4):1002483.
23. The Mathworks, Inc. MATLAB Version 8.6.0.267246 (R2015b). Natick, Massachusetts: The Mathworks, Inc; 2015.

Paper C

A new conceptual toxicokinetic model to assess synergistic mixture effects between the aromatic hydrocarbon β-naphthoflavone and the azole nocodazole on the CYP1A biomarker in a fish cell-line

Shirin Fallahi, Marie Mlnaíková, Charlotte Alvord, Guttorm Alendal, Håvard G. Frøysa, Torbjörn Lundh and Malin C. Celander
Revision submitted to: Environmental Science & Technology

A new conceptual toxicokinetic model to assess synergistic mixture effects between the aromatic hydrocarbon β-naphthoflavone and the azole nocodazole on the CYP1A biomarker in a fish cell-line

Shirin Fallahi [1], Marie Mlnaříková [2, §], Charlotte Alvord [2], Guttorm Alendal [1], Håvard G. Frøysa [1], Torbjörn Lundh [3] and Malin C. Celander [2, *]

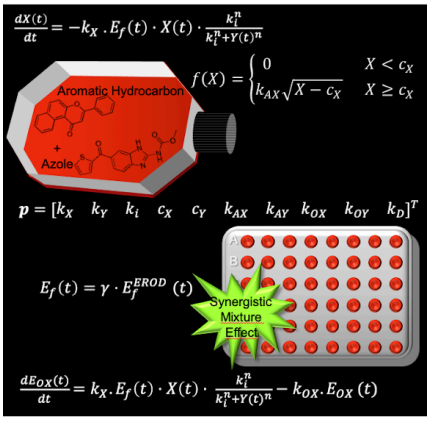
[1] Department of Mathematics, University of Bergen, P.O. Box 7803, N 5020 Bergen, Norway

[2] Department of Biological and Environmental Sciences, University of Gothenburg, P.O. 463, SE 405 30 Gothenburg, Sweden

[3] Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, SE 412 96 Gothenburg, Sweden

ABSTRACT

Toxicokinetic interactions with catabolic cytochrome P450 (CYP) enzymes can inhibit chemical elimination pathways and cause synergistic mixture effects. We have created a mathematical bottom-up-model for a synergistic mixture effect where we fit a multi-dimensional function to a given dataset using an auxiliary non-additive approach. The toxicokinetic model is based on data from a previous study in a fish cell-line, where the CYP1A enzyme activity was measured over time after exposure to various combinations of the aromatic hydrocarbon β-naphthoflavone and the azole nocodazole. To describe the toxicokinetic mechanisms in this pathway and how that affects the CYP1A biomarker, the model is using ordinary differential equations. Local sensitivity and identifiability analyses revealed that all the ten parameters estimated in the model were identified uniquely while fitting the model to the data for the CYP1A enzyme activity. The model has a good prediction power and is a promising tool to test the synergistic toxicokinetic interactions between different chemicals.



1. INTRODUCTION

Induction of cytochrome P450 1A (CYP1A) in fish can be used as a biomarker to assess exposure to aromatic hydrocarbons in the aquatic environment. Induction of CYP1A is typically measured as increased levels of transcription (*i.e.* CYP1A mRNA/CYP1A protein levels), increased CYP1A enzyme activities or a combination of both. Aromatic hydrocarbons activate the aryl hydrocarbon receptor (AhR), which results in an induction of the CYP1A synthesis. The AhR-CYP1A signaling is central in the chemical detoxification pathway in fish.[1] The aquatic environment is also contaminated with other anthropogenic chemicals, including pharmaceuticals.[2,3] Some pharmaceuticals can interfere with the AhR-CYP1A signaling pathway. Hence, fish populations in their natural habitats are exposed to mixtures of chemicals that can interact with the AhR-CYP1A signaling and thereby affect the CYP1A biomarker.[4]

The mixture effect of different chemicals is called an additive effect if there is no direct interaction between the chemicals when they exert their effects. Thus, the chemicals act independently of each other with similar modes-of-action (MoA) and the mixture effect can be explained by addition. If a mixture gives an effect higher than the additive prediction, this effect is called synergistic.[5] There are established models to assess additive mixture effects such as the concentration addition, the independent action and the generalized concentration addition models.[6] There are also models that can identify non-additive mixture effects from response patterns for endpoints.[7-10] These models are, however, empirical models and lack a mechanistic basis for prediction. For this reason, there is a need for models to assess non-additive mixture effects in a mechanistic manner, e.g. to describe synergistic effects in fish and other vertebrates.

Chemicals with the same or different MoA can interact with each other's detoxification mechanisms and cause adverse toxicokinetic interactions. The value of integrating toxicokinetics, for a better mechanistic understanding to predict interactions between different chemicals in mixtures, has been advocated by the European Commission.[11] A promising approach, using a mechanistic toxicokinetic and toxicodynamic model, was suggested to

describe synergistic mixture effects between azole fungicides and a pyrethroid insecticide in the invertebrate *Daphnia magna*. This model was based on the fact that the synergistic potential of adding azoles could be explained by the azoles occupying the CYP enzymes which reduces the biotransformation of the insecticide.[12]

Azoles (*e.g.* clotrimazole, ketoconazole, nocodazole (NOC), omeprazole, prochloraz and propiconazole) have been shown to interact with the CYP system in fish and fish cells.[13-21] Most azoles act as potent inhibitors of CYP1A enzyme activities, but weak inducers of CYP1A transcription in fish.[17-19] Previous studies in the *Poeciliopsis lucida* hepatocellular carcinoma (PLHC-1) cell line show that the benzimidazole and microtubule disassembling drug NOC alone acted as a potent inhibitor of the CYP1A enzyme activity, and a weak inducer of CYP1A expression. Compared to BNF, NOC is one order of magnitude less potent and about 50 times less efficient to induce CYP1A expression.[20,21] However, a synergistic mixture effect with NOC and the prototypical AhR agonist, β-naphthoflavone (BNF), was seen as increased induction of CYP1A.[21] Although BNF and NOC have different MoA, they seem to share a common elimination pathway. Thus, PLHC-1 cells exposed to NOC were more sensitive to BNF exposures compared to cells exposed to BNF alone.[21] The synergistic mixture effect is based on the hypothesis that NOC occupies the CYP1A enzymes, inhibiting the CYP1A dependent metabolism of BNF which in turn enhances the BNF-mediated AhR-CYP1A signaling (Figure 1).

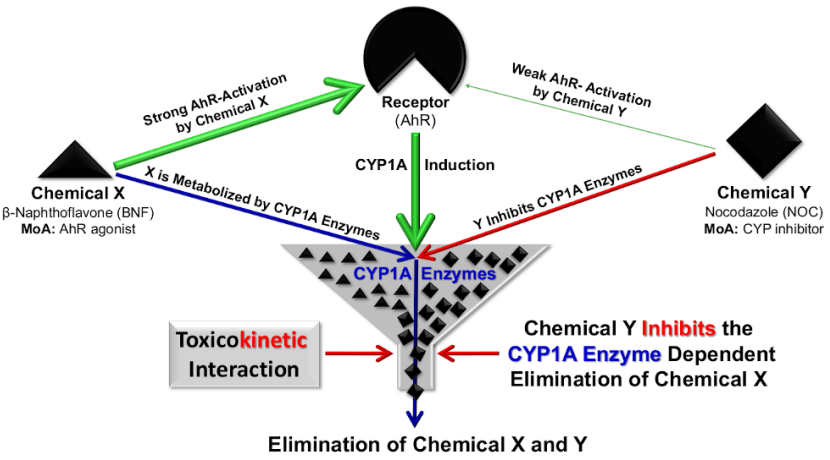


Figure 1. Illustration of the pathway that the model describes. Chemical X (β-naphthoflavone (BNF) black triangles) and chemical Y (nocodazole (NOC) black squares) belong to different chemical classes, each with different modes-of-action (MoA). Induction of cytochrome P450 1A (CYP1A) is mediated via ligand activation of the aryl hydrocarbon receptor (AhR). Chemical X is a strong ligand to the AhR (illustrated by a thick green arrow), which results in strong induction of CYP1A. Chemical Y is a weak ligand to the AhR (illustrated by a thin green arrow), which results in low induction of CYP1A. Both X and Y bind to the CYP1A enzymes, where X is metabolized by CYP1A enzymes (blue arrow) and Y inhibits CYP1A enzymes (red arrow). Thus, chemical Y occupies the CYP1A enzymes in the elimination pathway (the gray funnel shape), delaying biotransformation of chemical X which in turn results in increased biological half-life of chemical X. This toxicokinetic interaction between X and Y results in a synergistic mixture effect. The conceptual model, presented in this study, describes the toxicokinetic interaction between chemical X and Y in this pathway.

The aim of this study was to create a new conceptual toxicokinetic model to describe the toxicokinetic interaction between BNF and NOC, where the dominant effect of NOC is direct inhibition of CYP1A enzymes and the dominant effect of BNF is induction of CYP1A transcription via activation of AhR. This is a first attempt to create a mathematical bottom-up-model for synergistic mixture effects where we fit a multi-dimensional function to a given experimental dataset from a previous study using an auxiliary non-additive model.[21,22] The time dynamics is a key factor in the toxicokinetic interactions. We hypothesized that by constructing a model using ordinary differential equations (ODEs), we can describe how the concentrations of the chemicals change over time and their resulting effect on the CYP1A biomarker.

2. METHODS

2.1 Development of the mathematical model. To describe the mixture effect on the CYP1A biomarker, we construct a model using a set of ODEs. The model also explains the individual effects of each chemical by setting the initial concentrations of the other compounds equal to zero. There are five concentration state variables in the equations, where X is unbound BNF, Y is unbound NOC, E_f is free CYP1A enzymes, E_{OX} is CYP1A enzymes occupied by BNF and E_{OY} is CYP1A enzymes occupied by NOC. The state variables, their initial values and the parameters in the model (*i.e.* constants) are listed with units in Table 1.

