Modelling of Metabolic Pathways for Biomolecule Production in Cell-Free Systems

Anamya Ajjolli Nagaraja

Modelling of Metabolic Pathways for Biomolecule Production in Cell-Free Systems

By

Anamya AJJOLLI NAGARAJA

Prepared at LE2P-*EnergyLab* and DSIMB

Doctoral Thesis Submitted to the University of La Réunion

Doctoral School of Science, Technology and Health

Speciality: Bioinformatics

Thesis to be presented on the 14th May 2020 in the presence of the jury panel composed of:

Prof. Sowdhamini Ramanathan, Professor, National Centre for Biological Sciences	**President**
Prof. Manuel Dauchez, Professor, University of Reims Champagne Ardennes	**Rapporteur**
Prof. Marie-Hélène Mucchielli-Giorgi, Professor, University of Paris-Saclay	**Rapporteur**
Dr. HDR. Cedric Damour, Associate Professor, University of la Réunion	**Examiner**
Prof. Brigitte Grondin-Perez, Professor, University of la Réunion	**Director**
Prof. Frédéric Cadet, Professor, University of la Réunion	**Co-director**
Dr. Philippe Charton, Associate Professor, University of la Réunion	**Co-supervisor**

Modélisation des Voies Métaboliques pour la Production de Biomolécules dans les Systèmes Acellulaires

par

Anamya AJJOLLI NAGARAJA

Préparée au LE2P-*EnergyLab* and DSIMB

Thèse de Doctorat Présentée à l'Université de la Réunion

Ecole Doctorale des Sciences Technologies et Santé

Discipline : Biologie Informatique

Thèse présentée le 14 mai 2020 devant le jury composé de:

Prof. Sowdhamini Ramanathan, Professeur, National Centre for Biological Sciences	**Présidente**
Prof. Manuel Dauchez, Professeur, Université de Reims Champagne Ardennes	**Rapporteur**
Prof. Marie-Hélène Mucchielli-Giorgi, Professeur, Université de Paris-Saclay	**Rapportrice**
Dr. HDR. Cedric Damour, Maître de conférences, Université de la Réunion	**Examinateur**
Prof. Brigitte Grondin-Perez, Professeur, Université de la Réunion	**Directrice**
Prof. Frédéric Cadet, Professeur, Université de la Réunion	**Co-directeur**
Dr. Philippe Charton, Maître de conférences, Université de la Réunion	**Co-superviseur**

Dedicated to my uncle Madhura Upadhya,

who nurtured my dreams

Acknowledgements

The list of people whom I must thank for helping me to complete this thesis is quite long and there are no words which can adequately express my gratitude to each and every one of them. Here is my best attempt to acknowledge those who made this Ph.D. journey successful.

First and foremost, I would like to thank Prof. Frédéric Cadet who has been a great mentor for my academic career. His constant urge for exploring and learning new things motivated me to do my best during the whole process of this thesis. When I was pessimistic about the results, he was uplifting and made me look into the results from new perspectives. His optimism not only influenced my research but also personal life. He has always been patient and supportive when my work moved slowly. He not only supported my research, but also took care of all administrative works before and after my arrival at the lab.

Dr. Phillippe Charton and Dr. Cédric Damour have strongly influenced my thinking and research. The regular interactions with them were enormous help throughout my study. At the end of every discussion, we were full of new ideas. All three of them constantly supported me to improve my research and presentation. There were moments when I was unable to convey myself, but they encouraged me to explain to myself better and was more than forgiving.

I heartily thank Prof. Brigitte Perez for her constant support throughout the thesis. I must thank my thesis committee members Prof. Bernard Offmann, Prof. Jean-Pierre Chabriat, Dr. Fabrice Gardebien and Dr. Cédric Damour for their constructive feedback and appreciations for my research.

I also had the great pleasure of working with Dr. Birgit Wiltschi and Ms. Alena Voit from ACIB - Austrian Centre of Industrial Biotechnology. They were very informative and forthcoming to help us in experiments.

I would like to convey my special regards to Dr. Kshitish Acharya, who inspired me for the research during my Master's degree. He believes in independent thinking and encouraged to discover personal interest than falling in the race with the current "hot topics" of research. I also like to express my sincere gratitude to Prof. Sowdhamini, who stimulated me with her multidisciplinary research interests and encouraged me to apply for the doctoral fellowship.

My sincere thanks to my thesis reviewers, Prof. Manuel Dauchez and Prof. Marie- Helene Mucchiell-Giorgi for their patience and persistence while reviewing the manuscript. I also thank

the rest of the jury Prof. Sowdhamini Ramanathan, Dr Cédric Damour, for their valuable insight during my thesis defence.

Most of the time, it is not straightforward when we decide to move to the new place with a vast difference in culture and language. This transition was comfortable for me because of Akhila, who helped me throughout the process of applying to this position to my arrival at to this Island. And later also, she had been caring and supportive. There are no words to express my gratitude to her. Moaz Adam, Gabriel Hoarau, Guillaume Beck and Ophélie Lo-Thong who were more than colleagues and I would not have survived my initial days without them. Without Guillaume and Ophélie it would have been challenging to manage stress. Many thanks to Anna, Salwa, Sumaiya and Gary Mares for all the memorable moments.

This Ph.D. did not allow me to grow only as a researcher but also a better person. In Réunion, I met many amazing people who were gentle, accommodating and helped me to explore a different culture and grow as an individual. My sincere thanks to the family of Ophélie, who introduced me to many others as their other daughter.

In life, we meet people unexpectedly, some influence more you than the other, some change your perspective about life. Sandee Moutoussamy was one of such people, who changed my views significantly. Unfortunately, he had to tolerate all my stress and anger outbursts more frequently than the others. But he has never been anything less than a supportive, understanding, forgiving friend and well-wisher. He introduced me to most wonderful people like Damayanti Velachy and Mme. Dolène Moutoussamy. My sincere thanks to Devi Moutoussamy, who has been a constant support during the COVID-19 crisis. I am indebted to Moutoussamy and Velachy families for their unconditional love.

Tedjee Hoarau was not only my driving teacher but also an awesome friend. He was the one with whom I could be crazy and adventurous. Thank you for believing in me, for the constant words of reassurance. And for successfully making me pass the licence exam.

I cannot begin to express my thanks to Vigna Chandrasekaran, who was my friend, sister, and most often a therapist. Lovida was more than a friend and due to her, I missed my home not much. Thank you both for always being there for me.

This Ph.D. and meeting all these wonderful people, exploring new places, and culture would be unachievable without my uncle Madhura Upadhya, and aunt Usha Upadhya. Both of them

supported my education and stood with me all the time. Without my uncle, it would have been even impossible to have a dream in the first place.

I am obliged to Teerna, who showered me with constant encouragement, love, and support. She was always there when I needed her. My sister from another mother, Neha who offered me valuable advice personally and professionally, never hesitated to be an excellent critic when it was required. I also wish to thank Yughandhar and Shafi for their support. The memories of NASTY's will be cherished forever. Special thanks to Sanjeet and Sowmya for encouraging me to take up academic research. I wish to express my deepest gratitude to my best friends in India- Sindhoora, Satvic, and Priyamedha for being constant during all ups and downs in my life.

Most importantly none of this could have been possible without my family's support. With all the troubles and sacrifices in life, my parents never forced me to choose the path anything less than what I wanted. Thanking them for everything they have done for me, sounds almost ridiculous. Even when, it was difficult for them to understand my work, they were supportive and encouraging. My sister and brother always believed and encouraged my dreams and convinced my parents to send me for a country so far away. And now in addition to siblings, an astonishing brother-in-law who is supporting the path I chose should be equally acknowledged. I owe them every little thing I achieved so far.

Table of Contents

ABSTRACT

Cell-free systems (CFS) are emerging as a powerful platform for biomanufacturing. The optimisation of the cell-free system is important to achieve maximum yield. The experimental optimisation is time-consuming and expensive. Different kinds of modelling emerged in the last decades, helping to optimise the pathway of interest in a shorter time at a low cost. In this study, we tested two approaches: systemic through the implementation of neural networks, and analytical through the use of differential equations. In the first step, an artificial neural network model was built to predict the flux through the pathway, and in the second step, a new methodology termed GC-ANN was developed to select optimum and cost-efficient enzyme balances for higher flux. This approach showed unexpected betterment of flux estimation, up to 63%. In the third step, a kinetic model was built and estimation of kinetic parameters for selected enzymes was achieved to replicate experimental conditions. Finally, linked to one of the most demanding chemicals, malate synthesis pathway was successfully modelled in the cell-free system. Even though many studies have been performed, biomanufacturing has not yet been possible for malate. The combination of the cell-free system and modelling could help achieve the biomanufacturing of malate. Overall, this thesis explores different mathematical modelling approaches, and their limits, for optimising metabolic pathways.

RESUME

Les systèmes acellulaires sont en train de devenir une puissante plateforme de biofabrication. L'optimisation de systèmes acellulaires est importante pour obtenir un rendement maximal. L'optimisation expérimentale, en laboratoire humide, est longue et coûteuse. Différents types de modélisations permettant d'optimiser la voie d'intérêt, en un temps plus court et à moindre coût, sont apparus au cours des dernières décennies. Dans cette étude, nous avons testé deux approches : systémique à travers la mise en œuvre de réseaux de neurones, et analytique à travers l'utilisation d'équations différentielles . Dans une première étape, un modèle à réseau de neurones artificiels a été construit pour prédire le flux de métabolites à travers la voie. Dans une seconde étape, une nouvelle méthodologie, appelée GC-ANN, a été développée pour sélectionner des équilibres enzymatiques optimaux, et rentables, pour des valeurs de flux plus élevées. Cette approche a permis une amélioration inattendue du flux, jusqu'à 63%, validée *in vitro*. Dans une troisième étape, un modèle cinétique a été construit, et l'estimation des paramètres cinétiques pour les enzymes sélectionnées a été réalisée, afin de reproduire les conditions expérimentales. Enfin, liée à l'un des produits chimiques les plus exigeants en termes de production, la voie de synthèse du malate a été modélisée avec succès dans un système acellulaires. Même si de nombreuses études ont été réalisées, la biofabrication a grande échelle n'est pas encore possible pour le malate. La combinaison du système acellulaire et de la modélisation pourrait aider à réaliser la bioproduction du malate. De manière plus générale, cette thèse explore différentes approches de modélisations mathématiques, et leurs limites, pour l'optimisation de voies métaboliques.

ABBREVIATIONS

CFS ---------- Cell-Free System

BRENDA ---------- BRaunschweig ENzyme Database

SBML ---------- Systems Biology Markup Language

COPASI ---------- COmplex PAthway Simulator

ANN ---------- Artificial Neural Network

RMSE ---------- Root Mean Square Error

R^2- ---------- Coefficient of determination

PCA ---------- Principal Component Analysis

3D ---------- 3-Dimensional

k_{cat} ---------- Catalytic Constant or Turnover Number

K_m ---------- Enzyme affinity constant or Michaelis-Menten Constant

K_{eq} ---------- Equilibrium Constant

K_i ---------- Inhibition Constant

List of Peer-Reviewed Publications

1. Ajjolli Nagaraja A, Fontaine N, Delsaut M, Charton P, Damour C, Offmann B, *et al.* (2019) *Flux prediction using artificial neural network (ANN) for the upper part of glycolysis.* PLoS ONE 14(5): e0216178. https://doi.org/10.1371/journal.pone.0216178 (**Chapter 3**)
2. Ajjolli Nagaraja, A., Charton, P., Cadet, X.F., Fontaine, N., Delsaut, M., Wiltschi, B., *et al.* (2020) *A Machine Learning Approach for Efficient Selection of Enzyme Concentrations and Its Application for Flux Optimization*. Catalysts: 10, 291. https://doi.org/10.3390/catal10030291 (**Chapter 4**)

Chapter 1 Introduction

With the evolution of technology and industrialisation, there has been increased utilisation of fossil fuel which led to the historical rise in atmospheric carbon dioxide (CO_2). Population growth and increasing deforestation has led to a decrease in natural CO_2 fixation which in turn has increased the global average atmospheric CO_2 to 405 ppm (https://www.iae.org/topics/climatechange). CO_2 is recognised as one of the major causes of global warming and finding ways to decrease the level of CO_2 is an emergency for the survival of life on this planet. Using CO_2 for the production of different chemical molecules is one of the ways to reduce the net CO_2 concentration in the atmosphere.

Currently, four major methods are utilised for the CO_2 fixation into different products-chemical, electrical, photochemical and biological methods (J. Shi *et al.*, 2015; Singh *et al.*, 2018). The carbon atom in CO_2 is in a higher oxidation state and therefore requires energy to be converted into other products. Hence, industrial-scale production is limited. The chemicals currently synthesised by CO_2 fixation are limited by market size and the electrical and photochemical fixation has not matured enough yet for large-scale production. The biological CO_2 fixation is greatly observed naturally in plants and autotrophic microorganisms for photosynthesis. The biological method is the most appealing approach, due to the mild condition and high yield (Mistry, Ganta, Chakrabarty, & Dutta, 2019). However, low solar energy utilisation efficiency limits photosynthesis reaction by the enzymes (Y. H. Percival Zhang, 2013). Numbers of autotrophs have been identified, characterised and engineered for the CO_2 fixation into chemical molecules, including ethanol (Dexter & Fu, 2009), lactic acid (Angermayr, Paszota, & Hellingwerf, 2012), isobutyraldehyde (Atsumi, Higashide, & Liao, 2009), 1,3-propanediol (Hirokawa, Maki, Tatsuke, & Hanai, 2016), *etc*.

Many of the biomolecules like organic acids (Alsaheb et al., 2015; Cherrington, Hinton, Mead, & Chopra, 1991), antibiotics (Awan *et al.*, 2017; Haris *et al.*, 2018; Weissman & Leadlay, 2005), bioethanol, *etc*. (Khattak *et al.*, 2014; Y. H. Percival Zhang, Sun, & Zhong, 2010) are used in the pharmaceutical and food industries, and as energy sources. Biomolecule production is attracting the attention of biologists and industries, due to the decrease in non-renewable resources and global warming (Yim *et al.*, 2011; Y. H. Percival Zhang, 2010). For decades, scientists have been successfully producing different chemical molecules through microbial fermentation by optimising the process (Xiulai Chen, Wang, Dong, Hu, & Liu, 2017; Lee et al., 2012; Martínez, Bolívar, & Escalante, 2015; Yim et al., 2011). Synthetic biology and systems biology helped obtain the highest yield of biomolecules from the source (Anderson, Islam, & Prather, 2018; Lee et al., 2012; Zeng et al., 2018).

In the following section, I discuss different fields of synthetic biology, systems biology, and the most commonly used databases and tools to study metabolic pathways.

1.1 Synthetic Biology

Synthetic biology is the interdisciplinary field of biology which uses engineering and chemistry for the synthesis of biomolecules in substantial quantity. Synthetic biology uses natural functional parts of the biological systems, and redesigns or reassembles the parts to maximise their potential. The most important goal of synthetic biology is the cost-effective production of chemicals, drugs, and fuels (Clomburg, Crumbley, & Gonzalez, 2017; Dudley, Karim, & Jewett, 2015; Yi Heng Percival Zhang, Sun, & Ma, 2017). Synthetic biology could be broadly classified into two types, *in vivo,* and *in vitro* as described hereafter.

1.1.1 *In vivo* Synthetic Biology

In vivo, synthetic biology or cell-based system utilises the living entities for production; the field has greatly developed with the advancement of DNA sequencing (Y. Chen, Banerjee, Mukhopadhyay, & Petzold, 2020; Choi *et al.*, 2019). Traditional methods such as gene knockouts, gene editing, or metabolic engineering methods help assemble the biosynthetic pathway. The expression of heterologous genes which helps channel the pathway in particular directions, substrate channelling where reactants are directed to the active site of enzymes (Wheeldon *et al.*, 2016; Y. H. Percival Zhang, 2011)), quorum sensing which is capable of sensing a signal and responding with the gene expression (Tan & Prather, 2017), enzyme engineering, *etc* lead to improved yield. *Escherichia coli* and *Saccharomyces cerevisiae* organisms which are mainly engineered to produce alcohols, alkanes, and alkenes (Fatma *et al.*, 2018; X. Song, Yu, & Zhu, 2016; Y. Zhang, Nielsen, & Liu, 2018). However, microbial biosynthesis has its disadvantages- cell toxicity, low productivity, and the possible coproduction of by-products, which requires complex and protracted product recovery processes (Lu, 2017). Also, the synthesis of some molecules on an industrial scale through microbial fermentation is not cost-effective due to expensive substrates.

1.1.2 *In vitro* Synthetic Biology

Nobel laureate Eduard Buchner had laid the foundation for the cell-free system (CFS) of biomolecule production by converting sugar into ethanol in 1897. *In vitro* synthetic biology or cell-free system, constructs the artificial synthetic pathway for the conversion of a substrate to the product outside the cell. These systems consist of minimised parts of the metabolic pathway

achieving high efficiency in production. The cell-free systems are the reconstruction of biological pathways and minimise the cellular process. *In vitro* system can be of two types, either cell lysate based, where the cells are lysed and used for the production (Schoborg, Hodgman, Anderson, & Jewett, 2014; Shrestha, Holland, & Bundy, 2012), or pure enzyme-based: a mixture of purified enzymes and cofactors are in the system to produce the desired product (Y. H.Percival Zhang, 2010). CFS has been successfully used in the synthesis of many products like bio-hydrogen (Fontaine, Grondin-Perez, Cadet, & Offmann, 2015; Xinhao Ye *et al.*, 2009), bio-ethanol (Khattak *et al.*, 2014; Yi Heng Percival Zhang, 2015), antibodies (Huang, Sheng, Xu, Zhu, & Cai, 2014), vaccines (Junhao Yang *et al.*, 2005), proteins (Lu, 2017), *etc*. The advantages of the cell-free system are: no by-product formation, high volumetric productivity, high product titre, high tolerance to toxicity, untroubled process control and optimisation, *etc*.

1.2 Computational Biology

The complexity of the biological system and the availability of a large amount of data lead to using computers in biology, and encourage computational biology to be developed as a multidisciplinary field. Computational biology developed fast in the past decades with the blend of computer science, applied mathematics, statistics, and engineering to understand biological problems. Computational biology is extensively used for data analysis, molecular modelling and prediction, and simulation, *etc*.

1.2.1 Study of Metabolic Pathways

The metabolic pathways are extensively studied using synthetic biology approaches. The cell-free systems partly eliminate the problem regarding the performance, compared to the cell-based system, by avoiding unnecessary reactions while protein engineering is successfully used to improve the performance of enzymes (Erb & Zarzycki, 2016; C. Li, Zhang, Wang, Wilson, & Yan, 2020). However, the implementation of experimental processes for the biotransformation can be lengthy and costly for production. *In silico* approach for optimising the process will require less time and cost.

The construction of the model of the metabolic network consists of four steps:

i. Identification of constituents of the system: To improve a metabolic network by *in silico* approach, first, it is necessary to conceptualise the system in a model that can reproduce the behaviour of the real system. The selection of constituents of the system is the first step in modelling. The selection of the system is with regards to the interest of metabolites or pathway. Even though this step looks simple, it is crucial and tough because, if important components are missing, then the final results will be compromised. Alternatively, if it contains too many components, the model could be overparameterized.
ii. Identification of topology and regulators: This step requires substantial prior knowledge of the pathway. The biochemical research over the past 100 years has documented information about many of the metabolic pathways. The machine learning or graph-based approaches require minimal prior knowledge, but experimental data are necessary.
iii. Choice of mathematical representation: Depending on the availability of data and knowledge of the system, one could choose the type of the model to represent the system.
iv. Parameter estimation: Parameter estimation depends on the type of the constructed model. When experimental data are available, parameter values are usually determined using inverse problem. Depending on the model chosen, the simultaneous estimation of many parameters can be a difficult task.