The rate by which the unbound BNF (X) is biotransformed is modeled as:

$$\frac{dX(t)}{dt} = -k_X \,.\, E_f(t) \cdot X(t) \cdot \frac{k_i^n}{k_i^n + Y(t)^n} \quad (1)$$

In this equation the change in numbers of unbound BNF (X) molecules over time is described as a function of the turnover number of the CYP1A enzyme for biotransformation of BNF (k_X) together with the number of molecules for free CYP1A enzymes (E_f), BNF (X) and NOC (Y). The time unit (t) is hours. A Hill function is used to describe that BNF and NOC molecules compete for binding to the free CYP1A enzymes. Consequently, the change in numbers of unbound BNF molecules is affected by the number of NOC molecules over time, which creates a delay in the BNF-elimination pathway. The parameters k_i and n in the Hill function describe the competition between the BNF and NOC molecules for the free CYP1A enzymes. The parameter k_i is the concentration of NOC occupying half of the binding sites of the CYP1A enzymes.

The change in numbers of unbound NOC (Y) molecules over time is described by second-order kinetics using the turnover number of the CYP1A enzyme for biotransformation of NOC (k_Y), number of molecules for free CYP1A enzymes (E_f) and NOC (Y) giving:

$$\frac{dY(t)}{dt} = -k_Y \,.E_f(t) \cdot Y(t) \quad (2)$$

The change in numbers of free CYP1A enzymes (E_f) over time depends on the numbers of BNF (X) and NOC (Y) molecules that are occupying the CYP1A enzymes. Hence, binding of BNF and NOC molecules to the CYP1A enzymes result in increased numbers of occupied CYP1A enzymes (E_{OX} and E_{OY}) and decreased numbers of free CYP1A enzymes with the degradation rate constant k_D. The rate of change of free CYP1A enzymes is modeled as:

$$\frac{dE_f(t)}{dt} = -k_X.E_f(t) \cdot X(t) \cdot \frac{k_i^n}{k_i^n + Y(t)^n} - k_Y.E_f(t) \cdot Y(t) + f(X) + g(Y) +$$

$$k_{OX}.E_{OX}(t) + k_{OY}.E_{OY}(t) - k_D.E_f(t) \quad (3)$$

The parameters k_{OX} and k_{OY} are the biotransformation rate constants of BNF and NOC molecules whereas the functions $f(X)$ and $g(Y)$ describe the activation of the AhR-CYP1A signaling by BNF and NOC, respectively.

The activation of AhR is controlled by the number of BNF (X) and NOC (Y) molecules in the cells, in particular the number of BNF molecules. This is because BNF is more than 50 times more effective and around ten times more potent compared to NOC in activating the AhR-CYP1A signaling.[20,22] The functions $f(X)$ and $g(Y)$ are therefore included to describe this dependency for the activation of AhR, as described in equations (4) and (5):

$$f(X) = \begin{cases} 0 & X < c_X \\ k_{AX}\sqrt{X - c_X} & X \geq c_X \end{cases} \quad (4)$$

$$g(Y) = \begin{cases} 0 & Y < c_Y \\ k_{AY}(Y - c_Y) & Y \geq c_Y \end{cases} \quad (5)$$

The values of c_X and c_Y are the respective threshold-concentrations of BNF and NOC required to activate AhR-CYP1A. Equivalently, the parameters k_{AX} and k_{AY} describe the rates of BNF-induced and NOC-induced CYP1A enzymes, respectively.

The rate of change for occupied CYP1A enzymes by BNF (E_{OX}) and NOC (E_{OY}) are modeled as:

$$\frac{dE_{OX}(t)}{dt} = k_X . E_f(t) \cdot X(t) \cdot \frac{k_i^n}{k_i^n + Y(t)^n} - k_{OX} . E_{OX}(t) \quad (6)$$

$$\frac{dE_{OY}(t)}{dt} = k_Y . E_f(t) \cdot Y(t) - k_{OY} . E_{OY}(t) \quad (7)$$

The BNF and NOC molecules that occupy CYP1A enzymes are being biotransformed by the CYP1A enzymes. Next, their metabolites are released from the CYP1A enzymes and the previously occupied CYP1A enzymes (E_{OX} and E_{OY}) become free. The biotransformation rate constants are k_{OX} and k_{OY}. The number of free CYP1A enzymes (E_f) consequently increases and are available for the next cycle of biotransformation. The CYP1A biotransformation reduces the numbers of BNF and NOC molecules, and when there are too few molecules to activate AhR, no more free CYP1A enzymes are being synthesized. The remaining CYP1A will be degraded and the numbers of E_f will decrease.

Table 1. Model parameters and their descriptions.[a]

Parameter	Description	Unit	Value

State			
Variables			
X	concentrations of unbound BNF molecules	µM	
Y	concentrations of unbound NOC molecules	µM	
E_f	concentrations of free CYP1A enzymes (not used in model equations)	µM	
E_f^{EROD}	EROD activity of free CYP1A enzymes	pmol · (min · mg)$^{-1}$	
E_{OX}	concentrations of CYP1A enzymes occupied by BNF	µM	
E_{OY}	concentrations of CYP1A enzymes occupied by NOC	µM	
t	time	h	
Initial values			
	initial concentration of BNF	µM	0.1 and 1
	initial concentration of NOC	µM	1, 10 and 25
	initial EROD activity of free CYP1A enzymes	pmol · (min · mg)$^{-1}$	0
	initial concentration of CYP1A enzymes occupied by BNF	µM	0
	initial concentration of CYP1A enzymes occupied by NOC	µM	0
Constants			

Fixed values			
γ	Proportionality constant between EROD activity and concentration of free CYP1A	min · mg · L^{-1}	1
n	Hill coefficient	-	4
Estimated values			
k_X	turnover number of the CYP1A enzyme for BNF	(µM · h)$^{-1}$	0.033 ± 0.010
k_Y	turnover number of the CYP1A enzyme for NOC	(µM · h)$^{-1}$	0.039 ± 0.009
k_i	IC_{50} for NOC on EROD activity	µM	1.37 ± 0.049
c_X	minimum concentration of BNF to induce EROD activity	µM	0.063 ± 0.008
c_Y	minimum concentration of NOC to induce EROD activity	µM	0.542± 0.061
k_{AX}	rate of number of CYP1A enzymes induced by BNF	h^{-1}. (µM)$^{1/2}$	1.339± 0.060
k_{AY}	rate of number of CYP1A enzymes induced by NOC	h^{-1}	0.252 ± 0.050
k_{OX}	rate of biotransformation of BNF molecules	h^{-1}	0.375± 0.059
k_{OY}	rate of biotransformation of NOC molecules	h^{-1}	0.060 ± 0.012

k_D	rate of CYP1A enzyme degradation	h^{-1}	0.043 ± 0.003

[a] The state variables are defined by the equations (4), (5) and (9)-(13) in the text. The initial values of the state variables are the different doses added to the cell cultures. The constants are the model parameters, which are either fixed or estimated. The estimated parameters are given as the parameter estimate ± standard error. The parameters in the model are fitted to experimental EROD data from a previous study where the cells were exposed to different mixtures of BNF and NOC in a time-course study (Table S1).[21,22]

The measured data for the free CYP1A enzymes is the diagnostic ethoxyresorufin-*O*-deethylase (EROD) activity that is assumed to be proportional to the concentration of free CYP1A enzymes. This assumption is justified by the fact that only free CYP1A enzymes can carry out the EROD activity. For this reason, we express the concentration of the free CYP1A enzymes as:

$$E_f(t) = \gamma \cdot E_f^{EROD}(t) \tag{8}$$

The parameter γ is a proportionality constant and $E_f^{EROD}(t)$ is the EROD activity of the free enzymes that can be measured. The model equations then become:

$$\frac{dX(t)}{dt} = -k_X \cdot \gamma \cdot E_f^{EROD}(t) \cdot X(t). \frac{k_i^n}{k_i^n + Y(t)^n} \tag{9}$$

$$\frac{dY(t)}{dt} = -k_Y \cdot \gamma \cdot E_f^{EROD}(t) \cdot Y(t) \tag{10}$$

$$\frac{dE_f^{EROD}(t)}{dt} = -k_X \cdot E_f^{EROD}(t) \cdot X(t) \cdot \frac{k_i^n}{k_i^n + Y(t)^n} - k_Y \cdot E_f^{EROD}(t) \cdot Y(t) + \frac{f(X)}{\gamma} + \frac{g(Y)}{\gamma} + \frac{k_{OX}}{\gamma} \cdot E_{OX}(t) + \frac{k_{OY}}{\gamma} \cdot E_{OY}(t) - k_D \cdot E_f^{EROD}(t) \tag{11}$$

$$\frac{dE_{OX}(t)}{dt} = k_X \cdot \gamma \cdot E_f^{EROD}(t) \cdot X(t) \cdot \frac{k_i^n}{k_i^n + Y(t)^n} - k_{OX} \cdot E_{OX}(t) \quad (12)$$

$$\frac{dE_{OY}(t)}{dt} = k_Y \cdot \gamma \cdot E_f^{EROD}(t) \cdot Y(t) - k_{OY} \cdot E_{OY}(t). \quad (13)$$

Hence, we model the EROD activity that can be compared with the data.[21,22] Equation (11) is the rate of change in EROD activity of free CYP1A enzymes (E_f^{EROD}) over time. Note that the two functions $f(X)$ and $g(Y)$ have not been changed and are given by equations (4) and (5).

2.2 Experimental data used. The PLHC-1 is an established cell line used in aquatic toxicology and suggested as an useful tool for mechanistic studies of regulation and function of CYP1A.[23,24] So far, only one CYP1A immunoreactive protein has been detected in PLHC-1 cells treated with BNF.[25] In addition, a partial CYP1A cDNA sequence was isolated from BNF treated PLHC-1 cells.[20] The data used to estimate the parameters in the model were obtained from a previous study using PLHC-1 cells that had been treated with the carrier vehicle and different doses of BNF (0.1 and 1 μM) and NOC (1, 10 and 25 μM), alone or mixed together, and measured at five different times (6, 12, 24, 48 and 72h).[21,22] The CYP1A-mediated EROD activities were analyzed in that study.[21,22] Data from four biological replicates (*i.e.* four separate experiments each of which representing the mean of four technical replicates) were used during the parameter estimation.[22] The raw data used to parameterize and validate the model are given in Table S1. The effects of BNF and NOC differs, with BNF 10 times more potent and 50 times more effective compared to NOC for activation of the AhR-CYP1A signaling.[20,22] However, the impact of NOC is still significant since the concentration of NOC added to the cells are up to 250 times higher than for BNF and NOC is a potent inhibitor of the CYP1A activity.[21,22]

2.3 Parameter estimation. The model was implemented using the R software.[26] The model has two parameters (n and γ) that were fixed before the estimation procedure. The parameter n is the Hill coefficient and represents the inhibition of the BNF biotransformation by the NOC molecules. We proposed a set {2, 3, 4, 6, 10} of possible values for n and performed the estimation procedure for each of them. Based on this, the value of n was set to 4 since that made the remaining parameters best able to fit the data. The proportionality constant γ in equation (8) was set to 1 min · mg · L^{-1}. The remaining ten parameters to be estimated are denoted by the vector:

$$\boldsymbol{p} = [k_X \quad k_Y \quad k_i \quad c_X \quad c_Y \quad k_{AX} \quad k_{AY} \quad k_{OX} \quad k_{OY} \quad k_D]^T.$$

The initial values for the concentrations of BNF (X) and NOC (Y) were set to the concentrations that the cells have been dosed with at $t = 0$. For each of the six treatments used to develop the model, the initial values of the state variables were therefore set to:

Treatment 1 **Treatment 2** **Treatment 3** **Treatment 4**

$$\begin{bmatrix} X^0 \\ Y^0 \\ E_f^{EROD\,0} \\ E_{OX}{}^0 \\ E_{OY}{}^0 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} X^0 \\ Y^0 \\ E_f^{EROD\,0} \\ E_{OX}{}^0 \\ E_{OY}{}^0 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} X^0 \\ Y^0 \\ E_f^{EROD\,0} \\ E_{OX}{}^0 \\ E_{OY}{}^0 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 10 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} X^0 \\ Y^0 \\ E_f^{EROD\,0} \\ E_{OX}{}^0 \\ E_{OY}{}^0 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 25 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Treatment 5 **Treatment 6**

$$\begin{bmatrix} X^0 \\ Y^0 \\ E_f^{EROD\,0} \\ E_{OX}{}^0 \\ E_{OY}{}^0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} X^0 \\ Y^0 \\ E_f^{EROD\,0} \\ E_{OX}{}^0 \\ E_{OY}{}^0 \end{bmatrix} = \begin{bmatrix} 1 \\ 25 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{14}$$

To estimate the parameters in the model, the EROD data at the five time points for the one single treatment and the five different mixture treatments have been used. To compare the

simulated EROD activity of free CYP1A enzymes from the model against its observations, we define a cost function, *cost(**p**)*, as:[27]

$$\text{cost}(\boldsymbol{p})=\sum_{i=1}^{6}\sum_{j=1}^{5}\sum_{k=1}^{4}[\log((E_{f,model}^{EROD})_j^i+1)-\log((E_{f,experiment}^{EROD})_j^{i,k}+1)]^2. \quad (15)$$

Equation (15) is the sum of squares of the logarithmic residuals of the EROD activity of free CYP1A enzymes from the model ($E_{f,model}^{EROD}$) versus its experimental value ($E_{f,experiment}^{EROD}$), represented by EROD data. It should be noted that in the cost function one is added to the values of $E_{f,model}^{EROD}$ and $E_{f,experiment}^{EROD}$ to avoid numerical instability. We summed over six treatments with four biological replicates at five time points. The estimated values for the parameters in $\boldsymbol{p}$ are those that minimize $\text{cost}(\boldsymbol{p})$. The *modFit* function from the flexible modelling environment (*FME*) package in R[28] was used to perform the box constraint optimization. This method is appropriate to use because of non-negativity constraints on the parameters. The parameters in the model (*i.e.* constants) and their fixed or fitted values are listed in Table 1.

2.4 Global Sensitivity Analysis (GSA). GSA is a tool to describe how the uncertainty in the model parameters can influence the uncertainty in the model output.[29] GSA was performed to provide an overview of the sensitivity of the EROD activity of free CYP1A enzymes to uncertainty in the parameter values. GSA identifies the most influential parameters on the model output and identifies parameters that the model output is insensitive to.

The global sensitivity of the EROD activity of free CYP1A enzymes to changes in each parameter (Table 1), was analyzed using experimental EROD data.[21,22] The GSA was performed using the *senseRange* function from the *FME* package in R.[28] For each of the parameters, a random sample of 1000 values was drawn using a log-uniform distribution on the interval from the estimated value divided by 10 to the estimated value multiplied by 10. By

using this distribution, the expected number of values in the sample below and above the estimated value will be equal.

Next, the global sensitivity of the EROD activity of free CYP1A enzymes was estimated by varying one parameter at the time using the sample described above and fixing the remained parameters at their nominal values. The five coupled ODEs in equations (9)-(13) were solved numerically using the *ode* function from the *deSolve* package in R.[30] This was carried out to calculate the EROD activity of free CYP1A enzymes over time for each parameter set, and for each of the six different treatments represented in equation (14).

2.5 Local Sensitivity Analysis (LSA). A parameter is practically non-identifiable if it is not possible to determine a unique value for it through fitting the model to the data. The two main sources of practical non-identifiability were analyzed: (1) lack of influence of a parameter on the EROD activity of free CYP1A enzymes (E_f^{EROD}) as the measurable model output and; (2) interdependence among the parameters.[31]

The LSA was performed to assess the sensitivity of the EROD activity of free CYP1A enzymes (E_f^{EROD}), as the measurable model output, with respect to small changes in the estimated parameters (Table 1). The sensitivity of the EROD activity of free enzymes to change in the parameter p_l, while all other parameters were fixed at their nominal values, was computed at the five time points for each treatment experiment through:

$$s_l = \frac{\partial E_f^{EROD}}{\partial p_l}, l = 1, \dots, 10. \quad (16)$$

The sensitivity values were estimated numerically using the *sensFun* function from the *FME* package in R.[28] The parameter value p_l was perturbed by 1%. In order to take into account

changes in time and across experiments the root mean squared sensitivity was computed for each parameter p_l:

$$s_l^{msqr} = \sqrt{\frac{1}{n}\sum_{q=1}^{n}(s_l)_q^2}\ , l = 1, \dots, 10. \quad (17)$$

This was summed over all five time points for each EROD data in the six different treatment experiments *i.e.* n=$\sum_{i=1}^{6}\sum_{j=1}^{5} 1$. The collinearities for all combinations of the ten parameters in $\boldsymbol{p}$ were tested using the *Collin* function from the *FME* package in R.[28]

3. RESULTS AND DISCUSSION

The aim of the present study was to provide a new mathematical bottom-up model to describe the synergistic mixture effect between two different classes of chemicals, on a mechanistic level. Experimental data on CYP1A biomarker responses in PLHC-1 exposed to BNF and NOC alone or in binary mixtures were used. The model successfully predicts the changes in CYP1A-mediated EROD activities of free CYP1A enzymes over time by fitting the model to experimental EROD data. In addition, ten parameters could be estimated in the model.

3.1 Model Predictions. The parameters in the model were first estimated using EROD data from six different treatment experiments.[21,22] The fitted values for the EROD activity of free CYP1A enzymes from the model are shown in Figure 2. The optimal value of cost function in equation (15) was equal to 13.63. This value corresponds to the mean value of the fraction $(E_{f,model}^{EROD}+1)/(E_{f,experiment}^{EROD}+1)$ equals to 1.05, which is good since it is close to a value of 1.00 that corresponds to a perfect fit.

The fitted values for the number of free CYP1A enzymes in cells co-treated with 1, 10 or 25 μM NOC, together with either 0.1 or 1 μM BNF, were in good agreement with the experimental data (Figure 2A-B). In addition, there is a satisfactory agreement between the fits for the numbers of free CYP1A enzymes over time in cells treated with 0.1 μM BNF alone and the experimental data[21,22] (Figure 2A). Since no significant induction of CYP1A activities could be measured in cells treated with 1, 10 or 25 μM NOC alone (Table S3), compared to that in vehicle-control cells, no comparison with fitted values were made for those treatments.

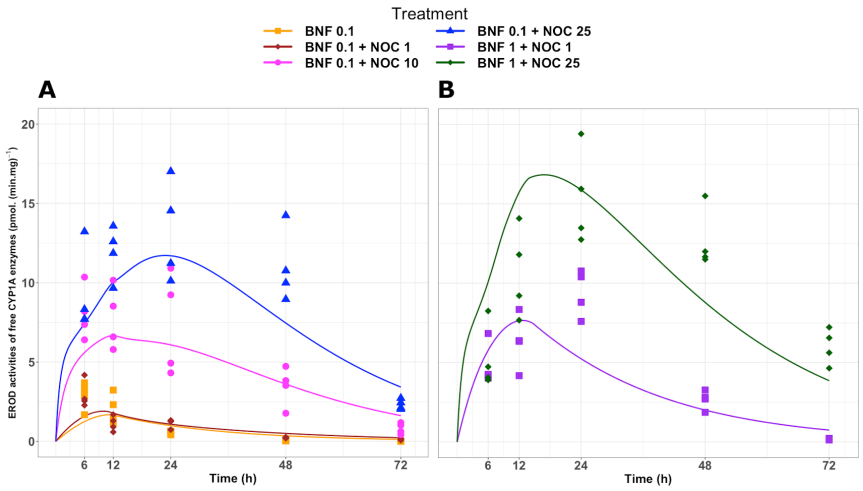


Figure 2. Fitted values for EROD activity of free CYP1A enzymes. The solid lines depict the EROD activity of free CYP1A enzymes from the model fitted to the experimental EROD data (pmol · (min · mg protein)$^{-1}$) (Table S1). The model is described by equations (4), (5), (9)-(13) and the parameters are listed in Table 1. The circles, triangles, squares and rhombuses represent EROD data from six independent experiments where the cells are exposed to: **A**. **Treatment 1)** 0.1 μM BNF (orange squares), **Treatment 2)** 0.1 μM BNF + 1 μM NOC (brown rhombuses), **Treatment 3)** 0.1 μM BNF + 10 μM NOC (pink circles) and **Treatment 4)** 0.1 μM BNF + 25 μM NOC (blue triangles), **B**. **Treatment 5)** 1 μM BNF + 1 μM NOC (purple squares) and **Treatment 6)** 1 μM BNF + 25 μM NOC (green rhombuses). In all treatments, the EROD activities have been measured at five different time points from 6 to 72h.[21,22] All experimental data are provided in Table S1.