These steps are followed by an analysis of consistency, model sensitivity, and stability. Once the model is optimised and validated, then it can be used to identify the regulatory points of the system, in order to improve its behaviour. The metabolic pathway modelling can be classified into two broad categories:

i. Knowledge-based model: This approach requires detailed knowledge of the system and its elements. This approach is also known as an analytical method.
ii. Data-based model: The data-based model requires experimental data from the system of interests. In this approach, the whole system is studied as only one entity and at the macroscopic level.

The choice of an approach, depends on the aim and information available about the system. Whichever approach is used, the model must reproduce the experimental behaviour of the

given system. The accuracy of the model is variable, depending on the type and the complexity of the system.

1.2.1.1 Knowledge-Based Model

The emergence of genomics, transcriptomics and proteomics**,** along with improvements in information technology, helped us to integrate the information, build the mathematical *in silico* model of a biological system and observe its behaviour (Fukushima, Kusano, Redestig, Arita, & Saito, 2009; Nookaew, Olivares-Hernández, Bhumiratana, & Nielsen, 2011; Pereira *et al.*, 2018; Schilling *et al.*, 2002; Stelling, 2004). Researchers became more interested in the mathematical modelling of biological systems due to the availability of data from "-omics" studies (Stelling, 2004). The modelling helps to organise the system information, to simulate and, hence, to optimise the experiment and to understand the system characteristics. The integration of different "-omics" data helped understand the genetic differences between the phenotypes, identify the molecular signature (Acharjee, Kloosterman, Visser, & Maliepaard, 2016; Wheelock et al., 2013), and use metabolic engineering (G. Q. Chen, 2016; Vemuri & Aristidou, 2005), *etc*. There have been many attempts to model biological systems**,** like *Saccharomyces cerevisiae* (Duarte, Herrgård, & Palsson, 2004; Förster, Famili, Fu, Palsson, & Nielsen, 2003; Nookaew *et al.*, 2011; Price, Reed, & Palsson, 2004)*, Escherichia coli* (Feist *et al.*, 2007; Reed & Palsson, 2003; Weaver, Keseler, Mackie, Paulsen, & Karp, 2014), other organisms (Pereira *et al.*, 2018), and many plant metabolic networks for observing and predicting the behaviour of a system using different methods (Rios-Estepa & Lange, 2007; Stelling, 2004).

Many kinds of mathematical models exist to study biological systems (Arturo & Mora, 2016; Friedman & Kao, 2012). From the data and constraints used, the mathematical modelling can be classified into two broad categories (Rios-Estepa & Lange, 2007; Stelling, 2004) i.e., constraint-based or stoichiometric modelling (Covert, Famili, & Palsson, 2003; Price et al., 2004; Vijayakumar, Conway, Lió, & Angione, 2017)**,** and kinetic modelling or mechanistic modelling (Almquist, Cvijovic, Hatzimanikatis, Nielsen, & Jirstrand, 2014; Srinivasan, Cluett, & Mahadevan, 2015; Steuer, Gross, Selbig, & Blasius, 2006).

The metabolic network modelling, always requires the knowledge of stoichiometry. The stoichiometry defines numbers of reactants and products involved in the pathway, and how the network is connected. The stoichiometric metric is built to summarise the system and the stoichiometric coefficients are used to describe the reactants and the connection with the

network. A negative value is assigned for the reactants as they are consumed and positive for the product as it is formed. The rows of the matrix refer to metabolite and columns to reactions.

The following equilibrium equation defines the steady-state:

Equation 1.1: S is a vector of concentration of metabolites, N is the matrix of stoichiometric coefficients and v is a vector of reaction rates.

$$\frac{dS}{dt} = Nv = 0 \qquad (1.1)$$

The knowledge-based modeling methods can be further classified into two types:

1. Constraint-based model
2. Kinetic model

1.2.1.1.1 Constraint-Based Model

The constraint-based model uses physio-chemical constraints like mass balance, thermodynamic constraints, *etc*., in the modelling, to observe and study the behaviour of the system (Covert *et al.*, 2003). Different constraint-based methods have been developed to study the metabolic pathways**,** like flux balance analysis (Orth, Thiele, & Palsson, 2010) or metabolic flux analysis (Wiechert, 2001).

Flux balance analysis is an approach to study biochemical networks on a genomic scale, which includes all the known metabolite reactions, and the genes that encode for a particular enzyme. The data from genome annotation, or existing knowledge, are used to construct the network (Edwards & Palsson, 2000; Schilling et al., 2002), and the physicochemical constraints are used to predict the flux distribution, considering that the total product formed must be equal to the total substrate consumed in steady-state conditions (Orth *et al.*, 2010). This method is used to predict the growth rate (Edwards & Palsson, 2000; Jamshidi & Palsson, 2007; Orth et al., 2010; Schilling et al., 2002) or the production of a particular metabolite (Claudia, Quintero, & Ochoa, 2015). Flux balance analysis aims to solve Equation 1.1 to calculate the intracellular fluxes. (The review of flux balance analysis can be refered (K. Raman & Chandra, 2009), https://en.wikipedia.org/wiki/Flux_balance_analysis,

Metabolic flux analysis is an experiment based method and allows the quantification of metabolites in the central metabolism using the Carbon-labelled substrate (Xuewen Chen, Alonso, Allen, Reed, & Shachar-Hill, 2011; Christensen & Nielsen, 2000; Wiechert, 2001).

The labelled substrate is allowed to distribute over the metabolic network, and is measured using NMR (Albert A. de Graaf, Mattias Mahle, Michael Mollney, Wolfgang Wiechert, Peter Stahmann, 2000) or mass spectrometry (Orth *et al.*, 2010).

Elementary modes (Schuster, Fell, & Dandekar, 2000) and extreme pathway analysis (Schilling, Edwards, Letscher, & Palsson, 2000) use the concept of metabolic path. An elementary mode is a unique metabolic path within a network, in the stationary state. An elementary mode consists of a minimum number of reactions and forms a unit of the network. The elimination of a reaction from an elementary mode reduces its functionality.

The extreme pathways are sub-parts of the paths forming elementary modes. The determination of an end path, for extreme pathway analysis is through the same rules as an elementary mode, with two additional features. The internal reactions to the system are broken down into two reactions, one forward direction and the other- reversible direction. An end channel is unique, it cannot be represented by the combination of other preexisting extreme channels of the network.

Constraint-based methods have an advantage as they do not require kinetic information. However, the constraint-based method does not provide information about the concentration of metabolites.

1.2.1.1.2 Kinetic Model

The kinetic model defines the reaction mechanism in the system using kinetic parameters to evaluate rate laws. Kinetic rate law defines how one or more reactants, products and effectors interact, and how fast the reaction takes place. Kinetic modelling of pathways helps to better understand their behaviour and replicate the system. These rate laws are defined from the experiment, assuming that the experimental conditions are similar to *in-vivo* conditions (A Cornish-Bowden and C W Wharton., 1988). To build a kinetic model, the system is made as simple as possible, while retaining the system behaviour. Several kinetic laws have been described, and often used to study the mechanism:

i. Law of mass action: The law of mass action was determined in 1867 by the Norwegian chemists Cato Guldberg and Peter Waage. It associates the equilibrium concentrations of metabolites with a constant $K_{T,}$ which only depends on the temperature.

   Equation 1.2: Relationship between constant K_T and the concentration of metabolite.

$$K_T = \frac{\prod[Product]}{\prod[Substrate]} \quad (1.2)$$

The law of mass action is used to express the kinetics of chemical reactions, but remains insufficient to write biochemical reactions. Indeed, it does not take the behaviour of the enzyme into account, and it does not describe the possible enzyme-substrate interactions.

ii. Michaelis-Menten equation: The German Leonor Michaelis and the Canadian Maud Menten in 1913 defined Michaelis-Menten's law. This is specific to enzymatic reactions which establish a link between the speed of the reaction with the concentration of substrates, and the kinetic parameters of the enzyme used. The rate of reaction *v,* for an irreversible enzyme which converts the substrate to product is:

S -> P

Equation 1.3: Michaelis-Menten rate equation for irreversible conversion of substrate S to product P.

$$v = \frac{V_{max} * [S]}{K_m + [S]} \quad (1.3)$$

Where, $V_{max} = k_{cat} * [E]$

k_{cat} is the catalytic constant of an enzyme, which refers to the maximum number of substrate converted to product per unit of time (usually time in seconds). [E] and [S] are the concentration of the enzyme and substrate. K_m is Michaelis-Menten constant which describes the affinity of an enzyme for the substrate.

The dynamic model, (i.e, kinetic model) allows a better-quality reproduction of the system in terms of precision, than the static models like constraint-based models. However, it requires the knowledge of the kinetic parameters of each enzyme. The modelling of enzymes like phosphofructokinase, which is involved in the glycolysis pathway, can be problematic and, might need more parameters than other enzymes (Teusink *et al.*, 2000). Determining the kinetic parameter is expensive and time-consuming (Hakenberg, Schmeier, Kowald, Klipp, & Leser, 2004); some parameters can be more difficult to measure (Bisswanger, 2014). Although many enzymatic assays are described in the literature, sometimes it is necessary to modify the assay for new enzymes, or to find a new one. Following enzyme reaction through spectrophotometer or spectrofluorimeter is difficult, due to no absorption or emission signals (Bisswanger, 2014) linked to the reactants. Most of the available kinetic data are obtained from *in vitro* studies

using purified enzymes, which might not represent the exact properties of *in vivo* enzymes (Steuer *et al.*, 2006). For example, the V_{max} value measured *in vitro,* may not represent the value of an *in vivo* system, due to the destruction of enzyme complexes, cellular organisation and the absence of an unknown inhibitor or activator (Albe & Wright, 1992; Wright, Butler, & Albe, 1992).

1.2.1.2 Data-Based Model

The data-based approach is not about analysing each elementary entity of the system, but aggregating the whole into functional sub-parts. Interactions between subparts and the environment around the system are to be taken into account in the modelling, and only interaction between subparts and the environment is considered, excluding the internal detail.

The input and output variables of the system are the only known and described variables whereas the internal entities and the mechanisms involved remain undetermined. This approach requires experimental data to build a model. The data-based modelling approach consists of two stages:

i. Learning phase: From the experimental data, the model parameters are estimated to reproduce the similar behaviour of the data. The model represents relationships and logical links between the conditions and the behaviours of the system, represented by the outputs.
ii. Validation phase: Using the established model from the learning phase, predict the behaviour of the system. The predicted behaviour is compared with experimental behaviour to assess the accuracy of the model. If the model has an acceptable accuracy, it can be used to predict the behaviour of unknown conditions.

Several data-based methods are used in different fields of science. These methods are developed on the basis of the learning and validation phase. The neural network is one of such data-based modelling methodologies used extensively in different disciplines.

1.2.1.2.1 Artificial Neural Network

The biological neural system inspired the artificial neural network (ANN) model structure. Neurologists developed and designed the first neural networks during the years 1940-1950 (Lettvin, Maturana, Mcculloch, & Pitts, 1959; Mcculloch & Pitts, 1990). The main aim of ANN is to mimic the human brain for its ability to process the information and acquire knowledge.

The neural network is an architecture, modelled on the brain**,** organised with neurons and synapses present in a structure of nodes (formal neuron) connected together (Hassoun, 1996; Jain, Mao, & Mohiuddin, 1996). The neuron is capable of receiving one or more signals (input) from the other neurons. It sums the information, analyses and processes the summed information and sends a response signal (output). By analogy to biological neuron, the formal neuron performs upon receipt of the input signals into a weighted, sum and processes this weighted sum by an activation function. If the threshold of the signal is reached, then the response will be transmitted.

Each signal as the input of a neuron is associated with a weighting coefficient to give different weights to the information arriving at the computation cell. Signals from output go to the input of other formal neurons or to the outside of the system.

Each numerical input corresponds to the input layer, and the value to predict (variable to explain) corresponds to the last level, the output layer. Between those two layers, intermediary nodes are present, built specifically and in sufficient numbers to model the problem; they form the hidden layer. The architecture of ANN can be summed up as in Figure 1.1:

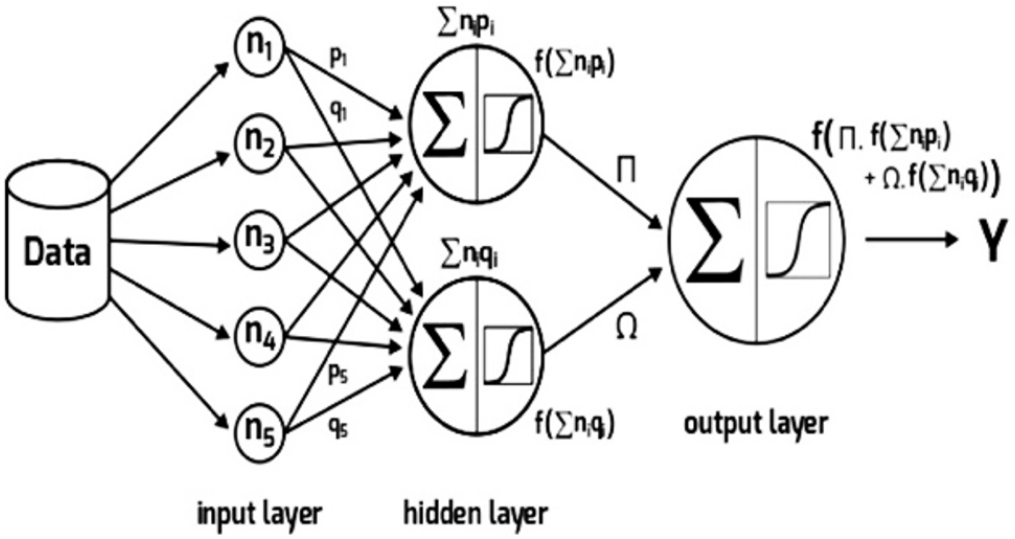


Figure 1.1: Assembly of a neural network in three layers. Information goes from the input cells to the output layer. pi represents the weighting coefficients of the signals coming from the neurons n1, n2..n5 going towards the hidden neuron 1. q_i represents the weighting coefficients of the signals coming from

the neurons $n_{1,}$ $n_{2..}n_5$ going towards the hidden neuron 2. □ and Ω is the weighting coefficients for the signal between the hidden layer and the output layer.

i. Input layer: The first layer is made up of a set of neurons, which are receivers of input signals, coming from outside, picked up by the system. Each signal activating the system is materialised by an input cell.

ii. Hidden layer: The second layer has a variable number of neurons. Each of these neurons receives signals from all the neurons in the input layer. These neurons process statistical information received. The weighting coefficients need to be estimated allowing the complex function, which represents the entire neural network, to reproduce the output signal. The number of hidden neurons used by the system is to be determined by a trial/error process to retain the one that most closely reproduces the actual behaviour of the system. The activation function calculates the weighted sum of inputs, add bias.

iii. Output layer: The third layer receives and processes all the information generated by the hidden layer to reproduce the behaviour specific to the system in response to the conditions encountered. A behaviour can be associated with several output signals. Each output signal from the system is materialised by an output neuron. It is possible to identify a neural network with several output neurons.

The neural network has been successfully applied in different fields of science including physics (Giuseppe Carleo, 2017; Kolanoski, 1995), environmental science (Ahmed Gamal El-Din, Daniel W. Smith, 2004; Liu ZeLin, Peng ChangHui, Xiang WenHua, Tian DaLun, Deng XiangWen, 2010; Pawul & Śliwka, 2016) and data mining (Eide, Johansson, Lindblad, & Lindsey, 1997; Kamruzzaman & Jehad Sarkar, 2011), for the prediction of different features in the system. The ANN is also core for deep learning (Schmidhuber, 2015). The artificial neural network can be used to predict the product outcome (i.e. flux through the pathway) when combined with flux balance analysis or other modelling approaches. In particular, the ANN has been used for the selection of optimised medium components in the fermentation process for producing different molecules such as lipids from *Chlorella vulgaris* (Morowvat & Ghasemi, 2016) and Spinosyns from *Saccharopolyspora spinose* (Lan, Zhao, Guo, Guan, & Zhang, 2015). ANN was employed, for instance, for the prediction of the flux through mammalian gluconeogenesis, using the simulated data from metabolite isotopic labelling (Antoniewicz, Stephanopoulos, & Kelleher, 2006).

The data-based model, like a neural network, is generally less precise than a dynamic analytical model, due to the abstraction of the internal behaviour of the system. However, it is easier to build because it requires only inputs and outputs. The property of the model depends on the time and the situations covered by the learning base used.

1.3 Databases and Tools Used in the Modelling

Analytical approach is the type of modelling most used by biochemists in the study of the metabolome. It only needs to determine the greatest number of molecular mechanisms present, which are the basic building blocks of the metabolic system. *In silico* reconstruction of metabolic networks makes it possible to identify and to analyse the molecular mechanisms involved in a physiological state of an organism. The first phase consists of retrieving the metabolic information of the reagents, enzymes and reactions involved in a metabolic pathway (Francke *et al.* (2005)). This phase requires the collection of data from publications and biological databases gathering related information to the metabolic pathways and genes, proteins, reactions involved. Many databases allow the collection of these data, examples:

i. The KEGG database (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa & Goto, 2000), is one of the first databases, on the internet, gathering information on genes, proteins, reactions and metabolic pathways from the sequencing of a large panel of organisms.
ii. The BRENDA database (BRaunschweig ENzyme Database) (Schomburg, Chang, & Schomburg, 2002) contains comprehensive enzymatic and metabolic data from various experiments updated regularly. It describes the biochemical property of each enzyme.
i. SABIO-RK database ((http://sabio.h-its.org/, (Wittig *et al.*, 2012)) contains comprehensive information about biochemical reaction and their kinetic property.

Several tools have been developed to model and to simulate the biological networks and process. To improve the compatibility of these tools, markup languages such as SBML (Systems Biology Markup Language, (Hucka et al., 2003) have been created. SBML is a simple XML-based software-independent language for representing biochemical reaction networks. The following are the tools used in this thesis for biochemical network analysis:

i. CellDesigner (Funahashi, Morohashi, Kitano, & Tanimura, 2003) is a process diagram editor for the gene-regulatory and biochemical network using SBML. It helps in modelling and simulating networks.

ii. COPASI (COmplex PAthway Simulator) (Hoops et al., 2006) is a biochemical simulator with the integration of diverse numerical methods to analyse the biochemical pathway.

1.4 Aim and Scope of the Thesis

Human activities have elevated the global atmospheric carbon dioxide and there is an urgent need to reduce the CO_2 emission and to develop the strategy to reduce atmospheric CO_2. One of the strategies is to convert CO_2 into biochemicals. The biological methods are more favourable with the mild reaction conditions and are easy to handle. The main goal of this thesis is to model the biological pathway which helps in fixing atmospheric CO_2 to a chemical molecule. In **Chapter 1**, I have discussed different methods to study the metabolic pathways. Recently, the cell-free systems (CFS) gained its attention to produce different biochemicals. CFS proved its efficiency in producing proteins, therapeutics and insulin. In this thesis, four studies are presented. For each one, I address different issues of metabolic engineering using the different types of modelling approaches to study the cell-free systems.

In **Chapter 2**, the applications of synthetic biology, modelling approaches, different concepts and differences between differet methodologies chosen are described breifly.

The glycolysis is a central carbon metabolism pathway, studied in different aspects such as diseases, bioprocessing, *etc*. Several biomolecules are produced by optimising the glycolysis pathway and the availability of experimental data provided the opportunity to use data-based modelling approach. Many methods are developed to determine or to estimate the flux through the pathways. But these methods require many parameters or constraints which can be expensive and time-consuming to determine. Hence, an artificial neural network (ANN) based method is developed in predicting flux using existing experimental data, and is discussed in **Chapter 3**.

Furthermore, the selection of optimum enzymes balance for the higher product is one of the challenges in bioprocessing. In **Chapter 4**, a new methodology for selecting optimum enzyme concentration in part glycolysis is explained. The developed methodology is based on the flux estimation through the pathway using ANN. ANN is known to be inefficient in extrapolating predictions outside the box: high predicted values will bump into a sort of “glass ceiling”. However, by careful selection of the enzyme balances from glass-ceiling space could yield better flux. The newly developed approach (termed GC-ANN) also helps us to select the economic enzyme balances with higher flux values.