The model predicts the four concentration state variables, X, Y, E_{OX} and E_{OY}, over time (Figure 3). The model shows that increasing the concentration of NOC (at t=0) from 1 to 10 or

25 μM results in slower BNF biotransformation rates. In fact, the model predicts no BNF biotransformation during the first 5-6h in cells co-treated with 25 μM NOC and BNF (Figure 3A). This supports our hypothesis that the presence of 25 μM NOC delays the elimination of BNF molecules, which means that more CYP1A enzymes are occupied by BNF after 6h compared to cells that have been co-exposed with a low concentration (1 μM) of NOC (Figure 3C). The model predicts that most of the CYP1A enzymes are being occupied by NOC molecules in a NOC dose-dependent manner during the first 6 hours (Figure 3D). In accordance, the model predicts that almost no CYP1A enzymes are occupied by BNF during the first 6h in the presence of either 10 or 25 μM NOC. The time delay for BNF to bind to free CYP1A enzymes is about 5h shorter for the cells co-treated with 1 μM NOC compared to cells treated with 25 μM NOC (Figure 3A and 3C). We conclude that it is the delayed elimination of BNF by NOC inhibition of CYP1A enzymes that causes the synergistic mixture effect on the CYP1A mediated EROD activity.

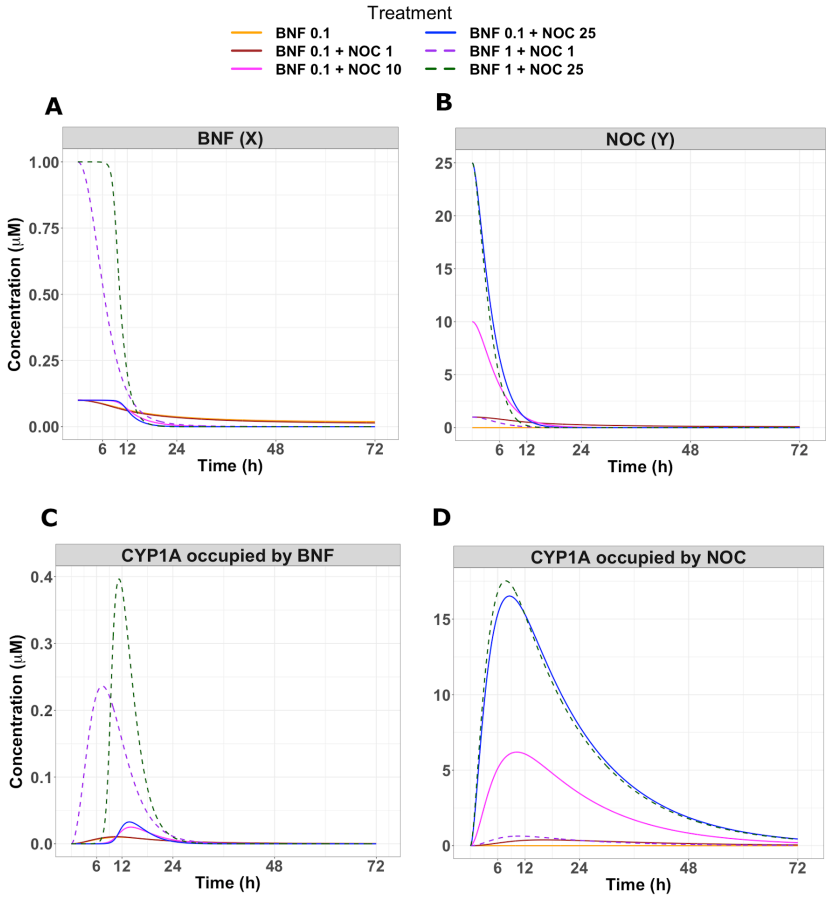


Figure 3. Model predictions for four state variables. A. Model predictions of BNF (X) concentrations; **B.** Model predictions of NOC (Y) concentrations; **C.** Model predictions of number of CYP1A enzymes occupied by BNF (E_{OX}); **D.** Model predictions of number of CYP1A enzymes occupied by BNF (E_{OY}). The plots depict the model predictions based on the parameters derived from the fit to experimental EROD data (Table S1) from six different treatment experiments: **Treatment 1)** 0.1 μM BNF (orange solid line), **Treatment 2)** 0.1 μM BNF + 1 μM NOC (brown solid line), **Treatment 3)** 0.1 μM BNF + 10 μM NOC (pink solid

line), **Treatment 4)** 0.1 μM BNF + 25 μM NOC (blue solid line), **Treatment 5)** 1 μM BNF + 1 μM NOC (purple dashed line) and **Treatment 6)** 1 μM BNF + 25 μM NOC (green dashed line). The model is described by equations (4), (5) and (9)-(13) and the parameters are listed in Table 1. In all treatments, the EROD activities have been measured at five different time points from 6 to 72h.[21,22]

3.2 Model Validation. The model was validated with data from a seventh experiment[21,22], **Treatment 7)** 1 μM BNF + 10 μM NOC shown in Figure 4, which was not used in the parameter estimation procedure.

Treatment 7 used for Model Validation

$$\begin{bmatrix} X^0 \\ Y^0 \\ E_f^{EROD\,0} \\ E_{OX}{}^0 \\ E_{OY}{}^0 \end{bmatrix} = \begin{bmatrix} 1 \\ 10 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

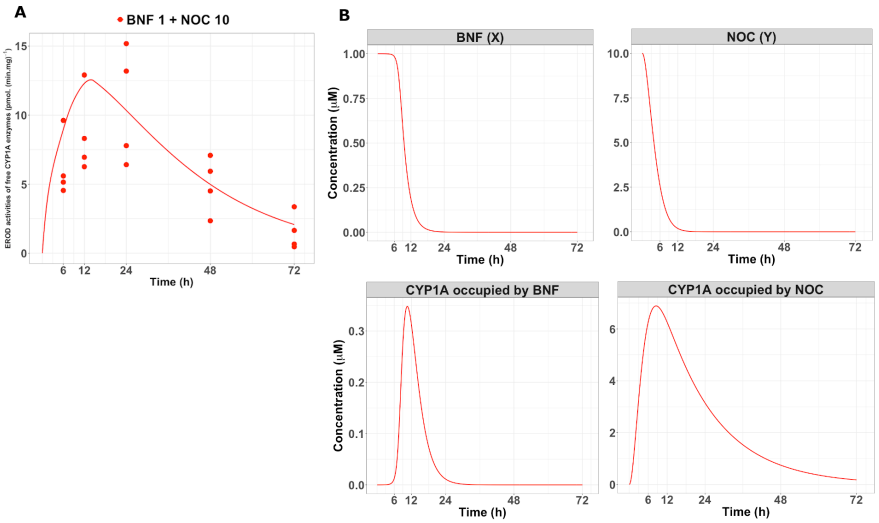


Figure 4. Model Validation. A. The model was validated using data from an additional experiment, **Treatment 7)** 1 μM BNF + 10 μM NOC (red circles). The model prediction for EROD activity of free CYP1A enzymes from the model (solid red line) is compared with the experimental EROD data (pmol · (min · mg protein)$^{-1}$) (Table S1).[21,22] **B**. Model predictions of changes in the concentrations of BNF (X), NOC (Y) molecules over time (top panel). Model predictions of the changes in numbers of CYP1A enzymes occupied by BNF (E_{OX}) and NOC (E_{OY}) over time (bottom panel). The model is described by equations (4), (5) and (9)-(13) and the parameters are listed in Table 1. The experimental data are provided in Table S1.

The model prediction for the EROD activity of free CYP1A enzymes is in good agreement with the experimental values of EROD activities from **Treatment 7** (Figure 4A). The model predicts how the numbers of CYP1A enzymes occupied by BNF and NOC change over time (Figure 4B, bottom panel). The model also predicts that BNF is not being

biotransformed by CYP1A enzymes during the first 5h in the presence of 10 µM NOC (Figure 4B, top panel).

The validation test shows that the model has good prediction power and can be used to test different combinations of two chemicals and their effect on the EROD activity of free CYP1A enzymes (Figure 4).

3.3 Global Sensitivity Analysis. The ranges from the GSA for the EROD activity of free CYP1A enzymes are illustrated using 5-95%, 25-75% and 50% (median) quantiles (Figure 5).

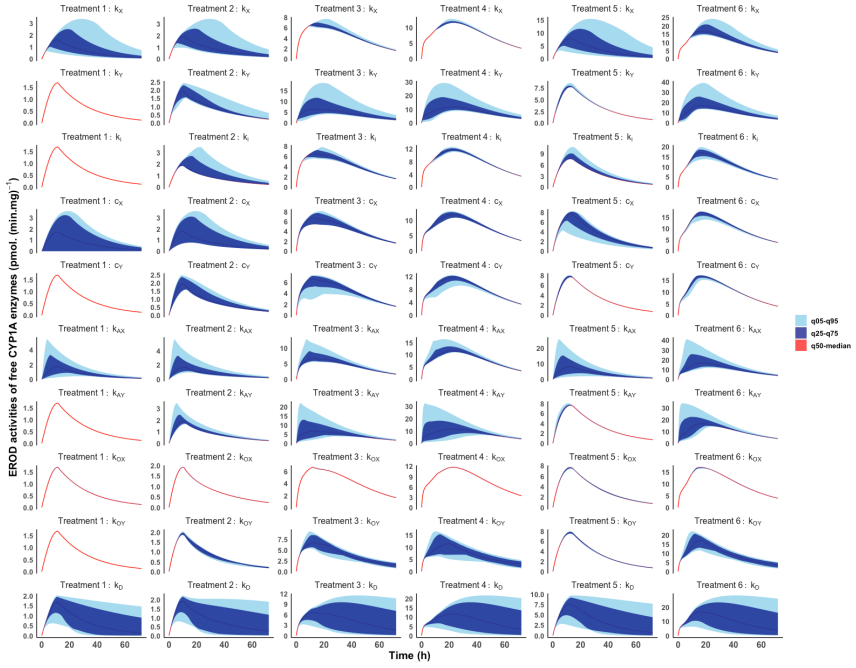


Figure 5. Global Sensitivity Analysis. Sensitivity range of the EROD activities[22] of free CYP1A enzymes over time to changes in one parameter per row is illustrated. The sensitivity ranges are depicted by using 5-95% (light blue), 25-75% (dark blue) and 50% (red) quantiles

of the EROD activities of free CYP1A values (vertical axis) estimated by the model on the time interval from dosing (t = 0) to t=72h (horizontal axis) for six different treatments: **Treatment 1)** 0.1 μM BNF; **Treatment 2)** 0.1 μM BNF + 1 μM NOC; **Treatment 3)** 0.1 μM BNF + 10 μM NOC; **Treatment 4)** 0.1 μM BNF + 25 μM NOC; **Treatment 5)** 1 μM BNF + 1 μM NOC and **Treatment 6)** 1 μM BNF + 25 μM NOC.

The GSA revealed a high sensitivity of the EROD activity of free CYP1A enzymes in all treatments to changes in the CYP1A degradation rate (k_D). To reduce the uncertainty in the model predictions, the value of k_D should therefore be estimated with low uncertainty. In addition, the GSA indicated that changes in the biotransformation rate of BNF (k_{OX}) has no or insignificant individual effect on the robustness of the dynamics of free CYP1A enzymes. To rank the k_{OX} as a non-influential parameter on the dynamics of free CYP1A enzymes, further analysis is required to calculate the joint effect of this parameter due to its interaction with the other parameters. This can be beneficial for model simplification, but was not the main focus of this study. Hence, the GSA revealed that k_D is an influential parameter. It also confirmed that the number of free CYP1A enzymes over time is more sensitive to the changes in the threshold concentration of BNF to induce EROD activity (c_X) compared to the changes in the threshold concentration of NOC to induce EROD activity (c_Y) (Figure 5).

3.4 Local Sensitivity Analysis and practical identifiability. The first source of practical non-identifiability is assessed by computing the sensitivity values of E_f^{EROD} using equation (16). The LSA indicated that the three parameters k_D, k_X and k_Y, have the largest negative average effects on the EROD activity of free CYP1A enzymes and are followed by c_X, c_Y and k_i. The other four parameters on average have positive effects on the EROD activity of free enzymes,

with k_{OX} being the parameter with the least effect (Table 2). This is in good agreement with the results from the GSA (Figure 5). The sensitivity values of the ten parameters at each time point for the six treatments are provided in Figure S1.

Table 2. Statistics of Local Sensitivity Analysis of EROD Activity of Free CYP1A Enzymes.[b]

Parameter	s_l^{msqr}	Mean
k_X	38.17	-14.61
k_Y	79.06	-29.28
k_i	0.45	-0.17
c_X	31.29	-12.19
c_Y	1.23	-0.47
k_{AX}	2.67	0.94
k_{AY}	14.69	5.32
k_{OX}	0.08	0.01
k_{OY}	15.12	6.20
k_D	147.29	-54.59

[b] The s_l^{msqr} are the root mean squared sensitivity measures defined in equation (17) and the mean values are the average of the sensitivity values illustrated in Figure S1.

A parameter with no or insignificant effect on the EROD activity of free CYP1A enzymes was classified as a practically non-identifiable parameter. It has been suggested that the threshold value classified as a non-identifiable parameter is four orders of magnitude lower than the maximum root mean squared value.[31] All of the ten parameters are above this cut-off value of 0.014 (Table 2). Hence, all the ten model parameters have significant effects on the

EROD activity of free CYP1A enzymes with the experimental EROD data. This indicates a strength of the model since the first source of practical non-identifiability is not a problem.

The parameters may also interfere with each other. Due to interdependence among the parameters, the possible effect of each parameter on the EROD activity of free CYP1A enzymes may be compensated by change(s) in other parameter(s), known as parameter collinearity. A parameter set with collinearity index above 20 is considered as a non-identifiable set.[32]

The maximum values of collinearity indices for sets inholding different combinations of the parameters two-ten were 5.60, 8.06, 9.54, 11.58, 14.43, 15.07, 15.58, 16.50 and 17.39, respectively. The collinearity index for all ten parameter combinations are provided in Figure S2. The collinearity analysis indicated that all sets with different combinations of parameters had a collinearity index below 20. This led us to conclude that by using experimental EROD data, unique values for each of the ten parameters in the model can be estimated simultaneously. Moreover, non-identifiability due to collinearity between parameters is not a problem.

Azoles have been shown to interact with CYP enzymes, including CYP1A in fish.[13-18] For example, ketoconazole was shown to act as a potent non-competitive inhibitor of microsomal CYP1A activities in Atlantic cod.[18] Co-exposure of ketoconazole that inhibits CYP1A and CYP3A enzymes resulted in increased sensitivity to ethinylestradiol exposure in rainbow trout.[19] Synergistic effects between antifungal azoles and insecticides have earlier been reported. Thus, the azole prochloraz inhibited the biotransformation of a pyrethroid, resulting in increased insecticide toxicity in honeybees co-exposed to prochloraz.[33] The azole mediated inhibition of CYP-dependent detoxification of pesticide was suggested being the main mechanism behind the synergizing effect of azoles on pesticide toxicity.[12] However, the synergistic effect of antifungal azoles on the pyrethroid insecticide toxicity was not correlated to azole inhibition strength on the CYP-mediated ethoxycoumarin-*O*-deethylase (ECOD) activity in two aquatic invertebrates, implying that the mechanisms behind the synergism is more complex.[34] There also seem to be species differences in enzyme susceptibility towards

azoles.[35] In PLHC-1 cells, co-exposure to the azole NOC delayed the response to BNF exposures indicating that NOC prevents biotransformation of BNF, presumably by inhibition of CYP1A enzymes.[21] This suggests a delayed CYP1A-mediated biotransformation of BNF in the presence of the NOC. Delayed biotransformation for benzo[*a*]pyrene was demonstrated in two rainbow trout cell lines (RTL-W1 and RTgutGC) co-exposed with the CYP1A (EROD) inhibitor α-naphthoflavone.[36] Hence, inhibition of CYP metabolism can increase the biological half-life of aromatic hydrocarbons resulting in increased sensitivity to exposures to aromatic hydrocarbons.

The synergistic mixture effect with α-cypermethrin and two azoles in *Daphnia magna* correlated with inhibition of the CYP-mediated EROD activity by prochloraz and to some extent with the inhibition by propiconazole. It was suggested that a toxicokinetic and toxicodynamic model could be a tool to test mechanisms of interactions between chemicals.[12] In the present study, we also use a toxicokinetic approach focusing on the inhibition of the CYP1A enzyme activity by NOC, by using a Hill function to model the competition between the azole NOC and the AhR agonist BNF. The time that these two chemicals are competing and the steepness of the rate of biotransformation of BNF depend on the Hill coefficient, which was fixed at a preselected value to optimize the model fit to the data. Here, the model might not give a real estimation on how the biotransformation of BNF is controlled by NOC. The model can be further refined in future studies by including chemical data.

The new model presented here successfully predicts the changes in EROD activities of free CYP1A enzymes over time by fitting the model to experimental EROD data with given mixtures of BNF and NOC. Ten parameters could be estimated in the model. We hypothesize that the synergistic effect is a result of NOC mediated inhibition of the CYP1A dependent clearance of BNF. Synergistic mixture effects were seen with two other azoles, clotrimazole and prochoraz in PLHC-1 cells. These azoles also acted as inhibitors of the EROD activities (having IC_{50} values below 10 μM). The azole omeprazole, on the other hand, did not

significantly inhibit EROD activities (having an IC_{50} value above 50 μM), and there was no synergistic mixture effect when BNF was mixed with omeprazole.[21] This supports our hypothesis that inhibition of CYP1A activities triggers a synergistic mixture effect. A similar synergistic mixture effect on the CYP1A biomarker has been observed in cells exposed to an another AhR agonist, the polycyclic aromatic hydrocarbon benzo[*a*]pyrene, in combination with an antifungal imidazole drug clotrimazole (Alvord, C.; Lundh, T.; Wiklander, K.; Bernhardsson, A.; Celander, M.C. data not shown). Hence, the model has a potential to be used for other chemical mixtures.

Sensitivity and identifiability analysis revealed that the parameter corresponding to the rate of CYP1A enzyme degradation is the most influential parameter on the dynamics of the EROD activity of free CYP1A enzymes predicted by the model. In contrast, the parameter related to biotransformation of BNF is the parameter with the least individual influence on this variable. Hence, to reduce the total uncertainty in the model predictions for the EROD activity of CYP1A enzymes the parameter corresponding to the rate of CYP1A enzyme degradation should be estimated with low uncertainty. The present study provides a new promising toxicokinetic model with predictive power to describe synergistic mixture effects between aromatic hydrocarbons and azoles.

ASSOCIATED CONTENT

Supporting Information

Table S1: Raw data for EROD activities in seven treatment groups from the previously published study.[21,22]

Table S2: Raw data from a dose response experiment with β-naphthoflavone (BNF) alone on ethoxyresorufin-*O*-deethylase (EROD) activities at 24 hours in vehicle-control (DMSO) treated cells and in cells treated with 0.1, 0.3, 0.9, 1, 3, 8 or 25 μM BNF alone from the previously published study.[22]

Table S3: Raw data from dose response experiment with nocodazole (NOC) alone on ethoxyresorufin-*O*-deethylase (EROD) activities at 6, 12, 24, 48 and 72h in vehicle-control (DMSO) treated cells and in cells treated with 1, 10 or 25 μM NOC alone from the previously published study.[21,22]

Figure S1: The sensitivity values of the ten parameters at each time point for the six treatments.

Figure S2: Collinearity index for all parameter combinations.

AUTHOR INFORMATION

Corresponding Author

*Phone: +46 (0)31 786 3693; E-mail: malin.celander@gu.se

ORCID

Shirin Fallahi: 0000-0002-9867-5596

Marie Mlnaříková: 0000-0002-7783-982X

Guttorm Alendal: 0000-0001-5757-4345

Håvard G. Frøysa: 0000-0002-5161-477X

Torbjörn Lundh: 0000-0001-7081-1384

Malin C. Celander: 0000-0002-4350-751X

Present Address

§ Formerly Marie Michelová. Recetox, Faculty of Science, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic

AUTHOR CONTRIBUTION

TL created the initial model. SF performed the mathematical modeling and mathematical analysis and co-wrote the manuscript. MM performed the time course experiments and the EROD analyses and commented on the manuscript. CA participated in the discussions and meetings, before leave of absence, and has approved the submission. GA, HF and TL participated in the discussions and commented on the manuscript. MC supervised the study and co-wrote the manuscript. SF, MM, HF, GA, TL and MC have participated in the revision process and have read and approved the final version of the manuscript.