ANN is a training-based model, the prediction of new data depends on the dataset used for the training of the model. The training dataset used to train the model in **Chapter 3** and **Chapter**

4 is small (121). Obtaining larger dataset by experiments is expensive and tedious. Simulating data, using the kinetic model can be a convenient choice to reduce the experiments. In **Chapter 5**, the kinetic model of the upper part of glycolysis is built using the experimentally measured kinetic parameter. And, we attempted to fit the model with experimental data in order to simulate the flux through the pathway. The main goal of **Chapter 5** is to perform *in silico* replication of the experimental condition, and by doing so, to obtain larger datasets to enhance the learning by ANN models.

Recently, i*n vitro* system has emerged as a platform for biomolecule production with advantages such as no byproduct formation, no cell toxicity due to substrate/product. Indeed, producing the biochemicals from cell-based fermentation emits CO_2 by respiration during the process. Global warming is the red flag to find ways to reduce the CO_2 emission to the environment. Malic acid is one of the chemicals which can utilise CO_2 during the synthesis with high demand in food, beverage and chemical industries. Currently, 40,000 tons of malate is produced annually whereas the global demand is 200,000 tons(Chi, Wang, Wang, Khan, & Chi, 2016). Malic acid is defined as one of the building blocks chemicals by the US department of energy (Werpy & Petersen, 2004). Currently, this dicarboxylic acid is mainly synthesised from chemical processes using the non-renewable resources. The depletion of these resources and environmental concerns encouraged to develop a green method for the synthesis. The detailed review of the synthesis of malic acid is explained in **Chapter 6**. Even after extensive studies of biosynthesis of malic acid using different microorganisms, sources and techniques, there is no successful method for the industrial-scale production using microorganisms. This provides an opportunity to select the malic acid as the molecule of interest in our study.

In this thesis, the synthetic pathway for malic acid synthesis is reconstructed. *In vitro* systems using the thermophilic enzyme for synthesising malic acid are getting attention in recent years. To optimise the malic acid synthesis system using thermophilic enzymes, the kinetic model is built and studied in **Chapter 7**.

Chapter 2 State of Art and Concepts Used in the Thesis

2.1 Synthetic Biology Techniques and Application

1911 by French biophysicist Stéphane Leduc laid the foundation for synthetic biology (Leduc, S. 1912. *La Biologie Synthétique*. A. Poinat, Paris.). Synthetic biology aims to engineer the biological system with the predictable behaviour of human needs. The long-term goal is to breakdown the complex system to interchangeable parts that can perform different functions. Synthetic biology gave rise to the development of many technologies that are implemented using the model microbial species *Escherichia coli* and *Saccharomyces cerevisiae* which includes biosensors that are capable of sensing the broad range of bioanalytes and responding with regulated expression, the engines for the production of biochemicals or performing complex logical functions.

2.1.1 Engineering Synthetic Parts

The proteins, RNA, and DNA are extracted from the natural producer and transferred to the host to develop a new system with a novel function. This feature is used to develop new pathways *in vivo* and *in vitro,* or by incorporating unnatural amino acid.

2.1.1.1 Genetic Circuit Design

The cells communicate and exhibit complex patterning by responding to signal with gene expression. Gene switches works at the different stage of central dogma molecular biology as signal responses and helps cells to adapt to the environment (Ausländer & Fussenegger, 2013; Bradley, Buck, & Wang, 2016; Brophy & Voigt, 2014). A typical gene switch consists of a sensor that detects the input signal and the regulatory unit controls the gene expression which is the output. Gene switches respond with different chemicals as a signal example, antibiotics (Fussenegger et al., 2000; Weber et al., 2002), metabolites, proteins, or physical signals such as light and temperature.

2.1.1.2 Protein Switches

Allostery is a feature of protein, where the binding of one ligand leads to the structural rearrangement and influences the other protein domain (Goodey & Benkovic, 2008; S. Raman, Taylor, Genuth, Fields, & Church, 2014). The general architecture of protein switches consists of a sensor domain that binds to ligands, which is coupled to a functional domain in a way that allows for its allosteric regulation. For example, the intracellular calcium receptors are fused with the two fluorescent proteins to readout the signal (Miyawaki et al., 1997).

2.1.1.3 Expansion of Genetic Code

By incorporating unnatural nucleic acids, improve the functionalities, and improves diversity. Since the modified nucleic acids are not targeted by the nucleases is promising therapeutic agents (Vater & Klussmann, 2015). The unnatural amino acids are incorporated in the proteins and site-specific introduction of protein modifications also used in therapeutic agents (Wals & Ovaa, 2014).

2.1.2 Genome Engineering

2.1.2.1 Site-Specific Genome-Editing

The DNA binding protein recombinases binds to a specific region of DNA and perform the alterations depending on the DNA recognition sites. The zinc finger nucleases(Kim & Chandrasegaran, 1994) and transcription activator-like effector nucleases (Boch et al., 2009; Moscou & Bogdanove, 2009) are programmed to bind to DNA sites in proximity within the genome and with endonuclease lead to the double-stranded breakage in the genome.

2.1.3 Application of Synthetic Biology

2.1.3.1 Metabolic Engineering

Metabolic engineering aims to convert the host into powerful factories which can use feedstocks to produce metabolites. The novel pathways are designed both *in vivo* and *in vitro* by incorporating the parts from different organisms. The variety of secondary metabolites are produced by the natural host are engineered for the industrial scale. The anticancer compounds, taxadiene (Ajikumar et al., 2010), and noscapine (Y. Li & Smolke, 2016) are produced successfully. The in-vitro systems are advanced recently for the production of different metabolites due to the disadvantages of the microbial system as discussed in the introduction chapter.

2.1.3.2 Biocomputing

The cellular biocomputing systems sense various input molecules and respond with a biological output which makes it good candidates for the control-systems in therapeutic, diagnostic. RNA-based gene switches used for the construction of logical gates in yeast and bacterial (C. C. Liu et al., 2012; Win & Smolke, 2008).

2.1.3.3 Other Applications of Synthetic Biology

The cell-free technology is used in diagnosis, for example, the capability of transcriptional factors to detect environmental signals which can serve as biosensors when added to the cell extracts used for the diagnosis of viral RNAs in diagnosis example Ebola (Pardee et al., 2014)- and Zika-specific (Pardee et al., 2016) sequences. The components that can be freeze-dried on a paper can be activated at the time of need.

The detailed review of synthetic biology techniques and application can be found in many reviews (Cameron, Bashor, & Collins, 2014; Dudley et al., 2015; El Karoui, Hoyos-Flight, & Fletcher, 2019; Keasling, 2012; Lu, 2017).

2.2 Modelling of the Biological System

A mathematical model describes the behaviour of the system and can be used for different purposes, such as predict the future behaviour of the system, estimate unmeasurable variables, used in a model-based control strategy, determine the optimal operating conditions, etc. The modelling is considered Succesful when the model, i. is accurate in representing the existing behaviour, ii. Should be able to predict the behaviour which is not already observed, iii. when the model can be used in another condition and iv. The model should be simple.

The modelling methods involve different steps that are described in Chapter 1 section Study of Metabolic Pathways**.**

Typically model consists of several components as bellow:

i. Variables: one variable in ordinary differential equations (ODEs), or more than one variable as in partial differential equation (PDEs) example time t and space (x,y,z).
ii. Unknown functions : which depends on the time example concentration of enzyme, substrate, and product {[E](t), [S](t), [P](t) respectively}.
iii. Parameters: This can be varied under the experimental condition which leads to the change in the system behaviour.
iv. Constants: which are fixed values, for example, Avogadro constant.

2.2.1 Type Of The Model

Deterministic vs Stochastic Model: in the deterministic model, the variable, parameter, and constants do not contain the randomness and are defined by a unique function. In the stochastic model, the variable, parameter, and constants contain the randomness, described by the probabilities.

The detailed description of types of models, differential equations used in systems biology can be found in the notes (Kuttler, 2009; Y. Zheng & Sriram, 2010) (Kuttler, 2009; Y. Zheng & Sriram, 2010), http://www.thep.lu.se/~henrik/bnf079/literature.html, http://www-m6.ma.tum.de/~kuttler/script1.pdf, http://www.sontaglab.org/FTPDIR/systems_biology_notes.pdf.).

2.3 Parameterisation Of Model

Parameter estimation aims to identify the model parameter which fits best the model to experimental behaviour of the system. The parameter to be estimated depends on the type of model constructed. For example, in the ANN model, the learning rate and weights are to be estimated based on the experimental data whereas in the kinetic model, the unknow kinetic parameters need to be calculated. In this section, I briefly discuss different methods mainly in the kinetic modelling of metabolic pathways. (The most of the information are taken from the review (Chou & Voit, 2009)).

2.3.1 Parameter Estimation (PE) of Kinetic Models

The measurements of kinetic parameters from individual experiments help to build the kinetic model of the system. The kinetic parameters can be obtained from databases such as BRENDA (Schomburg et al., 2002) or SABIO-RK (Wittig et al., 2012). If the measurement of the kinetic parameter is not available, it can be estimated using the experimental data, and estimation methods that use constraints from physical, chemical, or thermodynamic conditions to obtain the unique values (Chakrabarti, Miskovic, Soh, & Hatzimanikatis, 2013).

The parameters from the individual experiments might not be efficient enough to represent the real behaviour of the whole system because the parameter measurements are taken from different laboratories using different in-vitro conditions. The parameter estimation can be performed for the kinetic model either by simultaneous estimation of all the parameters in the model to experimental data or one by one parameter. The type of data required for the different

methods of estimation is different and based on the availability of data different approaches or combinations of approaches can be implemented to obtain a better model. The optimisation algorithms are used for parameter estimation which searches the large space of possible values under the constraints to search global optimum in a feasible time.

The currently available PE methods can be classified into the following classes:

1. Forward or bottom-up modelling
2. Using steady-state data
3. Inverse or top-down modelling

2.3.1.1 Forward or bottom-up modelling

In the bottom-up approach, the kinetic model was built using the individual parameters of the enzymes involved. The individual enzymes are purified, characterised, and studied and identified the parameter and represented in the mathematical rate laws and combining the information from rate law to build the mathematical model. The databases such as KEGG (Kanehisa & Goto, 2000), MetaCyc, and BRENDA (Schomburg et al., 2002) help to choose the topology of the pathway and kinetic information. The forward approach can lead to the qualitative representation of the system. The outline of this method involves, the representation of the model and estimating the parameters, then test the model behaviour and if required perform the refinement of the model structure and the parameter.

Even though this method looks simple and straightforward, the main disadvantage is the requirement of local kinetic parameters. And most of the time, the available parameters are from different organisms, experimental conditions therefore model might not represent the biological behaviour. This method can be laborious as it requires a series of refinements of the model.

2.3.1.2 Using steady-state data

The parameter estimation from the steady-state has been studied using stoichiometry and the flux. The estimation of the parameter from steady-state is by observing how the system behaves with small perturbations around the steady-state. Parameter values are obtained by changing the variable directly and measuring the flux. However, the data from these steady-state can be noisy and consists mainly few measurements (Chou & Voit, 2009).

2.3.1.3 Inverse or top-down modelling

The modern high throughput technologies such as include nuclear magnetic resonance (NMR), mass spectrometry (MS), high-performance liquid chromatography (HPLC), and flow cytometry helps to measure the concentration of the metabolite in the sequential points in time. And this data could be used for the modelling approaches and it is named the "top-down" or "inverse" approach of modelling. The advantage of inverse modelling over forward modelling is that the data can be obtained from the same organism with the same experimental condition, however, the data because of the complexity and non-linearity of the biological system (Chou & Voit, 2009).

2.4 Data-Based Modelling

The data-based modelling (DBM) requires the experimental data to study the behaviour of the model at the macroscopic level. Based on the data used, DBM can be classified as supervised and unsupervised modelling. DBM approach provides a faster predictive model. The various methods of DBM are highly used in biology and metabolic pathway analysis (Carbonell, Radivojevic, & García Martín, 2019; Cuperlovic-Culf, 2018; Mishra, Kumar, & Mukhtar, 2019; Zampieri, Vijayakumar, Yaneske, & Angione, 2019). Example, to study the enzyme turn-over number (Heckmann et al., 2018), to predict the model dynamics (Costello & Martin, 2018), Pathway prediction (Joseph M Dale, Liviu Popescu, 2010), Identify the essential genes(Plaimas et al., 2008).

2.4.1 Supervised Modellling

Supervised modelling uses the labelled data to learn from training data and predict the features for test data i.e, supervised modelling algorithms are designed to learn from the experience (https://en.wikipedia.org/wiki/Supervised_learning, https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590). The supervised learning algorithm can be written as Equation 2.1.

Equation 2.1 The learning function (f) maps input (X) to output (Y). Where Y is output and X is input variables.

$$Y = f(X)$$

Supervised learning can be classified into two types,

1. Classification: the model will be trained to classify data into different categories. The most common algorithm for classifications is Support Vector Machines, Decision Trees, K-Nearest Neighbor, Random Forest, etc.
2. Regression: the model will be trained to find the relationship between the dependent and independent variables. Example: linear regression, neural networks, etc.

2.4.2 Unsupervised Modelling

Unsupervised modelling inferences are drawn from the data consists of unlabelled input data. In unsupervised modelling, no training will be provided to the model, therefore the model is restricted to find the hidden pattern within the data (https://en.wikipedia.org/wiki/Unsupervised_learning, https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a, https://www.guru99.com/unsupervised-machine-learning.html).

Unsupervised modelling can be classified into two class

1. Clustering: the inherent groupings in the data are identified. Example: Hierarchical clustering, k-Means clustering.
2. Association: it allows establishing the association amongst the data object inside the dataset.

The main differences between supervised and unsupervised modeling are given in Table 2.1.

Table 2.1: The difference between supervised and unsupervised modelling.

Features	**Supervised learning**	**Unsupervised learning**
Input data	Uses known and labeled data as input	Uses unknown data as input
Computational complexity	Very Complex	Less Computational Complexity
Real-time	Uses off-line analysis	Uses real-time analysis of data
Number of classes	Number of Classes are known	Number of Classes are not known
Accuracy of results	Accurate and Reliable Results	Moderate Accurate and Reliable Results

2.5 Modelling Concepts and Difference Between the Approaches

In this section, I describe few basic concepts involved in different algorithms used in the thesis.

2.5.1 Artificial Neural Network

2.5.1.1 Feed Forward Vs Backpropagation

In Feedforward neural networks (https://en.wikipedia.org/wiki/Feedforward_neural_network, https://medium.com/machine-learning-for-li/explain-feedforward-and-backpropagation-b8cdd25dcc2f), the connections pass from the input layer to the hidden layer and then to the output layer and do not form a circle. Whereas in the backpropagation (https://en.wikipedia.org/wiki/Backpropagation, https://medium.com/machine-learning-for-li/explain-feedforward-and-backpropagation-b8cdd25dcc2f) the signal is passed in the feed forward, the error is computed and then propagate back to the earlier layer.

2.5.1.2 Importance of Order of Training Data in ANN

1. The order of data is important in ANN during the training phase. learning phase in the ANN model is not deterministic, so if we change the order of data, the output will differ during training. Setting seed, we assume that this change in the output is very low.
2. The ANN used in this thesis is not a recurrent neural network so it does not contain a dimension of time. So if we change the order of data, we can expect a slight change in the outcome during the training. Once the model is trained, it is deterministic: for a given input, the output is always the same.
3. Once the model is trained, it does not change the outcome of the test data.

2.5.2 Decision Tree Vs Random Forest

Decision tree	Random forest
Is built using the entire dataset	Several trees are built using the different subgroups of data, and each tree is ranked for each class
Is a simple method to explain	-
We know what variable/value is used to make the decision	It is a black-box approach, what variable/value used to make the decision is not known
-	Has high accuracy due to ranking

Chapter 3 Artificial Neural Network in Flux Prediction

3.1 Context

Glycolysis is the centre of the metabolic system in all living organisms. It is an anaerobic pathway present in almost all living cells and also helps in ATP generation. Glycolysis is established as the central core for fermentation**.** It contributes to the production of different metabolites**,** like citric acid, succinic acid, amino acids, *etc*., through pyruvate, the end product of glycolysis (J. Liu et al., 2017).

The ANNs were used earlier in predicting the fluxes from ^{13}C labelling of metabolites in mammalian gluconeogenesis by M.R. Antoniewicz *et al.* (Antoniewicz *et al.*, 2006). Three linear regression modelling methods, multiple linear regression (MLR), principal component regression (PCR) and partial least square regression (PLS) were performed on simulated data and compared to the ANN. The study showed that ANN, which requires the larger sample (>200), performed better with R^2 0.95 than the other methods (R^2 of 0.7) for flux prediction using new mass isotopomer data (Antoniewicz *et al.*, 2006).

Several approaches were developed to determine or estimate the flux through the metabolic pathway (Antoniewicz *et al.*, 2006; Fiévet, Dillmann, Curien, & de Vienne, 2006; Nikoloski, Perez-Storey, & Sweetlove, 2015) as explained in section Knowledge-Based Model in Introduction chapter. Due to the challenges in estimating the flux using different methods like constraint-based and kinetic-based modelling approaches (Allen, Libourel, & Shachar-Hill, 2009; K. Raman & Chandra, 2009; Rohwer, 2012; Vasilakou et al., 2016), a simple method using artificial neural networks was developed. This method is based purely on the existing experimental data, hence does not require kinetic parameters as in kinetic modelling and no prior information is required regarding the stoichiometry of the metabolic pathway. In this study, an artificial neural network was built to estimate the flux using enzyme concentrations for the upper part of glycolysis as input. Finding the optimum enzyme concentration, which gives the highest product through experiments is laborious and expensive. The neural network approach can help in choosing the optimum enzyme concentrations, which enhances the final product concentration without experimental setup, within a short period. Experiments are carried out to *i)* assess the structure of the ANN using three different approaches, *ii)* evaluate different activation functions, and iii) compare the prediction of flux of NADH to the fluxes predicted by Fiévet *et al.* (2006)

3.2 Materials and Methods

3.2.1 Principle of Artificial Neural Networks

The base element of ANN is the perceptron, defined in 1958 by Rosenblatt (Rosenblatt, 1958). A combination function computes a value from the input layer and some weight. This is a weighted sum $\sum n_i p_i$ (observed node) of the n_i values in the input layer. To define the output value, a function called activation function is applied to this value. The n_i is the node i, the weight p_i corresponds to the connection between node i, the observed node and the activation function f, associated with the observed node (Figure 3.1).

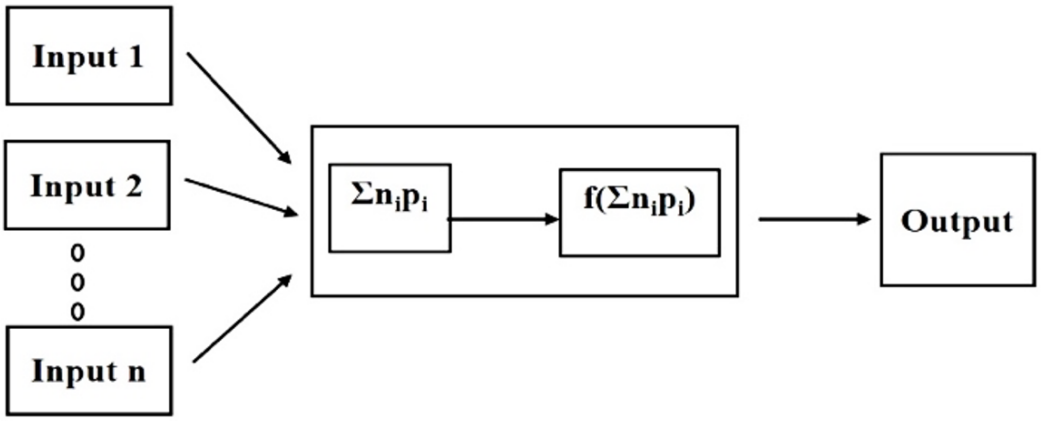


Figure 3.1: Architecture of the artificial neural network. The input layer consists of data provided, the middle layer is a hidden layer which consists of the number of neurons which consists of activation functions and an output layer which consists of processed information.

The structure of an ANN is defined by the numbers of layers and nodes, by the way, they are linked (activation function) and the method to estimate the weights.

3.2.1.1 Normalisation of Data

The data normalisation is recommended in an artificial neural network, which speeds up the learning and converges faster. The normalisations lead to the magnitude of the data on the same scale. Min-Max normalisation is the most common normalisation where the scale will be between 0-1. The following bellow Equation 3.1 is used for the normalisation of data:

Equation 3.1: Equation for min-max normalisation of data. Where Xn is normalised data, Xi is the data value; Xmax is the maximum value of data, Xmin is the minimum value of data.

$$X_n = \frac{X_i - X_{\min}}{(X_{max} - X_{\min})}$$

3.2.1.2 Cross-validation

Cross-validation is a method of model validation, to validate how well the model predicts the new data which was not used to build the model. To evaluate the model, some data will be removed from the training and used to test the model performance with it. In K-fold cross-validation, the data is divided into a k subset. Each time, a model is tested with one of the k subsets and a k-1 subset is used for the model training. And average errors across all k trials are computed. Leave-one-out cross-validation (LOOCV) is a K-fold validation where k equals the number of data points, N. The model will be trained with all data except one point for which prediction is made. The average error is computed and used to evaluate the model. The LOOCV is the nearest solution to the final model- as only one dataset is excluded from training. Parameter of the model are nearest to the final model however the limitation is that it might not be efficient in predicting the outliers.

3.2.2 Input for Building the ANN Model

The flux measurement data, from the *in-vitro* reconstructed upper part of glycolysis (Fiévet *et al.*, 2006) was used to build the artificial neural network (Figure 3.2). The input for the ANN model consists of concentrations of enzymes phosphoglucoisomerase (PGI), phosphofructokinase (PFK), fructose bisphosphate aldolase (FBA) and triosephosphate isomerase (TPI) in mg/l, and the output was flux J (µM/s) measured as the NADH consumption by glycerol-3-phosphate dehydrogenase (G3PDH). The flux measured through the upper part of glycolysis was indirect, and assumed that most of the NADH in the system was consumed during the measurement. The data were normalised using the min-max method before building the neural network.

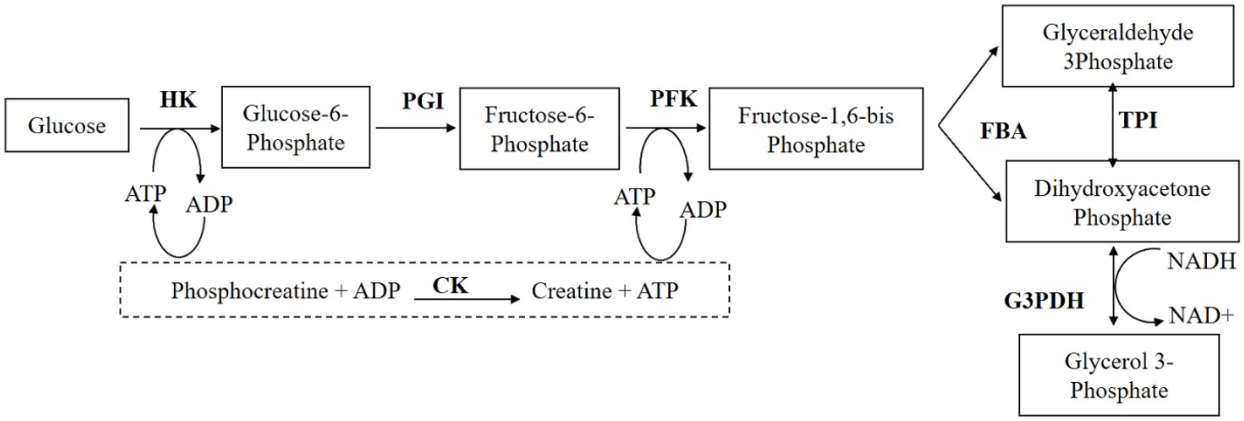


Figure 3.2: The upper part of glycolysis reconstructed *in vitro*. HK-hexokinase; PGI-phosphoglucoisomerase; PFK-phosphofructokinase; FBA-fructose bisphosphate aldolase; TPI-triosephosphate isomerase; G3PDH- glycerol-3-phosphate dehydrogenase, CK- Creatine kinase.

3.2.3 Experimental Details

The upper part of glycolysis was reconstructed *in vitro* (Figure 3.2), with a constant concentration of hexokinase and glycerol-3-phosphate dehydrogenase, while the other four enzymes (PGI, PFK, FBA and TPI) concentrations varied. The total enzyme concentration of the four enzymes (PGI, PFK, FBA and TPI) was constant at 101.9 mg/l. The NADH consumption using the glycerol-3-phosphate dehydrogenase was monitored every 2 seconds with the Uvikon-850 spectrometer at 390 nm from 60 to 120 seconds. The linear slope of NADH was calculated as the flux through the pathway. The assays were performed in triplicate by Fiévet *et al.* (Fiévet et al., 2006), at 25^0C**,** by adding 1mM ATP at pH 7.5 (Table 3.1)

Table 3.1: The flux measured in the experiment for enzyme balances used to build the ANN model. PGI: phosphoglucoisomerase; PFK: 6-phosphofructokinase; FBA: fructose bisphosphate aldolase; TPI: triosephosphate isomerase; Jobs: Experimentally observed flux with standard deviation (S.D).

mg/l				**µM/s**	
PGI	**PFK**	**FBA**	**TPI**	**Jobs**	**S.D.**
25	70	2	4.9	0.74	0.08
47.5	37.5	3.5	13.4	1.1	0.03
70	5	5	21.9	1.22	0.08
37.5	47.5	5	11.9	1.62	0.05
40	35	5	21.9	1.72	0.02
15	50	5	31.9	1.79	0
20	10	5	66.9	1.87	0.04
35	60	5	1.9	1.89	0.01
40	45	7	9.9	2.07	0.12
33	1	66.23	1.66	2.2	0.06
45	37.5	8.5	10.9	2.32	0.06
22.5	30	8.5	40.9	2.34	0.1
35	32.5	8.5	25.9	2.39	0.21
25	27.5	10	39.4	2.49	0.07
3.72	1.95	86.61	9.61	3.99	0.13
45	40	12	4.9	4.18	0.22
25	50	12	14.9	4.18	0.15
55	7.5	22.5	16.9	4.53	0.65
55	15	12	19.9	4.56	0.06
3.98	2.28	81.52	14.12	4.62	0.06
9.4	2.58	86.61	3.31	5.05	0.13
4.75	2.63	81.52	13	5.13	0.19
10	20	15	56.9	5.15	0.26
4.23	2.62	76.42	18.62	5.46	0.1
31	3	66.23	1.66	5.9	0.03
6.4	2.69	86.61	6.19	6.11	0.15
3.81	2.71	81.52	13.85	6.12	0.12
5.79	3.3	76.42	16.38	6.38	0.29
7.36	3.21	86.61	4.72	6.47	0.08
5.38	3.01	86.61	6.9	6.49	0.09
25	50	20	6.9	6.64	0.1
15	65	20	1.9	6.69	0.11
3.4	2.81	86.61	9.08	6.92	0.24

1	33	66.23	1.66	7.23	0.01
8.62	3.47	86.61	3.19	7.25	0.11
9.88	3.73	86.61	1.67	7.31	0.04
20	42.5	28.5	10.9	7.57	0.65
30	25	26	20.9	7.62	0.15
45.65	13.3	25.47	17.47	7.65	0.32
6.37	4.17	76.42	14.94	7.71	0.34
4.66	4.03	81.52	11.69	7.71	0.11
4.56	4.53	71.33	21.48	7.92	0.11
20	45	32.5	4.4	8.28	0.33
17.75	4.96	66.23	12.95	8.36	0.15
51.97	14.22	25.47	10.23	8.45	0.23
39.32	12.38	25.47	24.71	8.46	0.11
5.35	4.48	76.42	15.65	8.5	0.09
5.92	5.25	76.42	14.3	8.9	0.06
8.76	5.41	66.23	21.49	8.96	0.08
18.53	4.91	71.33	7.12	9.08	0.35
3.33	5.75	66.23	26.58	9.24	0.04
6	18	66.23	11.66	9.31	0.1
28	6	66.23	1.66	9.35	0.46
29.71	6.62	56.04	9.53	9.39	0.22
33.82	18.87	35.66	13.55	9.5	0.18
32.45	28.74	35.66	5.03	9.68	0.14
40	10	40	11.9	9.7	0.55
23.65	7.59	61.14	9.52	9.72	0.18
24.36	7.48	56.04	14.01	9.73	0.11
37.9	12.6	35.66	15.73	9.74	0.05
15.24	5.45	76.42	4.78	9.76	0.03
28.1	8.42	50.95	14.43	9.77	0.13
22.35	16.51	45.85	17.18	9.8	0.27
22.56	7.84	50.95	20.55	9.86	0.05
27.65	9.99	45.85	18.4	10.05	0.09
13.29	8.21	56.04	24.36	10.08	0.05
16	7	66.23	12.66	10.1	0.29
19	7	66.23	9.66	10.11	0.27
11.5	25	47.5	17.9	10.11	0.34
8	8	66.23	19.66	10.25	0.07
33.81	9.61	45.85	12.63	10.26	0.03
22.62	8.82	50.95	19.51	10.37	0.08

15.55	30.53	45.85	9.96	10.4	0.22
3	45	50	3.9	10.5	0.35
21.5	7.41	56.04	16.94	10.52	0.07
18.35	27.25	50.95	5.35	10.55	0.29
20.45	19.14	50.95	11.36	10.56	0.42
19.12	16.02	56.04	10.71	10.71	0.19
8.31	6.98	61.14	25.47	10.74	0.23
12	20.03	56.04	13.82	10.79	0.24
22	12	66.23	1.66	10.8	n,d,
11.15	8.5	61.14	21.11	10.82	0.19
24.06	26.13	40.76	10.95	10.88	0.3
25	9	66.23	1.66	10.9	0.14
9.9	28.14	56.04	7.82	10.95	0.26
11.17	25.48	50.95	14.3	11.01	0.16
5.26	11.19	71.33	14.12	11.03	0.16
19.19	17	56.04	9.67	11.05	0.29
23.46	10.7	50.95	16.79	11.08	0.25
11.23	21.97	50.95	17.75	11.11	0.07
12.2	27.64	35.66	26.39	11.19	0.22
8.44	9.53	66.23	17.69	11.22	0.1
6.82	14.17	76.42	4.49	11.33	0.38
9.01	8.8	66.23	17.85	11.39	0.24
10	15	66.23	10.66	11.45	0.49
12	13	66.23	10.66	11.45	0.21
9.06	16.44	56.04	20.35	11.49	0.1
20.6	7.18	66.23	7.89	11.52	0.08
30.21	11.88	45.85	13.96	11.54	0.07
6.79	20.43	56.04	18.63	11.55	0.16
9.64	16.24	66.23	9.79	11.56	0.23
10.61	18.22	50.95	22.12	11.57	0.29
12.34	16.16	56.04	17.36	11.58	0.06
14	18	66.23	3.66	11.6	0
23.34	7.8	56.04	14.72	11.63	0.14
18.07	18.85	56.04	8.93	11.64	0.05
4.24	25.75	66.23	5.67	11.7	0.3
16.4	17.41	50.95	17.14	11.75	0.1
18.13	15.35	56.04	12.38	11.79	0.08
16	16	66.23	3.66	11.85	0.21
13	13	66.23	9.66	11.9	0.14

10	10	66.23	15.66	12.05	0.07
15	40	45	1.9	12.07	0.81
15.66	23.52	45.85	16.86	12.15	0.22
7.06	14.74	76.42	3.67	12.23	0.13
5.08	15.5	71.33	9.99	12.28	0.13
7	12	66.23	16.66	12.35	0.21
11.04	15.68	61.14	14.03	12.47	0.17
11.51	13.37	66.23	10.79	12.63	0.15
4	25	66.23	6.66	12.65	0.21
15	15	55	16.9	12.9	0.53

3.2.4 Structure of ANN

The artificial neural network was built with a single layer of hidden units, (Hornik, Maxwell, & White, 1989) using statistical tool R **(**version 3.4.3**)** ((R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.), using three different packages: nnet (version 7.3-12) (Venables & B. D. Ripley, 2002), neuralnet (version 1.33) (Günther & Fritsch, 2010) and RSNNS (version 0.4-10) (Bergmeir & Benitez, 2012). The algorithm nnet, trains the feed-forward neural network general quasi-Newton optimisation (BFGS algorithm) procedure in one hidden layer (Bergmeir & Benitez, 2012), the neuralnet implements two types of resilient back-propagation which is one of the fastest algorithms (Günther & Fritsch, 2010) whereas in RSNNS; different architecture and learning functions are implemented.

The network consists of three layers: a) input (I), b) hidden layer (H) and c) output (O). These layers are connected by edges or neurons. The weighted sum of neuron inputs is submitted to a function which conditions neuron activation. There is no rule for deciding numbers of neuron units in a single hidden layer; to choose the best algorithm out of three (i.e. nnet, neuralnet and RSNNS), first chose number of hidden units according to Equation 3.2 and compared the RMSE (Equation 3.3) and coefficient of determination (R^2) (Equation 3.4) values between the three methods. The algorithm with the lowest RMSE value and highest R^2 value during the leave-one-out cross-validation was chosen as an algorithm of interest and the effect of numbers of hidden units on RMSE and R^2 was analysed between 1 and 25 hidden units.

Equation 3.2: Where Nh is number of hidden units; Ns: number of sampling in training data; Ni: number of input neurons; No: number of output neurons; α: arbitrary scaling factor 2-10. In this study, Ns= 120, Ni= 4, No = 1 and ∝=2 are used.

$$\text{Nh}= \frac{\text{Ns}}{(\propto *(\text{Ni}+\text{No})}$$

Equation 3.3: Where RMSE is the root mean square error, Yi is ANN predicted value; yi is experimental value; n is the number of predictions.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - y_i)^2}{n}}$$

Equation 3.4: Where R2 is the coefficient of determination; Yi, ANN is predicted values; yi is experimental value; n is numbers of predictions, $\bar{y}$ is the average of experimental values.

$$R^2 = \frac{\sum_i \left(y_i - Y_i\right)^2}{\sum_i (y_i - \bar{y})^2} \text{ where } \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

3.3 Results and Discussion

The three-neural network algorithms: nnet (Venables & B. D. Ripley, 2002), neuralnet (Günther & Fritsch, 2010) and RSNNS (Bergmeir & Benitez, 2012) were built with a hidden number of units ranging from 9 to12, as shown in Equation 3.2. The RMSE and coefficient of determination are compared between algorithms during leave-one-out cross-validation. Out of the three algorithms tested, neuralnet performed better than the other two (Table 3.2), allowing the option of choosing two different activation functions, *i.e.*, logistic (sigmoidal) and tanh **(**Equation 3.5 and Equation 3.6 respectively**)**.

Table 3.2: Comparison of RMSE and R-squared values during the leave-one-out cross-validation between neuralnet, nnet and RSNNS algorithm.

number of hidden units	**RMSE**				**R^2**			
	neuralnet: logistic	**neuralnet: tanh**	**nnet**	**RSNNS**	**neuralnet: logistic**	**neuralnet: tanh**	**nnet**	**RSNNS**

9	0.923	0.899	1.405	2.477	0.923	0.929	0.851	0.437
10	0.933	1.113	1.289	2.523	0.92	0.887	0.848	0.414
11	0.949	0.836	1.483	2.494	0.921	0.936	0.821	0.428
12	0.97	1.034	1.902	2.537	0.916	0.907	0.759	0.412

Equation 3.5: logistic activation function.

$$Logistic(x) = \frac{1}{1 + e^{-x}}$$

Equation 3.6: tanh activation function.

$$tan\,h(x) = \frac{2}{1 + e^{-2x}} - 1$$

Using the neuralnet model, with "logistic" and "tanh" activation functions, the effect of numbers of hidden units on RMSE and R^2 was studied (Figure 3.3) with a leave-one-out-cross-validation procedure. The logistic function with 13 hidden units gives an RMSE of 0.847, R^2 of 0.93 and tanh function RMSE of 0.804 and R^2 of 0.94 with 6 hidden units.

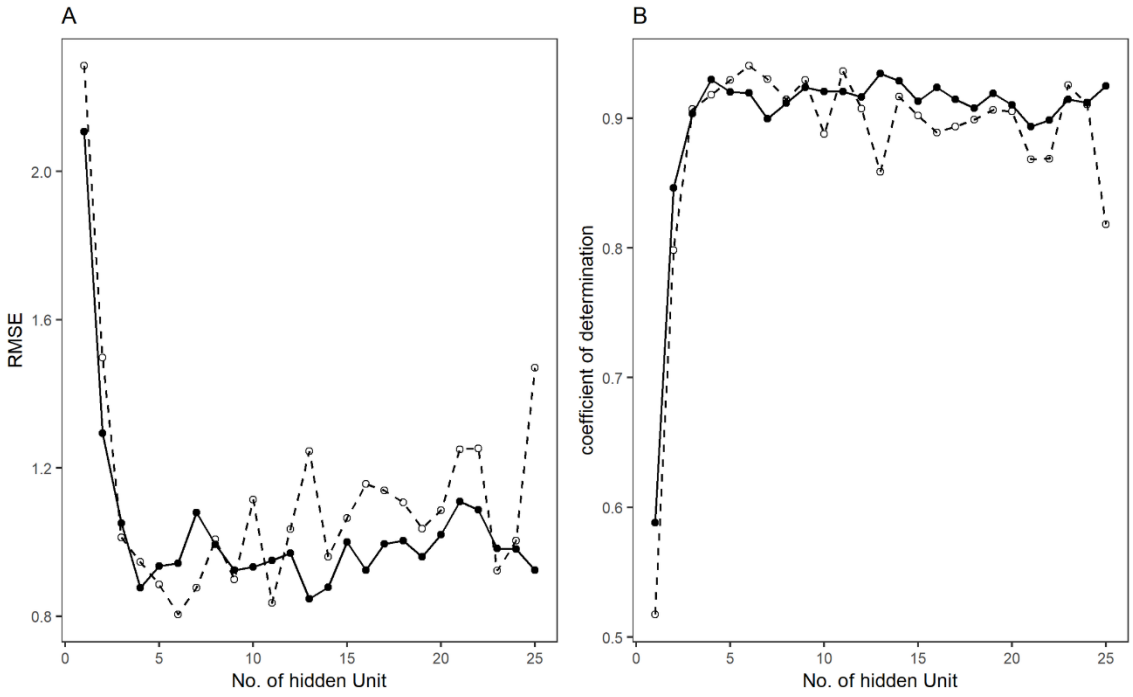


Figure 3.3: Effect of numbers of hidden units on RMSE (A) and coefficient of determination (B) activation function logistic (filled circle, solid line) and tanh (open circle and dotted lines).

Experimental flux was compared with the ANN predicted flux by leave-one-outcross-validation procedure, using chosen hidden units with logistic and tanh function (Figure 3.4). The effects of enzyme concentrations on the predicted flux and experimental flux were compared and found to follow a similar trend (Figure 3.5).