DISCLOSURE

The authors declare no conflicts of interest.

ACKNOWLEDGEMENT

This work is a contribution to the FORMAS project number 942-2015-605 to MC and TL and the dCod 1.0 project funded by Research Council of Norway through grant number 248840 to SF, HF, GA and MC. We thank Libe Aranguren-Abadía from the Department of Biosciences at University of Bergen for helping with preparing the graphics.

ABBREVIATIONS

AhR, aryl hydrocarbon receptor; BNF, β-naphtoflavone; CYP1A, cytochrome P450 1A; GSA, global sensitivity analysis; EROD, ethoxyresorufin-*O*-deethylase; FME, flexible

modelling environment; LSA, local sensitivity analysis; NOC, nocodazole; ODE, ordinary differential equation; PLHC-1, *Poeciliopsis lucida* hepatocellular carcinoma

REFERENCES

(1) Hinton, D.E.; Segner, H. Au, D.W.T.; Kullman, S.W.; Hardman, R.C. Liver Toxicity. *In:* The Toxicology of Fishes. Chapter 7. Edited by Richard T. Di Gulio and David E. Hinton. CRS Press. Taylor & Francis Group. **2008,** pp. 328-352.

(2) Fent, K.; Weston, A.A.; Caminada, D. Ecotoxicology of human pharmaceuticals. *Aquat. Toxicol*. **2006,** 76, 122-159.

(3) Corcoran, J.; Winter, M.J.; Tyler, C.R. Pharmaceuticals in the aquatic environment: A critical review of the evidence for health effects in fish. *Crit. Rev. Toxicol*. **2010,** 40, 287-304.

(4) Celander, M.C. Cocktail effects on biomarker responses in fish. *Aquat. Toxicol*. **2011,** 105S, 72-77.

(5) Groten, J.P.; Feron, V.J.; Sühnel, J. Toxicology of simple and complex mixtures. *Trends in Pharmacological Sciences*. **2001,** 22, 316-322.

(6) Hadrup, N.; Taxvig, C.; Pedersen, M., Nellemann, C.; Hass, U.; Vinggaard, A.M. Concentration addition, independent action and generalized concentration addition models for mixture effect prediction of sex hormone synthesis *in vitro*. *PLOS ONE*. **2013,** 8, e70490.

(7) Rajapakse, N.; Silva, E.; Scholze, M.; Kortenkamp, A. Deviation from additivity with estrogenic mixtures containing 4-nonylphenol and 4-tert-octylphenol detected in the E-SCREEN assay. *Environ. Sci. Technol*. **2004**, 6343-6352.

(8) Crofton, K.M.; Craft, E.S.; Hedge, J.M.; Gennings, C.; Simmons, J.E.; Carchman, R.A.; Carter Jr, W.H.; DeVito, M.J. Thyroid-hormone–disrupting chemicals: evidence for dose-dependent additivity or synergism. *Environ. Health Perspect*. **2005**, 1549-1554.

(9) Hass, U.; Scholze, M.; Christiansen, S.; Dalgaard, M.; Vinggaard, A.M.; Axelstad, M.; Metzdorff, S.B.; Kortenkamp, A. Combined exposure to anti-androgens exacerbates

disruption of sexual differentiation in the rat. *Environ. Health Perspect*. **2007**, 115, Suppl. 1, 122-128.

(10) Charles, G.D.; Gennings, C.; Tornesi, B.; Kan, H.L.; Zacharewski, T.R.; Gollapudi, B.B.; Carney, E.W. Analysis of the interaction of phytoestrogens and synthetic chemicals: an in vitro/in vivo comparison. *Toxicol. Appl. Pharmacol.* **2007**, 218, 280-288.

(11) Bobb, S.; Berggren, E.; Kienzler, A.; van der Linden, S.; Wort, A. Scientific methodologies for the assessment of combined effects of chemicals – a survey and literature review. Use of novel and alternative methods in the assessment of effects from combined exposure to multiple chemicals. *JRC Technical Reports.* EUR 27471 EN. **2015.**

(12) Cedergreen, N.; Dalhoff, K.; Gottardi, M.; Kretschmann, A.C. Can toxicokinetic modeling be used to understand and predict synergistic interactions between chemicals? *Environ. Sci. Technol.* **2017,** 51, 14379-14389.

(13) Bach, J.; Snegaroff, J. Effects of the fungicide prochloraz on the xenobiotic metabolism in Rainbow trout – *in vivo* induction. *Xenobiotica.* **1989,** 19, 1-9-

(14) Levine, S.L.; Czosnyka, H.; Oris, J.T. Effects of the fungicide clotrimazole on bioconcentration of benzo[*a*]pyrene in gizzard shad (*Dorosoma cepedianum*): In vivo and in vitro inhibition of CYP1A activity. *Environ. Toxicol. Chem*. **1997,** 16, 306-311.

(15) Egaas, E.; Sandvik, M.; Fjeld, E.; Källqvist, T.; Goksøyr, A.; Svensen, A. Some effects of the fungicide propiconazole on cytochrome P450 and glutathione *S*-transferase in brown trout (*Salmo trutta*). *Comp. Biochem. Physiol*. C. **1999,** 112, 337-344.

(16) Sturm, A.; Cravedi, J.P.; Perdu, E.; Baradat, M.; Segner, H. Effects of prochloraz and nonylphenol diethoxylate on hepatic biotransformation in trout: A comparative in vitro/in vivo-assessment using cultured hepatocytes. *Aquat. Toxicol*. **2001,** 53, 229-245.

(17) Hegelund, T.; Ottosson, K.; Tomberg, P.; Rådinger, M.; Celander, M.C. Effects of ketoconazole on CYP1A and CYP3A in rainbow trout and killifish. *Environ. Toxicol. Chem*. **2004,** 23, 1326-1334.

(18) Hasselberg, L.; Grøsvik, B.-E.; Goksøyr, A.; Celander, M.C. Interactions between xenoestrogens and ketoconazole on CYP1A and CYP3A in juvenile Atlantic cod (*Gadus morhua*). **2005,** *Comp. Hepatol.* 4:2

(19) Hasselberg, L.; Westerberg, S.; Wassmur, B.; Celander, M.C. Ketoconazole, an antifungal imidazole, increases the sensitivity of rainbow trout to 17α-ethynylestradiol exposure. *Aquat. Toxicol*. **2008,** 86, 256-264.

(20) Wassmur, B.; Gräns, J.; Norström, E.; Wallin, M.; Celander, M.C. Interactions of pharmaceuticals and other xenobiotics on key detoxification mechanisms and cytoskeleton in *Poeciliopsis lucida* hepatocellular carcinoma, PLHC-1 cell line. *Toxicol. In Vitro.* **2013,** 27, 111-120.

(21) Gräns, J.; Johansson, J.; Michelová, M.; Wassmur, B.; Norström, E.; Wallin, M.; Celander, M. C. Mixture effects between different azoles and β-naphthoflavone on the CYP1A biomarker in a fish cell line. *Aquat. Toxicol.* **2015,** 164, 43-51.

(22) Michelová, M. Cocktail effects of β-naphthoflavone and nocodazole in the *Poeciliopsis lucida* cell line. **2012,** Master-thesis, University of Gothenburg.

(23) Segner, H. Fish cell lines as a tool in aquatic toxicology. *In:* Fish Ecotoxicology. Eds. T. Branbeck, D.E. Hinton, B. Streit. Springer Basel AG. **1998,** pp. 1-38.

(24) Fent, K. Fish cell lines as versatile tools in ecotoxicology: assessment of cytotoxicity, cytochrome P4501A induction potential and estrogenic activity of chemicals and environmental samples. *Toxicol. In Vitro*. **2001,** 15, 477-488.

(25) Celander, M; Bremer, J.; Hahn, M.E.; Stegeman, J.J. Glucocorticoid-xenobiotic interactions: dexamethasone-mediated potentiation of cytochrome P4501A by ß-naphthoflavone in a fish hepatoma cell line (PLHC-1). *Environ. Toxicol. Chem.* **1997,**

16, 900-907.

(26) R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. **2009,** URL https://www.R-project.org/

(27) Tofallis C. A better measure of relative prediction accuracy for model selection and model estimation. *J. Oper. Res. Soc.* **2015**, 66, 1352-1362.

(28) Soetaert, K.; Petzoldt, T. Inverse modelling, sensitivity and monte carlo analysis in R using package FME. *J. Stat. Softw.* **2010,** 33, 1-28.

(29) Soetaert, K.; Herman, P.M. A practical guide to ecological modelling: using R as a simulation platform. Springer Science & Business Media. **2008**.

(30) Soetaert, K.E.; Petzoldt, T.; Setzer, R.W. Solving differential equations in R: package deSolve. *J. Stat. Softw.* **2010,** 33.

(31) Gábor, A.; Villaverde, A. F.; Banga, J.R. Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems. *BMC Syst. Biol.* **2017,** 11, 54.

(32) Brun, R., Reichert, P.; Künsch, H.R. Practical identifiability analysis of large environmental simulation models. *Water Resour. Res.* **2001,** 37, 1015-1030.

(33) Pilling, E.D.; Bromley-Challenor, K.A.C.; Walker, C.H.; Jepson, P.C. Mechanism of synergism between th pyrethroid insecticide lambda-cyhalothrin and the imidazole fungicide prochloraz, in the honeybee (Apis mellifera L.). *Pestic. Biochem. Physiol*. **1995**, 51, 1-11

(34) Gottardi, M.; Cedergreen, N. The synergistic potential of azole fungicides does not directly correlate to the inhibition of cytochrome P450 activity in aquatic invertebrates. *Aquat. Toxicol*. **2019**, 207, 187-196.

(35) Gottardi, M.; Tyzack, J.D.; Bender, A.; Cedergreen, N. Can the inhibition of cytochrome P450 in aquatic invertebrates due to azole fungicides be estimated with in silico and in vitro models and extrapolated between species? *Aquat. Toxicol.* **2018**, 201, 11-20.

(36) Stadnicka-Michalak, J.; Weiss, F.T.; Fisher, M.; Tanneberger, K.; Schirmer, K. Biotransformation of benzo[a]pyrene by three rainbow trout (Onchorhynchus mykiss) cell lines and extrapolation to derive fish bioconcentration factor. *Environ. Sci. Technol*. **2018**, 52, 3091-3100.

Appendix to paper C.