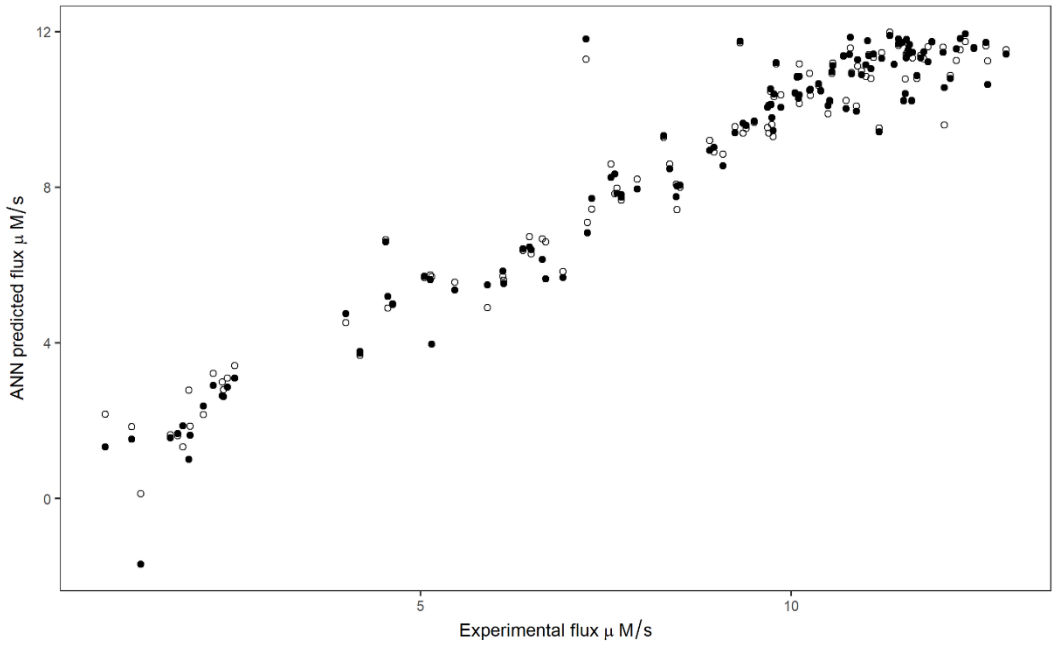


Figure 3.4: The relationship between flux predicted by leave-one-outcross-validation and experimental flux. Filled and open circles represent logistic and tanh activation functions respectively.

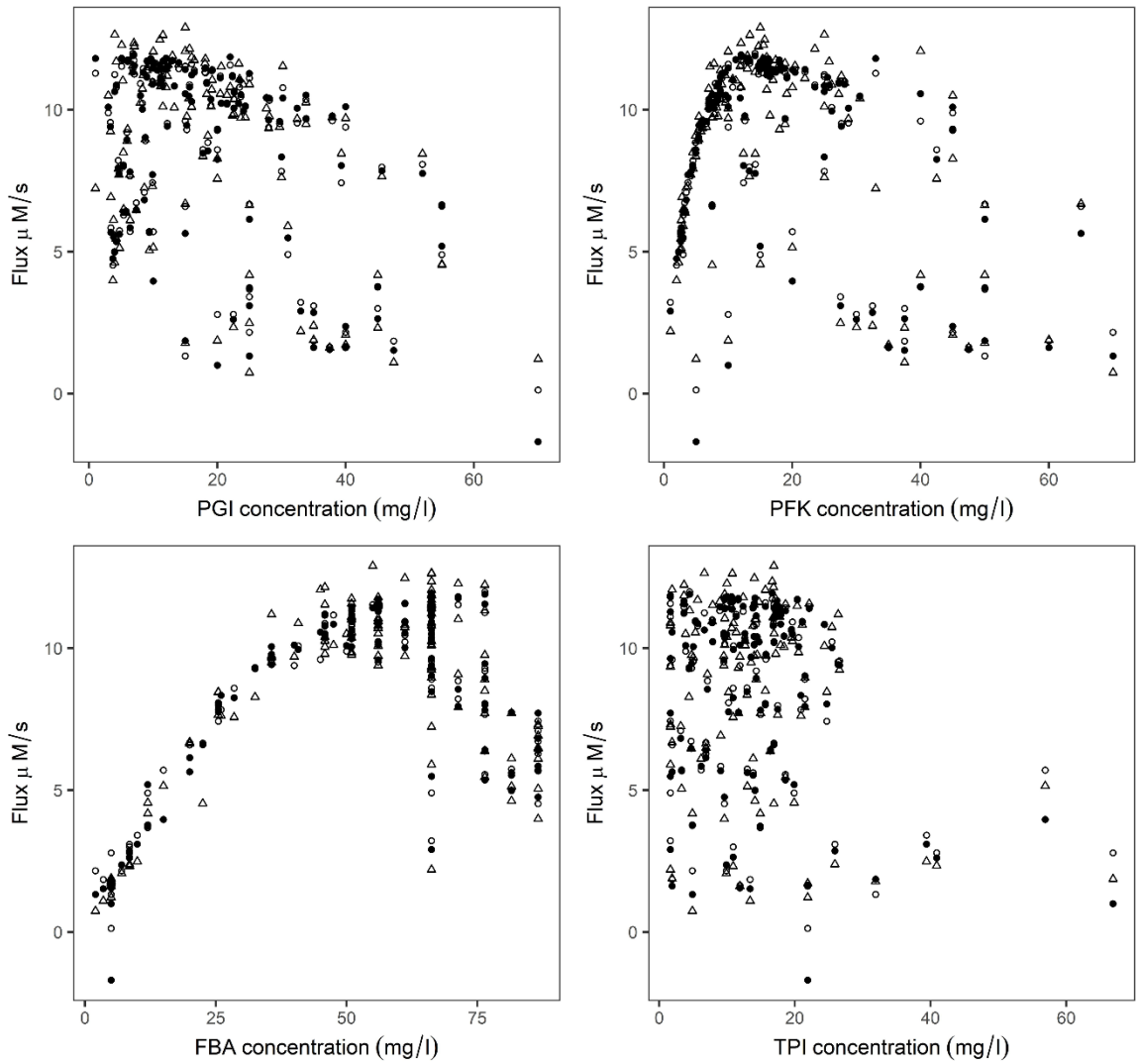


Figure 3.5: The relationship between the individual enzyme concentration with experimental and ANN predicted flux. Filled circle and open circle are enzyme concentration vs predicted flux with logistic and tanh activation functions respectively, open triangles represent the experiment.

During the cross-validation of the neural network model, a negative flux value is predicted for one combination of enzymes (Table 3.3: Index-3). This is because, during a leave-one-out procedure (LOOcv), one combination of the enzymes (concentration of PGI, PFK, FBA and TPI) was not included in the model training and the flux must be predicted for that particular combination. The negative value shows the poor ability of the ANN model to predict the outliers, i.e. a combination that was not close (in terms of PGI, PFK, FBA and TPI concentrations) to those included in the training data set.

The original study by Fiévet *et al.* (Fiévet *et al.*, 2006) developed a model to predict flux. As the authors mentioned in their article, their flux predictor overestimates the observed flux by a constant factor. The predicted flux, in their method, has an R^2 value of 0.86**,** whereas an ANN

approach with logistic function shows an R^2 value to be 0.93 and in case of tanh activation function, an R^2 of 0.94, obtained with leave-one-out cross-validation, which implies that the ANN approach is more efficient in predicting the flux than the method developed in the Fiévet study. The effect of enzyme concentrations on the predicted flux by both methods follows a similar trend.

The difference between actual flux and ANN predicted flux was an average of 0.57 µM/s for logistic and for tanh, with a standard deviation of 0.63 and 0.57 respectively (Table 3.3)**,** whereas the Fiévet *et al.* study showed an average of 3.3 and a standard deviation of 2.2 with actual predicted values (Fiévet *et al.*, 2006). Fiévet *et al.* stated that their method overestimates the flux values by a constant factor of 1.38 (Fiévet *et al.*, 2006). Hence, by dividing the predicted flux values by 1.38, corrected values were obtained. The new average corrected value is 1.04 and the standard deviation is 0.78 with the experimental value. This indicates that the ANN method performs better than the method described in the original study by Fiévet *et al.* This ANN-based method provides additional degrees of freedom over the method proposed in Fiévet *et al.* (Fiévet *et al.*, 2006). Indeed, numbers of degrees of freedom increase with numbers of hidden units. This makes it possible to obtain an important advantage regarding the error inherent to the learning phase. the logistic activation function was retained for further study.

Table 3.3: Comparison of flux values (in µM/S) between observed flux (J_{Exp}), J.B Fievet (J_{Fievet}) and ANN predicted flux with activation functions logistic ($J_{\{ANN:\ logistic\}}$) and tanh ($J_{\{ANN:\ tanh\}}$) and the standard deviation of observed flux (J_{SD}).

Index	**J_{Exp}**	**J_{SD}**	**J_{Fievet}**	**$J_{\{Fievet:\ Corrected\}}$**	**$J_{\{ANN:\ logistic\}}$**	**$J_{\{ANN:\ tanh\}}$**	**Difference_[J_{Exp} : J_{Fievet}]**	**Difference_[J_{Exp} : $J_{\{Fievet:Corrected\}}$]**	**Difference_[J_{Exp} : $J_{\{ANN:logistic\}}$]**	**Difference_[J_{Exp} : $J_{\{ANN:tanh\}}$]**
1	0.74	0.08	1.14	0.83	1.33	2.16	0.4	0.09	0.59	1.42
2	1.1	0.03	1.97	1.43	1.53	1.85	0.87	0.33	0.43	0.75
3	1.22	0.08	2.44	1.77	-1.69	0.13	1.22	0.55	2.91	1.09
4	1.62	0.05	2.79	2.02	1.56	1.64	1.17	0.4	0.06	0.02
5	1.72	0.02	2.78	2.01	1.66	1.62	1.06	0.29	0.06	0.1

6	1.79	0	2.76	2	1.86	1.32	0.97	0.21	0.07	0.47
7	1.87	0.04	2.6	1.88	1	2.79	0.73	0.01	0.87	0.92
8	1.89	0.01	2.8	2.03	1.62	1.85	0.91	0.14	0.27	0.04
9	2.07	0.12	3.86	2.8	2.37	2.16	1.79	0.73	0.3	0.09
10	2.2	0.06	3.08	2.23	2.91	3.22	0.88	0.03	0.71	1.02
11	2.32	0.06	4.63	3.36	2.64	3	2.31	1.04	0.32	0.68
12	2.34	0.1	4.54	3.29	2.62	2.79	2.2	0.95	0.28	0.45
13	2.39	0.21	4.59	3.33	2.86	3.09	2.2	0.94	0.47	0.7
14	2.49	0.07	5.26	3.81	3.1	3.41	2.77	1.32	0.61	0.92
15	3.99	0.13	4.9	3.55	4.76	4.52	0.91	0.44	0.77	0.53
16	4.18	0.22	6.4	4.64	3.73	3.68	2.22	0.46	0.45	0.5
17	4.18	0.15	6.43	4.66	3.78	3.75	2.25	0.48	0.4	0.43
18	4.53	0.65	8.4	6.09	6.6	6.65	3.87	1.56	2.07	2.12
19	4.56	0.06	5.98	4.33	5.2	4.9	1.42	0.23	0.64	0.34
20	4.62	0.06	5.52	4	5	4.99	0.9	0.62	0.38	0.37
21	5.05	0.13	6.77	4.91	5.71	5.68	1.72	0.14	0.66	0.63
22	5.13	0.19	6.27	4.54	5.62	5.74	1.14	0.59	0.49	0.61
23	5.15	0.26	6.98	5.06	3.97	5.7	1.83	0.09	1.18	0.55
24	5.46	0.1	6.09	4.41	5.36	5.55	0.63	1.05	0.1	0.09
25	5.9	0.03	7.75	5.62	5.49	4.9	1.85	0.28	0.41	1
26	6.11	0.15	6.72	4.87	5.84	5.71	0.61	1.24	0.27	0.4
27	6.12	0.12	6.17	4.47	5.53	5.61	0.05	1.65	0.59	0.51
28	6.38	0.29	7.51	5.44	6.42	6.38	1.13	0.94	0.04	0
29	6.47	0.08	7.77	5.63	6.47	6.72	1.3	0.84	0	0.25
30	6.49	0.09	7.09	5.14	6.4	6.29	0.6	1.35	0.09	0.2
31	6.64	0.1	10.19	7.38	6.14	6.67	3.55	0.74	0.5	0.03
32	6.69	0.11	9.95	7.21	5.64	6.59	3.26	0.52	1.05	0.1

33	6.92	0.24	6.2	4.49	5.68	5.84	0.72	2.43	1.24	1.08
34	7.23	0.01	6.3	4.57	11.8	11.29	0.93	2.66	4.57	4.06
35	7.25	0.11	8.36	6.06	6.83	7.09	1.11	1.19	0.42	0.16
36	7.31	0.04	8.92	6.46	7.72	7.44	1.61	0.85	0.41	0.13
37	7.57	0.65	13.45	9.75	8.26	8.59	5.88	2.18	0.69	1.02
38	7.62	0.15	12.05	8.73	8.34	7.83	4.43	1.11	0.72	0.21
39	7.65	0.32	10.72	7.77	7.84	7.98	3.07	0.12	0.19	0.33
40	7.71	0.34	8.27	5.99	7.74	7.74	0.56	1.72	0.03	0.03
41	7.71	0.11	8.86	6.42	7.81	7.67	1.15	1.29	0.1	0.04
42	7.92	0.11	8.59	6.22	7.95	8.21	0.67	1.7	0.03	0.29
43	8.28	0.33	15.02	10.88	9.32	9.28	6.74	2.6	1.04	1
44	8.36	0.15	10.79	7.82	8.48	8.6	2.43	0.54	0.12	0.24
45	8.45	0.23	10.92	7.91	7.76	8.08	2.47	0.54	0.69	0.37
46	8.46	0.11	10.49	7.6	8.03	7.43	2.03	0.86	0.43	1.03
47	8.5	0.09	8.97	6.5	8.05	8	0.47	2	0.45	0.5
48	8.9	0.06	10.03	7.27	8.94	9.2	1.13	1.63	0.04	0.3
49	8.96	0.08	10.55	7.64	9.03	8.91	1.59	1.32	0.07	0.05
50	9.08	0.35	10.97	7.95	8.56	8.85	1.89	1.13	0.52	0.23
51	9.24	0.04	8.74	6.33	9.4	9.55	0.5	2.91	0.16	0.31
52	9.31	0.1	15.81	11.46	11.75	11.71	6.5	2.15	2.44	2.4
53	9.35	0.46	12.43	9.01	9.64	9.39	3.08	0.34	0.29	0.04
54	9.39	0.22	12.52	9.07	9.59	9.51	3.13	0.32	0.2	0.12
55	9.5	0.18	14.66	10.62	9.7	9.67	5.16	1.12	0.2	0.17
56	9.68	0.14	15.84	11.48	10.05	9.53	6.16	1.8	0.37	0.15
57	9.7	0.55	13.13	9.51	10.11	9.39	3.43	0.19	0.41	0.31
58	9.72	0.18	13.77	9.98	10.53	10.46	4.05	0.26	0.81	0.74
59	9.73	0.11	13.23	9.59	10.13	10.11	3.5	0.14	0.4	0.38

60	9.74	0.05	13.24	9.59	9.79	9.61	3.5	0.15	0.05	0.13
61	9.76	0.03	11.75	8.51	9.45	9.3	1.99	1.25	0.31	0.46
62	9.77	0.13	13.58	9.84	10.4	10.33	3.81	0.07	0.63	0.56
63	9.8	0.27	16.29	11.8	11.2	11.16	6.49	2	1.4	1.36
64	9.86	0.05	12.94	9.38	10.05	10.38	3.08	0.48	0.19	0.52
65	10.05	0.09	13.89	10.07	10.42	10.43	3.84	0.02	0.37	0.38
66	10.08	0.05	13.11	9.5	10.84	10.83	3.03	0.58	0.76	0.75
67	10.1	0.29	13.13	9.51	10.29	10.29	3.03	0.59	0.19	0.19
68	10.11	0.27	13.34	9.67	10.37	10.15	3.23	0.44	0.26	0.04
69	10.11	0.34	16.89	12.24	10.85	11.17	6.78	2.13	0.74	1.06
70	10.25	0.07	12.73	9.22	10.5	10.92	2.48	1.03	0.25	0.67
71	10.26	0.03	13.82	10.01	10.52	10.36	3.56	0.25	0.26	0.1
72	10.37	0.08	13.68	9.91	10.66	10.62	3.31	0.46	0.29	0.25
73	10.4	0.22	17.96	13.01	10.48	10.47	7.56	2.61	0.08	0.07
74	10.5	0.35	12.1	8.77	10.1	9.89	1.6	1.73	0.4	0.61
75	10.52	0.07	13.05	9.46	10.23	10.2	2.53	1.06	0.29	0.32
76	10.55	0.29	19.26	13.96	10.97	10.93	8.71	3.41	0.42	0.38
77	10.56	0.42	17.95	13.01	11.12	11.19	7.39	2.45	0.56	0.63
78	10.71	0.19	17.85	12.93	11.37	11.37	7.14	2.22	0.66	0.66
79	10.74	0.23	11.69	8.47	10.02	10.23	0.95	2.27	0.72	0.51
80	10.79	0.24	17.82	12.91	11.41	11.41	7.03	2.12	0.62	0.62
81	10.8	n,d,	17.51	12.69	11.86	11.57	6.71	1.89	1.06	0.77
82	10.82	0.19	13.48	9.77	10.91	11.76	2.66	1.05	0.09	0.94
83	10.88	0.3	16.88	12.23	9.96	10.08	6	1.35	0.92	0.8
84	10.9	0.14	15.48	11.22	11.28	11.11	4.58	0.32	0.38	0.21
85	10.95	0.26	18.48	13.39	10.9	11	7.53	2.44	0.05	0.05
86	11.01	0.16	17.59	12.75	11.14	10.84	6.58	1.74	0.13	0.17

87	11.03	0.16	13.48	9.77	11.77	10.22	2.45	1.26	0.74	0.81
88	11.05	0.29	18.2	13.19	11.38	11.41	7.15	2.14	0.33	0.36
89	11.08	0.25	14.92	10.81	11.05	10.79	3.84	0.27	0.03	0.29
90	11.11	0.07	17.06	12.36	11.42	11.33	5.95	1.25	0.31	0.22
91	11.19	0.22	14.36	10.41	9.42	9.52	3.17	0.78	1.77	1.67
92	11.22	0.1	13.93	10.09	11.31	11.46	2.71	1.13	0.09	0.24
93	11.33	0.38	16.17	11.72	11.9	11.98	4.84	0.39	0.57	0.65
94	11.39	0.24	13.61	9.86	11.16	11.15	2.22	1.53	0.23	0.24
95	11.45	0.49	16.86	12.22	11.81	11.69	5.41	0.77	0.36	0.24
96	11.45	0.21	17.14	12.42	11.68	11.65	5.69	0.97	0.23	0.2
97	11.49	0.1	15.97	11.57	11.73	11.71	4.48	0.08	0.24	0.22
98	11.52	0.08	13.61	9.86	10.23	10.22	2.09	1.66	1.29	1.3
99	11.54	0.07	14.93	10.82	10.41	10.78	3.39	0.72	1.13	0.76
100	11.55	0.16	15.71	11.38	11.32	11.42	4.16	0.17	0.23	0.13
101	11.56	0.23	17.45	12.64	11.79	11.8	5.89	1.08	0.23	0.24
102	11.57	0.29	16.13	11.69	11.4	11.46	4.56	0.12	0.17	0.11
103	11.58	0.06	16.85	12.21	11.53	11.67	5.27	0.63	0.05	0.09
104	11.6	0	19.32	14	11.66	11.44	7.72	2.4	0.06	0.16
105	11.63	0.14	13.48	9.77	10.23	10.94	1.85	1.86	1.4	0.69
106	11.64	0.05	18.64	13.51	11.46	11.33	7	1.87	0.18	0.31
107	11.7	0.3	14.94	10.83	10.87	10.79	3.24	0.87	0.83	0.91
108	11.75	0.1	17.04	12.35	11.32	11.39	5.29	0.6	0.43	0.36
109	11.79	0.08	17.5	12.68	11.48	11.3	5.71	0.89	0.31	0.49
110	11.85	0.21	18.97	13.75	11.23	11.61	7.12	1.9	0.62	0.24
111	11.9	0.14	17.09	12.38	11.74	11.73	5.19	0.48	0.16	0.17
112	12.05	0.07	14.68	10.64	11.47	11.61	2.63	1.41	0.58	0.44
113	12.07	0.81	18.22	13.2	10.57	9.6	6.15	1.13	1.5	2.47

114	12.15	0.22	17.11	12.4	10.8	10.87	4.96	0.25	1.35	1.28
115	12.23	0.13	16.53	11.98	11.56	11.25	4.3	0.25	0.67	0.98
116	12.28	0.13	14.77	10.7	11.82	11.53	2.49	1.58	0.46	0.75
117	12.35	0.21	14.61	10.59	11.94	11.74	2.26	1.76	0.41	0.61
118	12.47	0.17	17.1	12.39	11.57	11.58	4.63	0.08	0.9	0.89
119	12.63	0.15	16.91	12.25	11.72	11.63	4.28	0.38	0.91	1
120	12.65	0.21	14.5	10.51	10.64	11.24	1.85	2.14	2.01	1.41
121	12.9	0.53	16.79	12.17	11.42	11.53	3.89	0.73	1.48	1.37
The average difference between observed and predicted							3.32	1.05	0.57	0.57
The standard deviation of the difference between observed and predicted							2.14	0.78	0.63	0.57

3.4 Conclusion

Kinetic modelling of metabolic pathways is challenging because of difficulties in estimating the kinetic parameters (Bisswanger, 2014; Vasilakou *et al.*, 2016) and is sometimes expensive because of the high-cost substrates and technologies involved (Gupta, Rathi, Gupta, & Bradoo, 2003; Hakenberg et al., 2004)**,** whereas the constraint-based model does not use any kinetic parameters but is efficient enough to predict the flux of metabolites. Choosing the optimum enzyme concentrations for the highest flux could be a challenge when conducting experiments. Using artificial intelligence with available experimental data can help us find a quicker and more cost-effective solution for biological problems.