TABLES

FIGURES

TABLES:

Table S1. Raw data for the ethoxyresorufin-*O*-deethylase (EROD) activity shown in Figures 2A, 2B and 4A. All times are given in hours. Treatments are denoted by 1-7. The same annotation is used in the model code. Data are originally published in figures as mean values ± standard deviation in Gräns et al. 2015 (reference 1) and Michelová M. Master-thesis 2012 (reference 2).

Time (h)	**EROD activity** (pmol · (min · mg protein)$^{-1}$)	**Treatment**	**Treatment number**
6	2.92190258	BNF 0.1 µM	1
6	3.69454764	BNF 0.1 µM	1
6	1.68262381	BNF 0.1 µM	1
6	3.3423465	BNF 0.1 µM	1
12	1.42052688	BNF 0.1 µM	1
12	2.31785379	BNF 0.1 µM	1
12	1.18018557	BNF 0.1 µM	1
12	3.23571203	BNF 0.1 µM	1
24	0.65168451	BNF 0.1 µM	1
24	0.5054846	BNF 0.1 µM	1
24	0.41898238	BNF 0.1 µM	1
24	0.61937856	BNF 0.1 µM	1
48	0.05222293	BNF 0.1 µM	1
48	0.08032718	BNF 0.1 µM	1
48	0.10036158	BNF 0.1 µM	1
48	0.03061984	BNF 0.1 µM	1
72	-0.0089312	BNF 0.1 µM	1
72	-0.0431119	BNF 0.1 µM	1
72	0.03056186	BNF 0.1 µM	1
72	0.00783302	BNF 0.1 µM	1
6	2.27896044	BNF 0.1 µM and NOC 1 µM	2

6	2.67142836	BNF 0.1 µM and NOC 1 µM	2
6	2.55450885	BNF 0.1 µM and NOC 1 µM	2
6	4.18042635	BNF 0.1 µM and NOC 1 µM	2
12	0.593757	BNF 0.1 µM and NOC 1 µM	2
12	0.91465491	BNF 0.1 µM and NOC 1 µM	2
12	1.29736125	BNF 0.1 µM and NOC 1 µM	2
12	1.69157719	BNF 0.1 µM and NOC 1 µM	2
24	1.25667054	BNF 0.1 µM and NOC 1 µM	2
24	1.28266786	BNF 0.1 µM and NOC 1 µM	2
24	0.74049556	BNF 0.1 µM and NOC 1 µM	2
24	1.32778702	BNF 0.1 µM and NOC 1 µM	2
48	0.27140646	BNF 0.1 µM and NOC 1 µM	2
48	0.2556882	BNF 0.1 µM and NOC 1 µM	2
48	0.2019469	BNF 0.1 µM and NOC 1 µM	2
48	0.21827206	BNF 0.1 µM and NOC 1 µM	2
72	0.1852296	BNF 0.1 µM and NOC 1 µM	2
72	0.14971151	BNF 0.1 µM and NOC 1 µM	2
72	-0.0035442	BNF 0.1 µM and NOC 1 µM	2
72	0.07820264	BNF 0.1 µM and NOC 1 µM	2
6	7.36722239	BNF 0.1 µM and NOC 10 µM	3
6	8.19225914	BNF 0.1 µM and NOC 10 µM	3
6	6.40147279	BNF 0.1 µM and NOC 10 µM	3
6	10.3568426	BNF 0.1 µM and NOC 10 µM	3
12	5.79331185	BNF 0.1 µM and NOC 10 µM	3
12	8.52501746	BNF 0.1 µM and NOC 10 µM	3
12	6.59014412	BNF 0.1 µM and NOC 10 µM	3
12	10.1562112	BNF 0.1 µM and NOC 10 µM	3
24	4.31962722	BNF 0.1 µM and NOC 10 µM	3
24	4.93575241	BNF 0.1 µM and NOC 10 µM	3
24	9.23688221	BNF 0.1 µM and NOC 10 µM	3
24	10.9139222	BNF 0.1 µM and NOC 10 µM	3
48	3.528014	BNF 0.1 µM and NOC 10 µM	3
48	1.77744018	BNF 0.1 µM and NOC 10 µM	3

48	4.73028336	BNF 0.1 µM and NOC 10 µM	3
48	3.83101002	BNF 0.1 µM and NOC 10 µM	3
72	0.61760342	BNF 0.1 µM and NOC 10 µM	3
72	0.42778956	BNF 0.1 µM and NOC 10 µM	3
72	1.18436321	BNF 0.1 µM and NOC 10 µM	3
72	1.02342749	BNF 0.1 µM and NOC 10 µM	3
6	7.71808341	BNF 0.1 µM and NOC 25 µM	4
6	8.32716239	BNF 0.1 µM and NOC 25 µM	4
6	7.69936717	BNF 0.1 µM and NOC 25 µM	4
6	13.2227151	BNF 0.1 µM and NOC 25 µM	4
12	11.8627936	BNF 0.1 µM and NOC 25 µM	4
12	12.5949582	BNF 0.1 µM and NOC 25 µM	4
12	9.66216727	BNF 0.1 µM and NOC 25 µM	4
12	13.5778835	BNF 0.1 µM and NOC 25 µM	4
24	10.1189154	BNF 0.1 µM and NOC 25 µM	4
24	11.2275265	BNF 0.1 µM and NOC 25 µM	4
24	14.5414457	BNF 0.1 µM and NOC 25 µM	4
24	17.0127185	BNF 0.1 µM and NOC 25 µM	4
48	14.2378479	BNF 0.1 µM and NOC 25 µM	4
48	9.99436876	BNF 0.1 µM and NOC 25 µM	4
48	10.7615408	BNF 0.1 µM and NOC 25 µM	4
48	8.9564097	BNF 0.1 µM and NOC 25 µM	4
72	2.44214987	BNF 0.1 µM and NOC 25 µM	4
72	2.12808275	BNF 0.1 µM and NOC 25 µM	4
72	2.72777848	BNF 0.1 µM and NOC 25 µM	4
72	2.03687811	BNF 0.1 µM and NOC 25 µM	4
6	4.25915447	BNF 1 µM and NOC 1 µM	5
6	4.15426073	BNF 1 µM and NOC 1 µM	5
6	4.02362663	BNF 1 µM and NOC 1 µM	5
6	6.83484415	BNF 1 µM and NOC 1 µM	5
12	6.33964085	BNF 1 µM and NOC 1 µM	5
12	8.34863858	BNF 1 µM and NOC 1 µM	5

12	4.17620219	BNF 1 µM and NOC 1 µM	5
12	6.38054672	BNF 1 µM and NOC 1 µM	5
24	10.405456	BNF 1 µM and NOC 1 µM	5
24	10.771562	BNF 1 µM and NOC 1 µM	5
24	7.59540869	BNF 1 µM and NOC 1 µM	5
24	8.8010075	BNF 1 µM and NOC 1 µM	5
48	2.70204088	BNF 1 µM and NOC 1 µM	5
48	3.27295458	BNF 1 µM and NOC 1 µM	5
48	2.8059326	BNF 1 µM and NOC 1 µM	5
48	1.8645738	BNF 1 µM and NOC 1 µM	5
72	0.12598001	BNF 1 µM and NOC 1 µM	5
72	0.19551305	BNF 1 µM and NOC 1 µM	5
72	0.2272445	BNF 1 µM and NOC 1 µM	5
72	0.14294592	BNF 1 µM and NOC 1 µM	5
6	4.05958302	BNF 1 µM and NOC 25 µM	6
6	4.73479989	BNF 1 µM and NOC 25 µM	6
6	3.9063292	BNF 1 µM and NOC 25 µM	6
6	8.25327557	BNF 1 µM and NOC 25 µM	6
12	9.21418146	BNF 1 µM and NOC 25 µM	6
12	11.803604	BNF 1 µM and NOC 25 µM	6
12	7.6698235	BNF 1 µM and NOC 25 µM	6
12	14.0849049	BNF 1 µM and NOC 25 µM	6
24	12.7508499	BNF 1 µM and NOC 25 µM	6
24	13.483253	BNF 1 µM and NOC 25 µM	6
24	15.9410841	BNF 1 µM and NOC 25 µM	6
24	19.4159461	BNF 1 µM and NOC 25 µM	6
48	15.5009285	BNF 1 µM and NOC 25 µM	6
48	11.9997667	BNF 1 µM and NOC 25 µM	6
48	11.5037771	BNF 1 µM and NOC 25 µM	6
48	11.6652826	BNF 1 µM and NOC 25 µM	6
72	6.55693615	BNF 1 µM and NOC 25 µM	6
72	5.61974464	BNF 1 µM and NOC 25 µM	6

72	4.64567565	BNF 1 µM and NOC 25 µM	6
72	7.22737517	BNF 1 µM and NOC 25 µM	6
6	4.54633972[a]	BNF 1 µM and NOC 10 µM	7
6	5.59647396[a]	BNF 1 µM and NOC 10 µM	7
6	5.1440204 [a]	BNF 1 µM and NOC 10 µM	7
6	9.61475399 [a]	BNF 1 µM and NOC 10 µM	7
12	6.26397851 [a]	BNF 1 µM and NOC 10 µM	7
12	8.31189843 [a]	BNF 1 µM and NOC 10 µM	7
12	6.94923936 [a]	BNF 1 µM and NOC 10 µM	7
12	12.9031626 [a]	BNF 1 µM and NOC 10 µM	7
24	6.41179367 [a]	BNF 1 µM and NOC 10 µM	7
24	7.78904675 [a]	BNF 1 µM and NOC 10 µM	7
24	13.1874913 [a]	BNF 1 µM and NOC 10 µM	7
24	15.1788122 [a]	BNF 1 µM and NOC 10 µM	7
48	4.5062918 [a]	BNF 1 µM and NOC 10 µM	7
48	2.3500834 [a]	BNF 1 µM and NOC 10 µM	7
48	7.08559516 [a]	BNF 1 µM and NOC 10 µM	7
48	5.93579771 [a]	BNF 1 µM and NOC 10 µM	7
72	0.66411867 [a]	BNF 1 µM and NOC 10 µM	7
72	0.46601484 [a]	BNF 1 µM and NOC 10 µM	7
72	1.65068662 [a]	BNF 1 µM and NOC 10 µM	7
72	3.36291459 [a]	BNF 1 µM and NOC 10 µM	7

[a] Data not included in the fit and used for validation.