In this study, a neural network model was tested with three algorithms and several architectures to determine the best configuration of the ANN model. Eventually, the neuralnet algorithm was retained for the study with two different activation functions: logistic (sigmoidal) and tanh, with RMSE and R^2 values of 0.847, 0.93 and 0.804, 0.94 respectively. The difference between actual flux and ANN predicted flux was an average of 0.57 for both activation functions. The Fiévet *et al.* method after the correction has an RMSE of 1.30, with a 1.05 difference between predicted and observed flux, which clearly indicates that the ANN method works better than the other method. It has not escaped our attention that the artificial neural network model depends on the diversity of the training data and hence training the model with a maximum of variability in the concentration of enzymes plays a crucial role.

Chapter 4 Artificial Neural Network for the Selection of Optimum Enzyme Balances.

4.1 Context

The selection of enzymes is crucial for *in vitro* metabolic engineering since low performing enzymes result in poor titer and yield. Homology based methodologies like Selenzyme (Carbonell *et al.*, 2018) are developed to select better performing enzymes. One of the main challenges of purified enzyme-based CFS is the selection of optimum enzyme concentrations for maximum product formation. The experimental selection of optimum enzyme concentrations is expensive and tedious. Hence, the development of a computational method for selecting optimum enzyme concentrations without detailed knowledge of their kinetic parameters, using other existing experimental data, is helpful.

Glycolysis, one of the central carbon metabolism pathways, is not only important for organisms, but is also of great importance in biotechnology for producing different biomolecules (J. Liu et al., 2017). Many chemicals such as organic acids (C. W. Song, Kim, Choi, Jang, & Lee, 2013; Jiangang Yang *et al.*, 2014) and biofuels (Clomburg & Gonzalez, 2010; X. Yang, Xu, & Yang, 2015) have been successfully produced with high titer using engineered microorganisms including *Saccharomyces cerevisiae* or *Escherichia coli.* Glycolysis is widely studied from various perspectives. The availability of data from Fievet *et al.* (Fiévet *et al.*, 2006) for flux prediction with different enzyme concentrations makes it a good candidate for developing a new approach to select optimum enzyme concentrations.

In Chapter 3, ANN was used to predict the flux through the upper part of glycolysis using enzyme concentrations, i.e., phosphoglucoisomerase (PGI), phosphofructokinase (PFK), fructose biphosphate aldolase (FBA), and triosephosphate isomerase (TPI) as the input to the model. The predicted flux has a root mean square error (RMSE) of 0.84 and an R^2 of 0.93, with 13 hidden units. Since the ANN is a training-based method, the new prediction depends on the training dataset. Since ANN is not efficient in extrapolating predictions (Balabin & Smirnov, 2012; Minns & Hall, 1996). the new predictions will always lie in the range of the known output predictions; in other words, we could say that they will remain “in-the-box”. High predicted output values will bump into a sort of “glass ceiling”. The working hypothesis was that, in reality, actual flux values could be higher than the predicted ones. So, to explore this “glass ceiling” space, a new methodology (GC-ANN, for glass ceiling ANN) was developed to predict the flux for the upper part of glycolysis, given enzyme concentrations using an artificial neural network.

4.2 Materials and Methods

All enzymes as well as phosphocreatine, glucose-6-phosphate, fructose-6-phosphate and fructose-1,6-bisphosphate were purchased from Sigma-Aldrich (St. Louis, MO, USA). D-Glucose, ATP, NADH, and NADP were obtained from Carl Roth GmbH (Karlsruhe, Germany). Hexokinase (HK), phosphoglucoisomerase (PGI), triose-phosphate isomerase (TPI), and glucose-6-phosphate dehydrogenase (G6PDH) originated from baker's yeast; fructose biphosphate aldolase (FBA), glycerol-3-phosphate dehydrogenase (G3PDH), and creatine kinase (CK) were obtained from rabbit muscle and phosphofructokinase (PFK) originated from *Bacillus stearothermophilus*. The enzymes were obtained as lyophilized powder except for PGI and TPI, which were ammonium sulphate suspensions. Detailed information on the enzymes used is provided in Table 4.1.

4.2.1 Determination of Protein Concentration

Protein concentrations were determined using the Bradford protein assay (Marion M. Bradford, 1976) from Bio-Rad Laboratories (Hercules, CA). The protein solutions of 10 µL was mixed with 200 µL of Bio-Rad Protein Assay Dye Reagent, incubated for 5 minutes at room temperature, and the absorbance was measured spectrophotometrically at 595 nm. A dilution series of 0.06–0.5 mg/ml BSA (Carl Roth GmbH) was used for calibration.

4.2.2 Enzyme Assays for the Determination of Kinetic Parameters

Enzyme assays were performed in 96-well UV-STAR® microplates (Greiner Bio-One GmbH, Kremsmünster, Austria) in a total volume of 100 µL at 25 °C. The reaction buffer contained 50 mM PIPES (pH 7.5), 100 mM KCl, and 5 mM magnesium acetate. The cofactors for the reactions were 1 mM ATP and 1 mM NADH or NADP. All reactions were monitored by recording the absorption at a wavelength of 340 nm (molar extinction coefficient $\varepsilon_{340\ nm,\ 25\ °C}$ 6.22 L $mmol^{-1}$ cm^{-1}). Lineweaver-Burk as well as Eadie-Hofstee representations were used For calculation of the kinetic parameters V_{max}, K_m, and k_{cat}.

Table 4.1: Enzymes used in this study for the upper part of glycolysis. All enzymes were bought from Sigma.

Enzyme	**EC**	**Acronym**	**Source**	**Sigma Cat. - No.**	**Lot No.**	**Purity**	**MW**	**active enzyme**	**MW/subunit**
Hexokinase	2.7.1.1	HK	baker's yeast	H4502-1KU	SLBT5451	mix of isoenzymes HXK1 & HXK2	110.0 kDa	homodimer	54 kDa
Phosphoglucoisomerase	5.3.1.9	PGI	baker's yeast	P5381-1KU	SLBW8689	n.i.	119.5 kDa	homodimer	61.3 kDa
Phosphofructokinase	2.7.1.11	PFK	Bacillus stearothermophilus	F0137-100UN	SLBW6641	n.i.	136.5 kDa	homotetramer	34 kDa
Fructose bisphosphate aldolase	4.1.2.13	FBA	rabbit muscle	A2714-100UN	SLBR7752V SLBV7445	>80%	157.4 kDa	homotetramer	39.3 kDa
Triose phosphate isomerase	5.3.1.1	TPI	Baker's yeast	T2507-5MG	036H8025	n.i.	53.6 kDa	homodimer	26.8 kDa
Glycerol-3-phosphate dehydrogenase	1.1.1.8	G3PDH	rabbit muscle	10127752001	21866328	traces of other enzymes	75.2 kDa	homodimer	37.6 kDa
Glyceraldehyde-3-phosphate dehydrogenase	1.2.1.12	GAPDH	rabbit muscle	G2267-1KU	SLBR0602V	>80%	144.0 kDa	homotetramer	36 kDa
Creatine kinase	2.7.3.2	CK	rabbit muscle	10127566001	25998433	traces of other enzymes	86.2 kDa	homodimer	43.1 kDa
Glucose-6-phosphate dehydrogenase	1.1.1.49	G6PDH	baker's yeast	G6378-250UN	SLBP6152V	--	--	--	--

4.2.2.1 Hexokinase, HK

The hexokinase activity was assayed using glucose-6-phosphate dehydrogenase (G6PDH) in a coupled reaction (Figure 4.1). The substrate glucose was converted to 6-phosphogluconate, the formation of NADPH was followed spectrophotometrically at 340 nm.

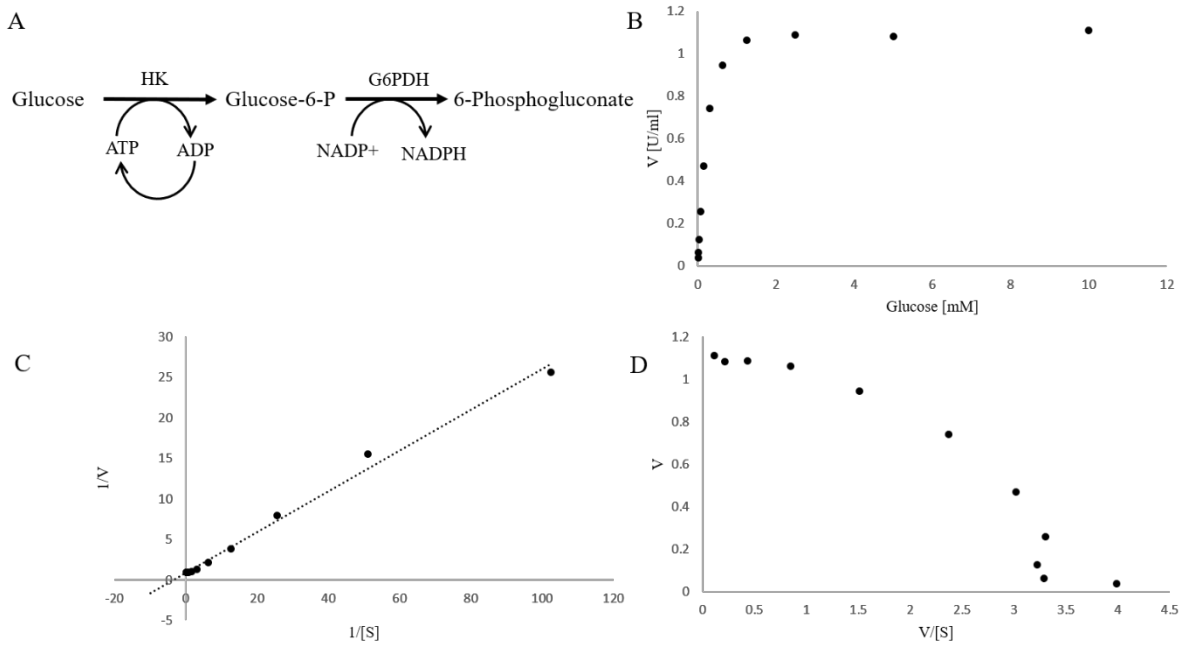


Figure 4.1: A) Coupled HK/G6PDH assay to assess the HK activity. (B) Michaelis-Menten kinetics. Mean of the 4 technical replicates. Corresponding (C) Lineweaver-Burk (goodness-of-fit R2=0.9923) and (D) Eadie-Hofstee (goodness-of-fit R2=0.9161) plots for the HK assayed with different concentrations of glucose.

4.2.2.2 Phosphoglucoisomerase, PGI

The phosphoglucoisomerase activity was assayed by coupling it to the reactions of PFK, FBA, TPI and G3PDH (Figure 4.2) The substrate glucose-6-phosphate was converted to glycerol-3-phosphate, the depletion of NADPH was followed spectrophotometrically at 340 nm.

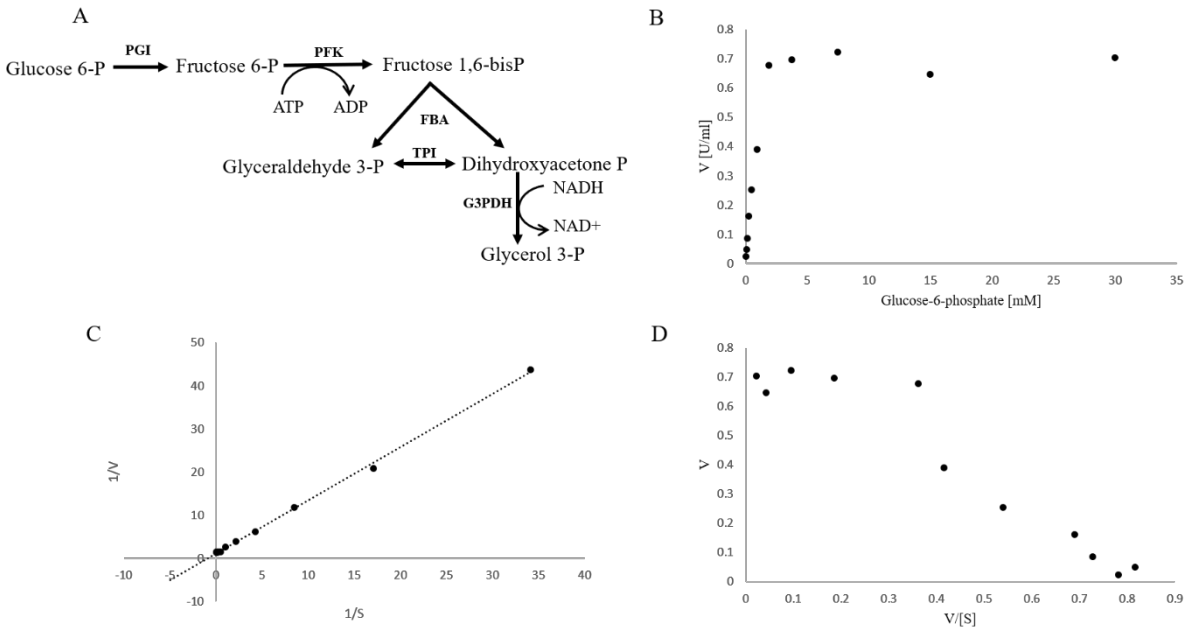


Figure 4.2: (A) Coupled PGI/PFK/FBA/TPI/GDH assay to assess the PGI activity. (B) Michaelis-Menten kinetics. Mean of the 4 technical replicates. Corresponding (C) Lineweaver-Burk (goodness-of-fit R_2=0.9987) and (D) Eadie-Hofstee (goodness-of-fit R_2=0.9123) plots for the PGI assayed with different concentrations of glucose-6-phosphate.

4.2.2.3 Phosphofructokinase, PFK

The phosphofructokinase activity was assayed by coupling it to the reactions of FBA, TPI and G3PDH (Figure 4.3). The substrate fructose-6-phosphate was converted to glycerol-3-phosphate, the depletion of NADPH was followed spectrophotometrically at 340 nm.

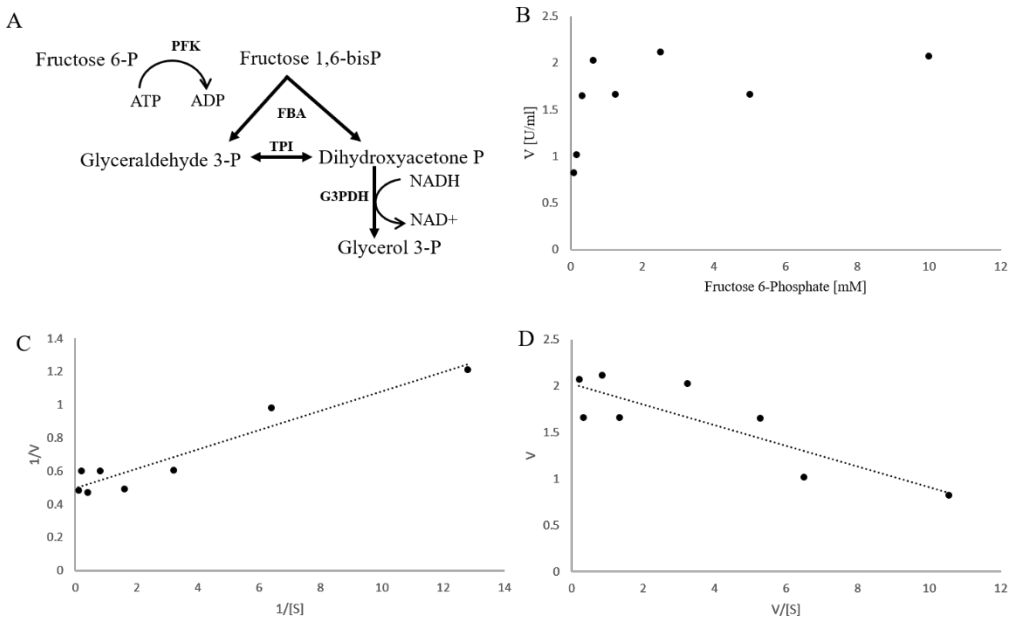


Figure 4.3: (A)Coupled PFK/FBA/TPI/GDH assay to assess the PFK activity. (B) Michaelis-Menten kinetics Mean of the 4 technical replicates. Corresponding (C) Lineweaver-Burk (goodness-of-fit R_2=0.9137) and (D) Eadie-Hofstee (goodness-of-fit R_2=0.7204) plots for the PFK assayed with different concentrations of fructose-6-phosphate.

4.2.2.4 Fructose bisphosphate aldolase, FBA

The fructose bisphosphate aldolase activity was assayed by coupling it to the reactions of TPI and G3PDH (Figure 4.4). The substrate fructose-1,6-bisphosphate converts to glycerol-3-phosphate and the depletion of NADPH is followed spectrophotometrically at 340 nm.

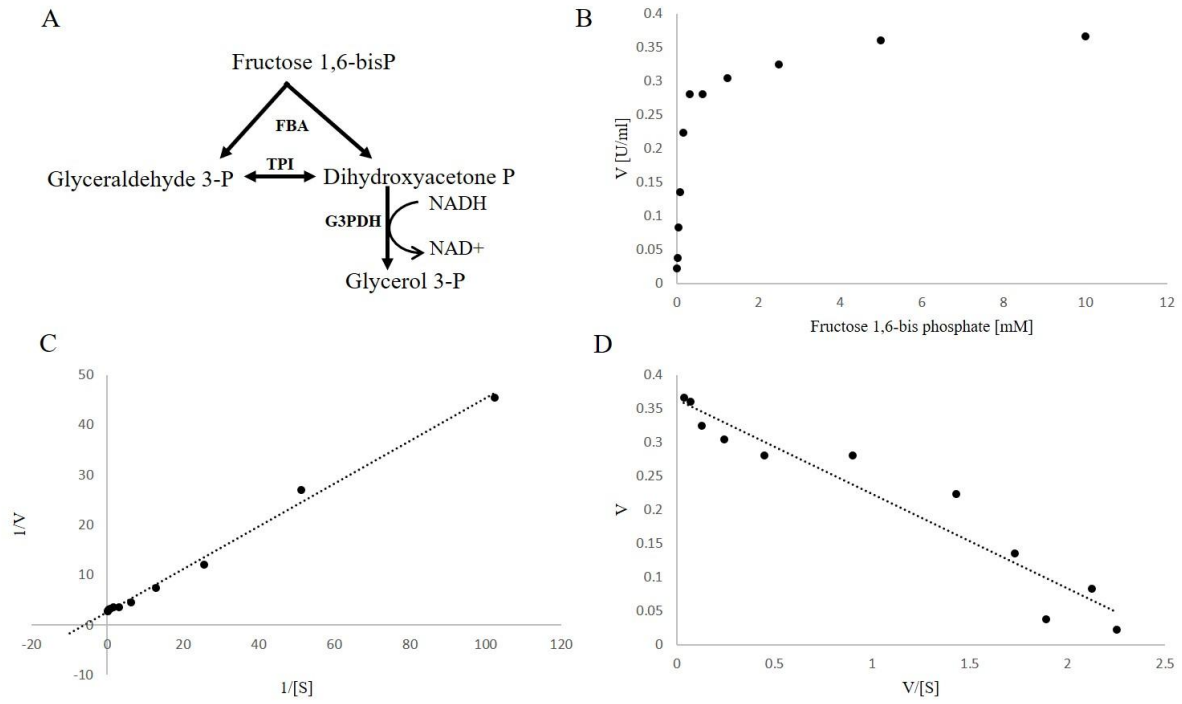


Figure 4.4: (A) Coupled FBA/TPI/GDH assay to assess the FBA activity. (B) Michaelis-Menten kinetics. Mean of the 4 technical replicates. Corresponding (C) Lineweaver-Burk (goodness-of-fit R_2=0.9940) and (D) Eadie-Hofstee (goodness-of-fit R_2=0.9274) plots for the FBA assayed with different concentrations of fructose-1,6-bisphosphate.

4.2.3 Flux Measurements

The total reaction volume of 100 µL contained fixed concentrations of 3 mM NADH, 20 mM phosphocreatine, 1 µM CK, 0.1 µM HK, and 1 µM G3PDH. The concentrations of PGI, PFK, FBA, and TPI were varied as indicated (Flux Determinations). The reactions were started with 1 mM ATP and 100 mM glucose. Blank reactions contained all ingredients except ATP and glucose. Each condition was measured in triplicates. The NADH decay was monitored every 3 s at 365 nm using a SynergyMxSMATBLD(+) Gen5 SW plater reader (SZABO-SCANDIC, Vienna, Austria). The slope of NADH decay was measured as the flux through the pathway (molar extinction coefficient $\varepsilon_{365\ nm,\ 25\ °C}$ 3.4 L $mmol^{-1}$ cm^{-1}).

4.3 Modifications and Calculations Used in the Study

4.3.1 Concentration Based on the Relative Activity

Our new methodology predicts the flux through the upper part of glycolysis based on the concentrations of the four enzymes, PGI, PFK, FBA and TPI. To make our prediction comparable to that of Fiévet *et al.*, it was necessary to employ relative enzyme activities rather than enzyme concentrations. Depending on the specific activity of the enzyme preparation, the concentration of the enzyme represents a particular activity, which can vary from batch to batch. To account for this, the enzyme concentrations were taken listed in Fiévet *et al.* (see Table 4.6, indices 1-10) and transformed them into enzyme activities (Concentration based on the relative activity) by employing the specific activities indicated in the paper (Table 4.2). Then, the assessed specific activities were used for the enzymes (Table 4.3 and Table 4.8) and transformed the enzyme activities back into enzyme concentrations. The enzyme concentrations in Table 4.6 were used for the prediction (ANN predicted flux, J_{ANN} and simulated flux J_{copasi}) while the concentrations indicated in Table 4.8, index 11-41 were used for the experimental assessment of the flux.

Equation 4.1: where C, enzyme concentration (mg/l); Uv, enzyme activity per volume (U/ml); Us, specific enzyme activity (U/mg).

$$C\left(\frac{mg}{l}\right) = 1000 * \frac{U_v\left(\frac{U}{ml}\right)}{U_s\left(\frac{U}{mg}\right)}$$

Table 4.2: Specification of enzymes used for the calculation of cost for the preparatory stage of glycolysis from sigma. Specific activities are calculated by Fiévet *et al.* MW: Molecular weight.

Enzyme	Origin	Price (EUR)	Sold units (kU)	Specific activity (calc.) (U/mg)	Units (calc.) (kU)	MW (active enzyme) (kDa)
Phosphoglucoisomerase	Baker's yeast	78.50	1.0	1370.0	1.0	119.5
Phospho-fructokinase	*Bacillus stearothermophilus*	178.00	0.1	70.0	0.1	136.5
Fructose bisphosphate aldolase	rabbit muscle	48.75	0.1	42.0	0.1	157.4
Triose phosphate isomerase	Baker's yeast	146.00	n.a.	14690.0	50.0	53.6

Table 4.3: The measured enzyme activities for the enzymes involved in the upper part of glycolysis.

Enzyme	EC No	Origin	Specific activity (U/mg)	Comments
Phosphoglucoisomerase	5.3.1.9	Baker's yeast	556	this study
Phosphofructokinase	2.7.1.11	*Bacillus stearothermophilus*	73	this study
Fructose bisphosphate aldolase	4.1.2.13	Rabbit muscle	10	this study
Triosephosphate isomerase	5.3.1.1	Baker's yeast	9500	Manufacturer value

4.3.2 Cost Calculation

The cost for µM/s of flux through the pathway was estimated as follows:

Cost per 1U of enzyme: For each enzyme (PGI, PFK, FBA, TPI), the cost was calculated in Equation 4.2 as below using Table 4.2.

Equation 4.2: Where P_U, the price per unit.

$$P_U = \frac{Price\ (EUR)}{Units\ sold\ (U)}$$

Cost per reaction through the whole pathway: the cost per 1 ml of reaction is calculated as follows in Equation 4.3:

Equation 4.3: CR, cost per reaction; Uv, enzyme activity per reaction volume (U/ml).

$$C_R\ (EUR) = \sum (U_v * P_U)$$

Cost per one µM/s flux: Cost for the conversion of 1 µM NADH in 1 second is calculated using Equation 4.4:

Equation 4.4: Cflux, cost per flux of 1 µM/s; f, estimated flux (µM/s).

$$C_{flux}\ (EUR/\frac{\mu M}{s}) = \frac{C_R\ (EUR)}{f\ (\frac{\mu M}{s})}$$

4.4 Methodology

4.4.1 Data for New Methodology

The data from Fiévet *et al.* were used to develop the new methodology of exploring glass-ceiling to select optimum enzyme balances using ANN (GC-ANN) for the upper part of glycolysis. A balance being defined as a mixture of the four enzymes PGI, PFK, FBA and TPI. The dataset consisted of 121 combinations of four enzymes (PGI, PFK, FBA, TPI) of the glycolysis upper part for a flux value of 0.74 to 12.9 µM/s (Table 4.1). The total enzyme concentration was kept constant for four enzymes at 101.9 mg/l. The flux was measured as NADH consumption through G3PDH.

4.4.2 ANN-Based Flux Prediction Workflow

The GC-ANN methodology is explained in three steps i.) Preparatory stage: the data dimension is reduced to find the possibly correlated variable, the rule for obtaining higher flux (> 12 µM/s) is derived from the data and a neural network model is built to predict the flux using the enzyme balances ii.) Execution stage: the new enzyme balances are generated using the rule obtained and the flux is predicted for the new concentration using ANN. iii.) Validation of methodology: The ANN predicted flux was validated using simulation and experiments.

4.4.3 The Workflow of the Proposed Methodology

Based on the data listed in Fievet *et al.* (Fiévet *et al.*, 2006), the ANN model was built to predict the flux using enzyme balances, and the rule for enzyme balances with higher flux was obtained by data classification. The fluxes for newly generated enzyme balances were predicted using the ANN model. The balances with a flux value > 12 µM/s (balances from the glass-ceiling) and the balances obeying the derived rule for higher flux were selected as potential higher flux balances. These selected balances were validated using the kinetic model and by experiments. The methodology that followed for exploring the glass-ceiling of ANN (GC-ANN) is represented diagrammatically in Figure 4.5.

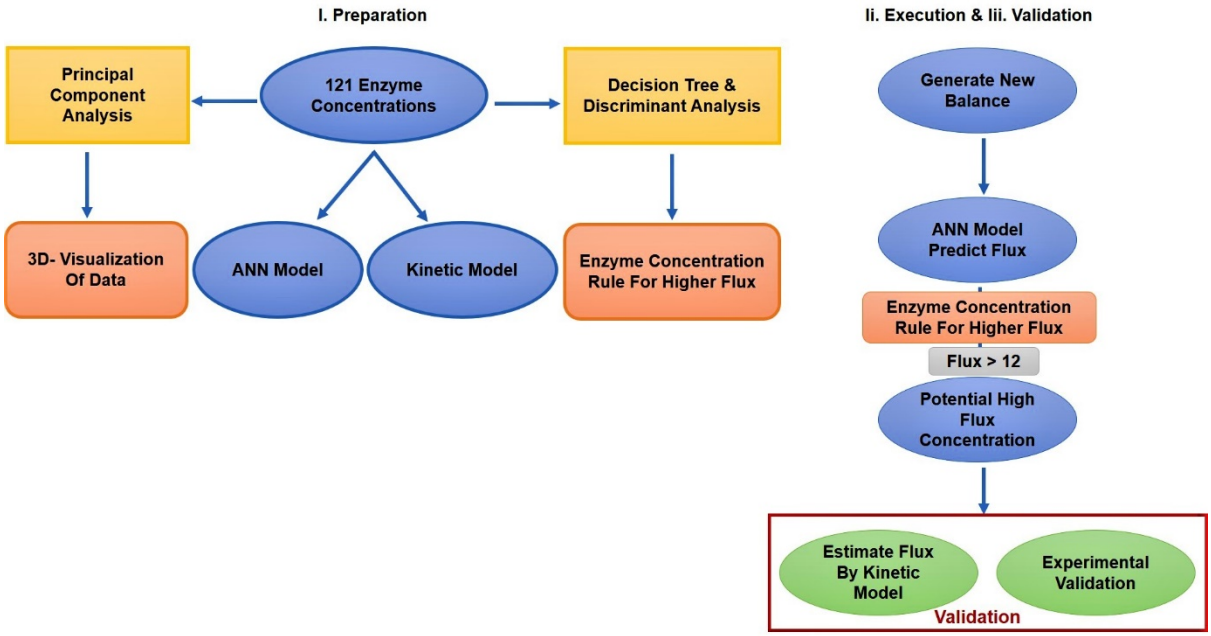


Figure 4.5: The methodology followed to obtain the new flux values from generated enzyme concentration.

4.4.3.1 Preparation Stage

4.4.3.1.1 Reduction of Dimensionality of Data

Principal component analysis (PCA) is one of the methods for the reduction of dimensionality of the dataset (Ringnér, 2008; Wold, Esbensen, & Geladi, 1987). For datasets with a high degree of freedom, PCA is useful to find possible correlations between the variables. The analysis of correlated variables helps find the relationship between the enzymes and the final flux. Finding correlated variables helps to understand the data in terms of flux distribution. PCA is performed using the R (V 3.4.3) (R Development Core Team (2008)) package FactoMineR (Le, Josse, & Husson, 2008).

4.4.3.1.2 Visualisation of Data

Three-dimensional viewing of data could provide insight into the distribution of flux in the space. Therefore, the fluxes in the 3D space of concentrations PGI, PFK, and TPI were visualized using R statistical packages plot3D (Soetaert, 2017) and plot3Drgl (Soetaert, 2016).

4.4.3.1.3 Classification of Data for Higher Flux (>12 µM/S):

Data classification is the process of categorizing data into various homogeneous groups or types based on common characteristics. Decision tree analysis is a method of data classification helping to search for possible associations within the dataset. The decision tree is a simple tree-like graph

method to understand and interpret the observations. The discriminant analysis helps to discriminate between the groups of data. The classification is supported by a discriminant analysis**.**

The data were classified into 5 groups, i.e., flux value from 0.728–3.17, 3.17–5.6, 5.6–8.04, 8.04–10.5 and 10.5–12.9. Approximately, 40% of the data are in the final group, which consists of higher flux concentrations (greater than 10.5 µM/s). The R packages klaR (Weihs, Ligges, Luebke, & Raabe, 2005) and rpart (Therneau & Atkinson, 2018) were used for discriminant analysis and decision tree respectively. The results from the decision tree and discriminant analysis were used to derive the concentration rule for higher flux values (> 12 µM/s) through the pathway.

4.4.3.1.4 Neural Network Model

The artificial neural network for predicting the flux through the upper part of glycolysis is built using the data described earlier in the section "Data for New Methodology". The model predicts flux as an NADH consumption through the pathway. The model is built using the R package neuralnet (Günther & Fritsch, 2010), which gives us the freedom to choose two different activation functions: logistic and tanh as explained in Chapter 3.

4.4.3.2 Execution Stage

4.4.3.2.1 Generation of New Enzyme Concentration

To explore the glass-ceiling space to obtain better balances with higher flux, new enzyme balances are generated. To limit the number of new balances generated, highest (PGI = 70, PFK = 70, FBA = 86.1, TPI = 66.1 mg/l) and lowest (PGI = 1, PFK = 1, FBA = 2, TPI = 1.66 mg/l) concentrations of the data from Fiévet *et al.* (Fiévet *et al.*, 2006) were used with the step size of 1 mg/l using R script. The total enzyme concentration of four enzymes was kept constant at 101.9 mg/l as in Fiévet *et al.* (Fiévet *et al.*, 2006). The newly generated concentrations are used in the additional analysis.

4.4.3.2.2 Flux Prediction Using ANN

Newly generated enzyme balances are fed to the ANN model to predict the flux. The data consisted of flux values ranging from 0.74 µM/s to 12.9 µM/s. Since ANN is not good for extrapolation, the prediction is limited to this range. Nevertheless, new enzyme balances could likely provide higher flux. However, ANN prediction will remain in the glass ceiling space. Hence, we decided to explore this space of squeezed flux, *i.e*. balances which lies in this particular space. Thus, for our study, fluxes >12 µM/s predicted by ANN and the balances which obeyed the rules derived from data classification were retained.

4.4.3.3 Validation of Methodology

The GC-ANN methodology of selection of enzyme balances is validated in two steps. In the first step, *the silico* model of the experimental system was built using the available kinetic parameters from Fiévet *et al.* (Fiévet *et al.*, 2006). In the second step, experiments were carried out to validate the selected balances.

4.4.3.3.1 Simulation of Upper Part of Glycolysis

In CellDesigner (ver 4.4) (Funahashi *et al.*, 2008, 2003), the kinetic model of the upper part of glycolysis was built using the kinetic parameters from Fiévet *et al.* (Fiévet *et al.*, 2006). The parameters for cofactors were chosen from the BRENDA (Schomburg *et al.*, 2002) database. The model was built to replicate the experimental condition with the Michaelis-Menten equation (Figure 4.6, Table 4.4). ATP is regenerated using the creatine kinase system. The hexokinase concentration was kept constant at 0.1 µM and flux was measured as NADH consumption, as catalysed by 1 µM of G3PDH. The concentrations of PGI, PFK, FBA and TPI are varied according to the selected balances from section "Flux Prediction Using ANN" (i.e. with concentrations which provide a flux ≥12 µM/s as predicted by the ANN model). The concentrations were converted from mg/l to µM using the molecular weight as suggested by Fiévet *et al.*

The model was simulated for 120 seconds using COPASI (Hoops et al., 2006) to measure NADH consumption. The slope of NADH decay between 60 to 120s was estimated as flux through the pathway. 182 enzyme balances yielding flux ≥15 µM/s from simulation using an *in silico* model were selected as the potential higher flux balances.

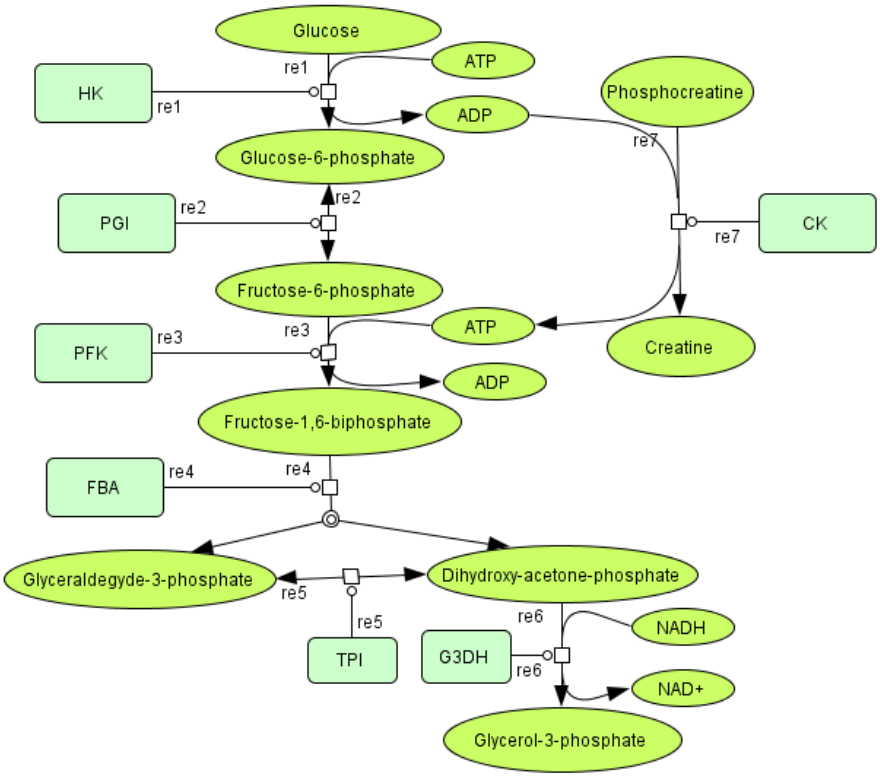


Figure 4.6: CellDesigner diagram for the upper part of glycolysis which replicates the experimental conditions described by Fiévet *et al.* (Fiévet *et al.*, 2006).

Table 4.4 The kinetic equations and parameters used to build the kinetic model of the upper part of glycolysis. Glu: glucose; G6P: glucose-6-phosphate; F6P: fructose-6-phosphate; FBP: fructose bisphosphate; DHAP: dihydroxyacetone phosphate. k_{cat}: turnover number in s^{-1}; K_m: Michaelis-Menten Constant in µM and K_i: inhibition constant in µM K_{eq}: equilibrium constant without units.