Table S2. Raw data from a dose response experiment with β-naphthoflavone (BNF) alone on ethoxyresorufin-*O*-deethylase (EROD) activities at 24 hours. Dimethylsulphoxide (DMSO) was used as in vehicle-control. Six different doses of BNF was tested 0.1, 0.3, 0.9, 3, 8 and 25 μM. The data are originally presented in a diagram as mean values ± standard deviation in Michelová M. Master-thesis, 2012 (reference 2).

Time (h)	**EROD activity** (pmol · (min · mg protein)$^{-1}$)	**Treatment**
24	0.046551324	DMSO
24	0.060342951	DMSO
24	0.039346828	DMSO
24	0.084340084	DMSO
24	0.670409585	BNF 0.1 μM
24	0.674327937	BNF 0.1 μM
24	0.684057217	BNF 0.1 μM
24	0.730882299	BNF 0.1 μM
24	2.155646156	BNF 0.3 μM
24	2.998096302	BNF 0.3 μM
24	6.882963038	BNF 0.3 μM
24	7.118508328	BNF 0.3 μM
24	9.826610315	BNF 0.9 μM
24	9.640790055	BNF 0.9 μM
24	13.20528880	BNF 0.9 μM
24	15.66574736	BNF 0.9 μM
24	10,47750098	BNF 1 μM
24	15,75083724	BNF 1 μM
24	11,37043434	BNF 1 μM
24	11,46827036	BNF 1 μM

24	12.61086424	BNF 3 μM
24	16.82920780	BNF 3 μM
24	14.21727784	BNF 3 μM
24	14.37496398	BNF 3 μM
24	8.667026192	BNF 8 μM
24	12.63653731	BNF 8 μM
24	8.916652597	BNF 8 μM
24	11.58314181	BNF 8 μM
24	4.356204610	BNF 25 μM
24	7.245004896	BNF 25 μM
24	3.180976352	BNF 25 μM
24	5.051250052	BNF 25 μM

Table S3. Raw data from dose response experiment with nocodazole (NOC) alone on ethoxyresorufin-*O*-deethylase (EROD) activities at 6, 12, 24, 48 and 72h. Dimethylsulphoxide (DMSO) was used as vehicle-control. Three different doses of NOC was tested, 1, 10 1nd 25 μM. No significant induction of CYP1A activities could be measured in cells treated with 1, 10 or 25 μM NOC alone compared to that in vehicle-control treated (DMSO) cells and no comparison with fitted values were made for those treatments. The data are originally presented in a diagram as mean values ± standard deviation in Michelová M. Master-thesis, 2012 (reference 2).

Time (h)	**EROD activity** (pmol · (min · mg protein)$^{-1}$)	**Treatment**
6	0.11377082	DMSO
6	0.25665065	DMSO
6	0.16774007	DMSO
6	0.01560841	DMSO
12	0.11538702	DMSO
12	0.1016562	DMSO
12	0.2325335	DMSO
12	0.09802283	DMSO
24	-0.0063485	DMSO
24	0.05044173	DMSO
24	-0.0005397	DMSO
24	-0.0105922	DMSO
48	-0.0274927	DMSO
48	0.03330583	DMSO
48	0.05984807	DMSO
48	-0.0017851	DMSO
72	0.0029007	DMSO
72	0.06646949	DMSO
72	0.0272952	DMSO
72	0.0057347	DMSO

6	0.02561141	NOC 1 μM
6	0.03276828	NOC 1 μM
6	0.07516854	NOC 1 μM
6	0.24248221	NOC 1 μM
12	-0.0246521	NOC 1 μM
12	-0.0079321	NOC 1 μM
12	0.18523452	NOC 1 μM
12	0.18955516	NOC 1 μM
24	0.00743391	NOC 1 μM
24	-0.0256403	NOC 1 μM
24	-0.0031455	NOC 1 μM
24	0.03193376	NOC 1 μM
48	-0.0303212	NOC 1 μM
48	0.04061289	NOC 1 μM
48	0.07123973	NOC 1 μM
48	0.0127288	NOC 1 μM
72	0.0027803	NOC 1 μM
72	-0.0623571	NOC 1 μM
72	0.0249061	NOC 1 μM
72	0.0216266	NOC 1 μM
6	0.0763394	NOC 10 μM
6	0.04654934	NOC 10 μM
6	0.26674799	NOC 10 μM
6	0.06454803	NOC 10 μM
12	0.07035758	NOC 10 μM
12	0.05555987	NOC 10 μM
12	0.17346336	NOC 10 μM
12	0.20638625	NOC 10 μM
24	0.0681884	NOC 10 μM
24	-0.0734886	NOC 10 μM
24	0.14514875	NOC 10 μM
24	0.10303029	NOC 10 μM

48	0.16856512	NOC 10 µM
48	0.10049923	NOC 10 µM
48	0.6412272	NOC 10 µM
48	0.08100902	NOC 10 µM
72	0.29324048	NOC 10 µM
72	0.0263505	NOC 10 µM
72	0.0459174	NOC 10 µM
72	0.03089082	NOC 10 µM
6	0.00804636	NOC 25 µM
6	-0.0439197	NOC 25 µM
6	0.27521007	NOC 25 µM
6	0.15776044	NOC 25 µM
12	-0.1723871	NOC 25 µM
12	0.06644439	NOC 25 µM
12	0.12006505	NOC 25 µM
12	0.06413321	NOC 25 µM
24	-0.0644523	NOC 25 µM
24	0.00814383	NOC 25 µM
24	0.10558674	NOC 25 µM
24	0.01805869	NOC 25 µM
48	0.16261439	NOC 25 µM
48	0.09551795	NOC 25 µM
48	0.03861299	NOC 25 µM
48	0.00371416	NOC 25 µM
72	0.09099329	NOC 25 µM
72	0.15649775	NOC 25 µM
72	0.0309007	NOC 25 µM
72	0.06562328	NOC 25 µM

FIGURES:

Figure S1. Local Sensitivity Analysis. Effects of the ten parameters on the EROD activity of free CYP1A enzymes (E_f^{EROD}) as the model output. The model is described by equations (4), (5), (9)-(13) and the parameters are listed in Table 1. The diagram shows sensitivity of E_f^{EROD} to small perturbations in the nominal values of the ten parameters. It shows that k_D has the largest negative average effects on the E_f^{EROD} values whereas k_{OX} has the least positive average influence.

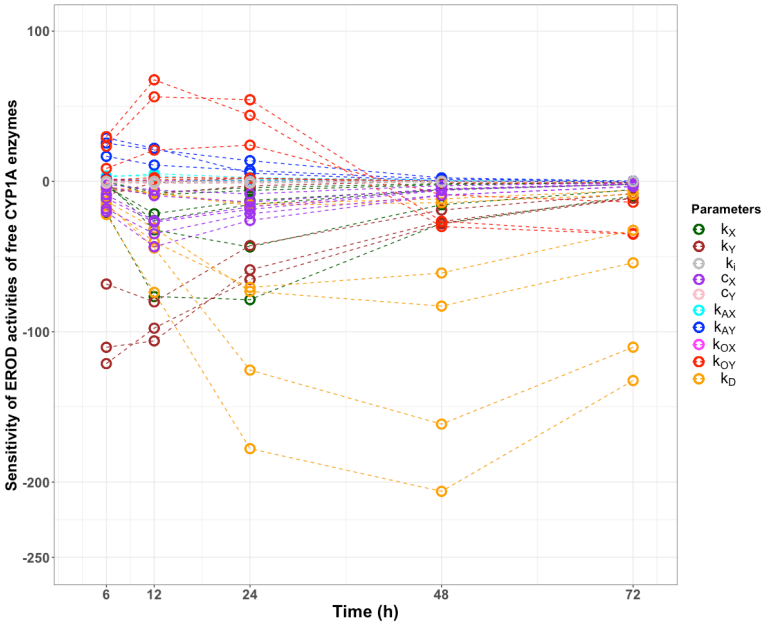
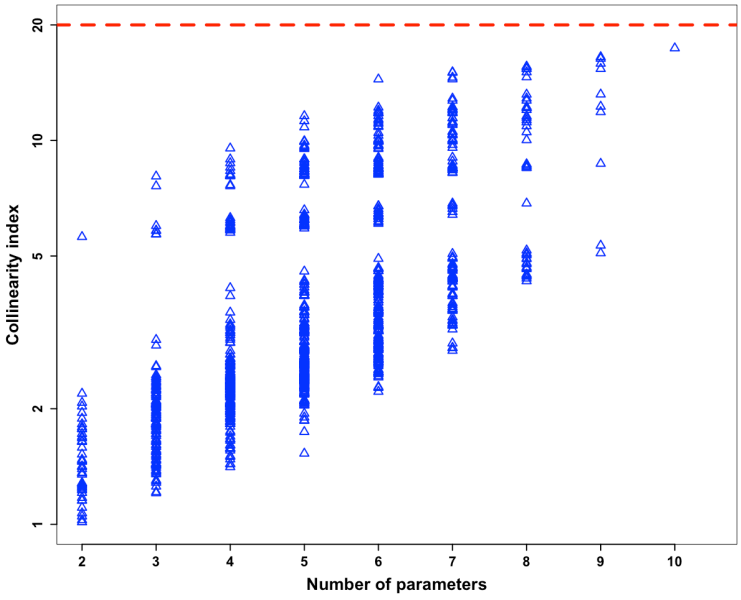


Figure S2. Collinearity index for all parameter combinations. The dashed red line denotes the threshold value of 20, for the collinearity index. All sets inholding different parameter combinations in the model have a collinearity index below 20.



REFERENCES

(1) Gräns, J.; Johansson, J.; Michelová. M.; Wassmur, B.; Norström, E.; Wallin, M.; Celander, M.C. Mixture effects between different azoles and β-naphthoflavone on the CYP1A biomarker in a fish cell line. *Aquat. Toxicol.* **2015.** 164, 43-51.

(2) Michelová, M. Cocktail effects of β-naphthoflavone and nocodazole in the *Poeciliopsis lucida* cell line. **2012.** Master-thesis. University of Gothenburg.

Paper D

A mathematical model of allosteric regulation in the cytosolic fatty acid synthesis pathway

Shirin Fallahi, Rune Kleppe, Anders Goksøyr, Guttorm Alendal
September 28, 2020

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no