Reaction catalysed by	Kinetic Equation	Kinetic Parameters
Hexokinase (HK)	$v = \frac{\mathrm{kcat_{HK}} * HK * Glu * ATP}{(\mathrm{Glu} + \mathrm{KmGlucose}) * (\mathrm{ATP} + \mathrm{KmATP})}$	$kcat_{HK}$ = 72; $Km_{Glucose}$ = 120; Km_{ATP} = 100
Phosphoglucoisomerase (PGI)	$v = \frac{\left(\mathrm{kcat_{PGI_F}} * \mathrm{PGI} * \left(\frac{\mathrm{G6P}}{\mathrm{Kmg6p}}\right) - \mathrm{kcat_{PGI_R}} * \mathrm{PGI} * \left(\frac{\mathrm{F6P}}{\mathrm{KeqPGI} * \mathrm{Kmf6p}}\right)\right)}{\left(1 + \frac{\mathrm{G6p}}{\mathrm{Kmg6p}} + \frac{\mathrm{F6P}}{\mathrm{Kmf6p}}\right)}$	$kcat_{PGIF}$ =1410; $kcat_{PGIR}$ =3720; Kmg6p = 1650; Kmf6g = 4100; KeqPGI = 31
Phosphofructokinase (PFK)	$v = \frac{\mathrm{kcat_{PFK}} * \mathrm{PFK} * \mathrm{F6P^{nH}}) * \mathrm{ATP} /}{((\mathrm{Kmf6p^{nH}} + \mathrm{F6P^{nH}}) * (\mathrm{Km_{atp}} + \mathrm{ATP}))}$	kcat =41.7; Km_F6P = 33; nH = 1.1; Kmatp =120
Aldolase (ALD)	$v = \frac{\left(\mathrm{kcat_{ALDF}} * \mathrm{FBA} * \left(\frac{\mathrm{FBP}}{\mathrm{Km_{FrucBPhosp}}}\right) - \mathrm{kcat_{ALDR}} * \mathrm{FBA} * \left(\frac{\mathrm{glyc3pho} * \mathrm{DHAP}}{(\mathrm{Km_{gap}} * \mathrm{Km_{dhap}})}\right)\right)}{\left(1 + \frac{\mathrm{FBP}}{\mathrm{Km_{FrucBPhosp}}} + \frac{\mathrm{glyc3pho}}{\mathrm{Km_{gap}}} + \frac{\mathrm{DHAP}}{\mathrm{Km_{dhap}}} + \frac{\mathrm{FBP} * \mathrm{glyc3pho}}{(\mathrm{Km_{FrucBPhosp}} * \mathrm{Ki_{g3p}})} + \frac{\mathrm{glyc3pho} * \mathrm{DHAP}}{(\mathrm{Km_{gap}} * \mathrm{Km_{dhap}})}\right)}$	$kcat_{ALDF}$=7.59; $kcat_{ALDR}$=720; $Km_{FrucBPhosp}$ = 12; Km_{gap} =2000; Km_{dhap} = 2400 ;Ki_{g3p}= 10000
Triose-phosphate Isomerase (TPI)	$V = \frac{\mathrm{kcat_{TPI}} * \mathrm{TPI} * \mathrm{glyc3pho}}{\mathrm{Km_{gap}} + \mathrm{glyc3pho}}$	$kcat_{TPI}$ =6680; Km_{gap} = 238
Glycerol-3-phosphate dehydrogenase (G3PDH)	$v = \frac{\mathrm{kcat_{G3PDH}} * \mathrm{G3PDH} * \frac{\mathrm{DHAP}}{\mathrm{Km_{DHAP}}} * \frac{\mathrm{NADH}}{\mathrm{Km_{NADH}}}}{\left(1 + \frac{\mathrm{DHAP}}{\mathrm{Km_{DHAP}}} + \frac{g3p}{\mathrm{Km_{G3P}}}\right) * \left(1 + \frac{\mathrm{NADH}}{\mathrm{Km_{NADH}}} + \frac{\mathrm{NAD}}{\mathrm{Km_{NAD}}}\right)}$	$kcat_{G3PDH}$= 189.1 s^{-1}; Km_{DHAP} =75; Km_{G3P} = 909; Km_{NADH} = 22; Km_{NAD} = 83
Creatine kinase (CK)	$v = \frac{\mathrm{kcat_{CK}} * CK * \mathrm{phosphocreatine} * \mathrm{ADP}}{\left(\left(1 + \frac{\mathrm{phosphocreatine}}{\mathrm{Km_{PhosphoCrea}}} + \frac{\mathrm{Creatine}}{\mathrm{Km_{Creatine}}}\right) * \left(1 + \frac{\mathrm{ADP}}{\mathrm{Km_{ADP}}} + \frac{\mathrm{ATP}}{\mathrm{Km_{ATP}}}\right)\right)}$	$kcat_{G3PDH}$= 189.1; Km_{DHAP} =75; Km_{G3P} = 909; Km_{NADH} = 22; Km_{NAD} = 83

4.4.3.3.2 Experimental Validation

The upper part of glycolysis was reconstructed as described in Fiévet *et al.* (Fiévet *et al.*, 2006). The *in vitro* system consisted of varied concentrations of PGI, PFK, FBA and TPI. The HK and G3PDH were kept constant and creatine kinases were used to regenerate ATP in the system. The NADH decay was measured as flux through the pathway and the slope of the linear NADH decay was used to calculate the flux in μM/s.

4.5 Application and Results

4.5.1 Preparation

4.5.1.1 Data Dimension Reduction

PCA identifies new variables, the principal components, which are linear combinations of the original variables. This new variables can be used for further step in the methodology.

In our study, PCA does not provide much information regarding the data. The total four-enzyme concentration is constant in the system, which reduces the degree of freedom to limit the enzyme concentrations to three. . If total enzyme concentration had not been constant or the dataset presented a high degree of freedom, PCA would have been more useful for obtaining uncorrelated variables.

4.5.1.2 Visualisation of Data

After the PCA, data is visualised in 3D (Figure 4.7). In Figure 4.7 the quite distinct higher flux (red dots) can be observed. The distinction of flux in space indicates that a quantitative method could be applied to select the best balance for obtaining higher flux and should provide good results. Indeed, this is verified in the section “Flux Prediction Using ANN” (Figure 4.7). In this methodology, the space around those higher flux concentrations was explored to obtain new concentrations of PGI, PFK and TPI.

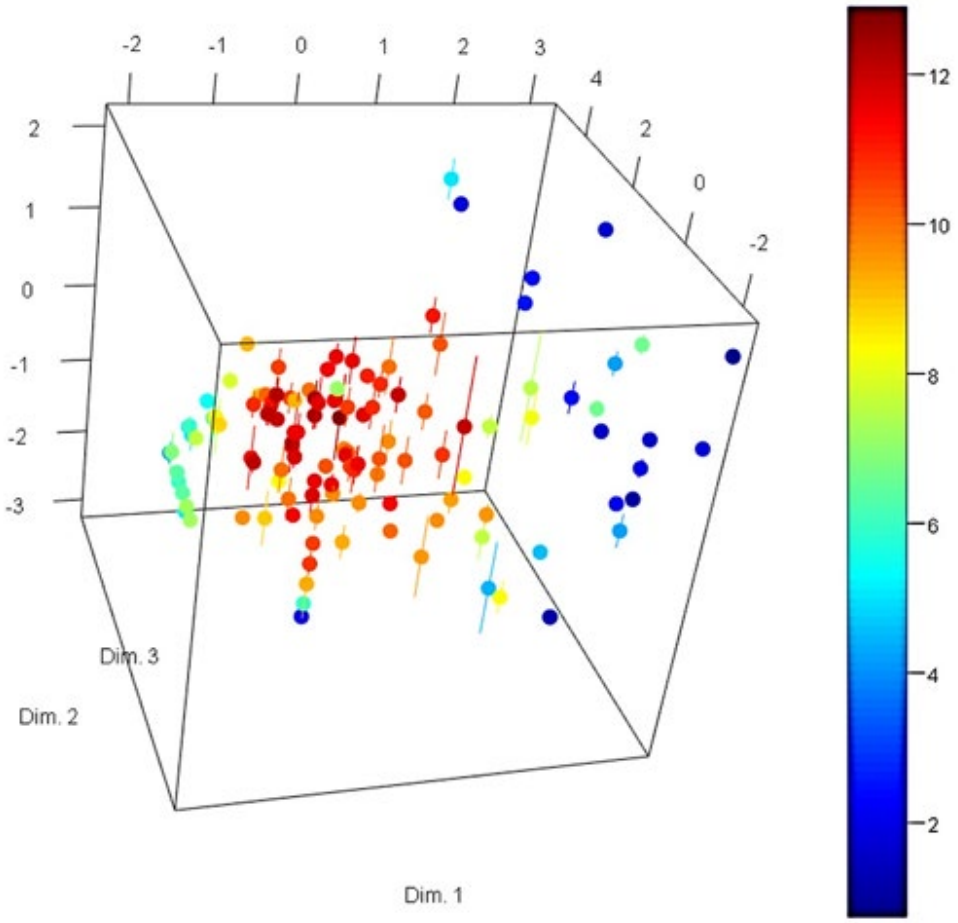


Figure 4.7: Three-dimensional visualization of Fievet *et al.* (Fiévet *et al.*, 2006) enzyme balances after PCA (Dim1: 43.55%; Dim2: 23.78% Dim 3: 17.56%). The change from blue to red indicates the gradient from low to high fluxes, respectively. Standard deviation of experimental flux is represented on the third-dimension.

4.5.1.3 Enzyme Concentration Rule

Decision tree analysis was performed using the R package, rpart, by dividing the data into five groups; this provides with the best compromise on the gain in inter-class inertia. The five groups are determined using K-means clustering. Figure 4.8 represents the classification of data where the percentage of data belongs to the branch of tree and fraction represents the distribution into different groups. For example, 89% of the data has FBA concentration >11 and distributed in five groups as fraction of 0.01, 0.09, 0.17, 0.29 and 0.44 (Figure 4.8, node 3).

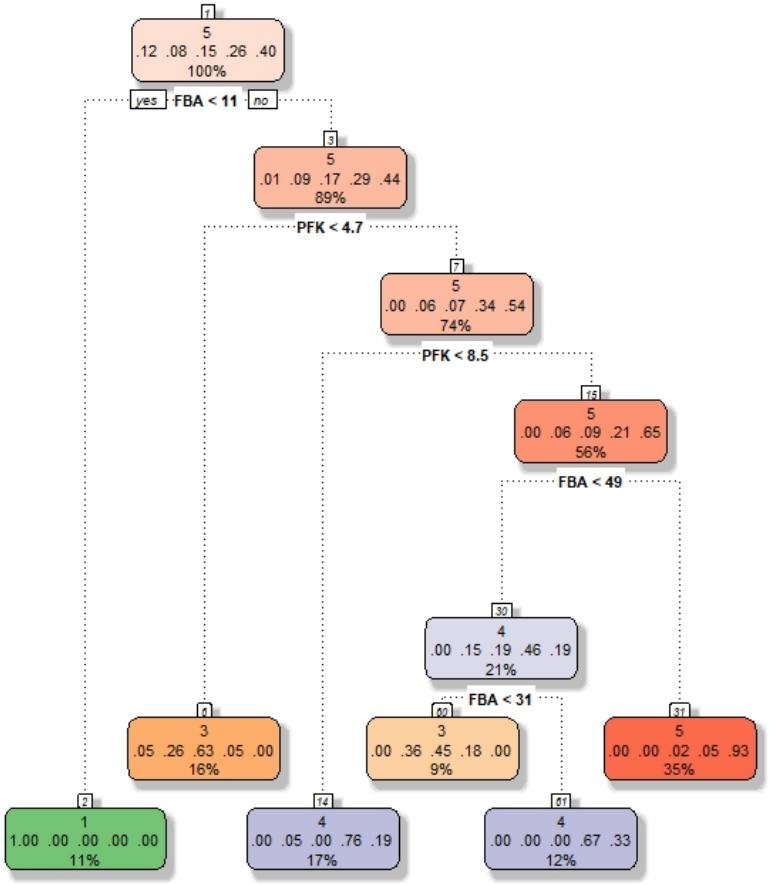


Figure 4.8: Decision tree for the Fiévet *et al.* (Fiévet *et al.*, 2006) data to obtain the rule for higher flux (≥12 µM/s). The data is classified into 5 groups (i.e., flux value from 0.728-3.17, 3.17-5.6, 5.6-8.04, 8.04-10.5, 10.5-12.9).

Among the different methods of discriminant analysis studied, rpart performed the best with an approximate error rate of 0.1. The different methods studied were lda (linear discriminant analysis), qda (quadratic discriminant analysis), sknn (simple k nearest neighbours), rda (regularized discriminant analysis) and naïve Bayes (under R package). For the sknn method, the error rate is low but it leads to an over-classification (

Annexe 2**).** Figure 4.8 represents the discriminant analysis for the classification of data from Fievet *et al.* (Fiévet *et al.*, 2006) using the rpart (Therneau & Atkinson, 2018) method from R.

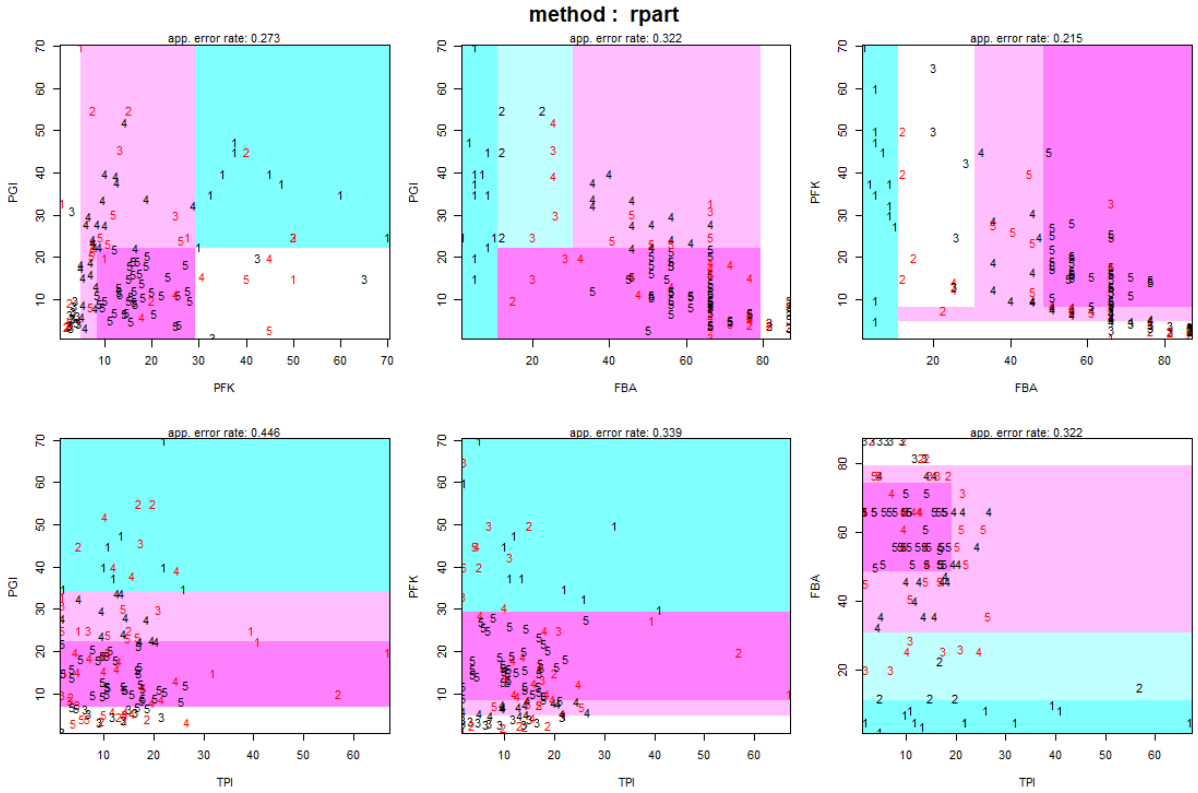


Figure 4.9: Discriminant analysis for the classification of data from Fievet *et al.* (Fiévet et al., 2006) using the rpart (Therneau & Atkinson, 2018) method from R. Color code according to the feature space of data, where group 1 (flux: 0.728-3.17 µM/s) is shown in light blue, group 2 (flux: 3.17-5.6 µM/s) in dark blue, group 3 (flux: 5.6-8.04 µM/s) in white, group 4 (flux: 8.04-10.5 µM/s) in light pink and group 5 (flux: 10.5-12.9 µM/s) in dark pink. Numbers in black represent the data classified to the same group, and in red represent data misclassified into the other groups.

After using the decision tree and discriminant analysis, the following rule was derived to obtain a flux ≥12 µM/s:

PGI < 11; 10 < PFK < 16; TPI < 18; 59 >FBA (mg/l) which corresponds to

PGI < 15.07 U/ml

0.7 U/ml > PFK < 1.12 U/ml

TPI < 264.42 U/ml

2.48 U/ml > FBA.

The conversion from mg/l to U/ml is given in 4.3.1. The derived rule is applied for the selection of the best concentrations of the enzymes PFK, PGI, TPI and FBA to obtain a high flux through the pathway.

4.5.1.4 Neural Network Model

ANN is a training-based method, the structure of the neural network needs to be chosen carefully since it depends on the number of inputs, sampling in the training dataset and the outputs. The structure was determined based on our previous study as described in Chapter 3. The neuralnet package from R statistical tool with the logistic activation function was used. It has 13 hidden units in a single layer. The ANN model used has an RMSE value of 0.84 and an R^2 value of 0.93, using leave-one-out cross-validation (Chapter 3).

4.5.2 Execution

4.5.3 Generation of New Enzyme Concentrations

The new concentrations of PFK, PGI, TPI and FBA were generated as explained in the methodology section. These new balances were used for further analysis to predict the flux.

4.5.3.1 Flux Prediction Using ANN

The new balances were fed into the previously built neural network to predict the flux. The ANN predicted flux from the newly generated data was visualised in 3-dimensions (Figure 4.10).

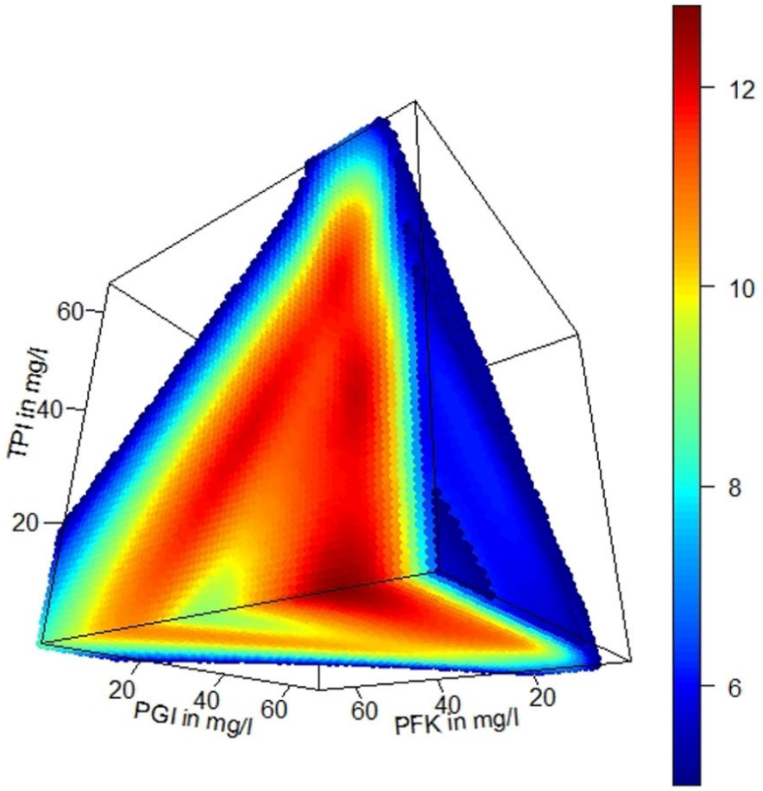


Figure 4.10: Three-dimensional visualisation of flux predicted by ANN for newly generated enzyme concentration. The colour gradient is from the lowest (blue) to the highest (red) predicted flux.

As expected, the new prediction remained in the box (see the maximum value of the colour gradient bar in Figure 4.10) since the ANN is a training-based method, which depends on the training dataset. The high predicted values bump into the "glass ceiling". Hypothesise that even though they remain in the roof of the "glass ceiling", the experimental values could be higher than the predicted ones. By exploring this space, new balances with higher flux values could be obtained.

To explore the "glass ceiling" space, the new methodology (GC-ANN) using the artificial neural network was developed to predict the flux through the upper part of glycolysis for given enzyme balances. This study showed that (see below in the section Validation) by careful selection of enzyme balances from the "glass ceiling" space, it is possible to obtain higher flux values "out-of-the-box".

For all the enzyme balances generated between minimum and maximum of experimental data, only flux values above 12 µM/s predicted by the neural network and only enzyme balances

(total of 335 balances) obeying the enzyme concentration rules are selected as potential high-flux balances.

4.5.4 Validation

The methodology for exploring the glass-ceiling using ANN (GC-ANN) was validated in two steps: first using the kinetic model and second, *in vitro*.

4.5.4.1 Simulation of Upper Part of Glycolysis

The kinetic model is built using CellDesigner (Funahashi *et al.*, 2008, 2003) (Figure 4.6) and validated with COPASI (Hoops et al., 2006) using the 121 concentrations from Fiévet *et al.* (Fiévet *et al.*, 2006). The model has an RMSE value of 1.58 and R^2 of 0.84 in a cross-validation procedure, compared to the experimentally determined flux (Figure 4.11). The highest flux predicted by the kinetic model of the reconstituted upper part of glycolysis was 14.93 µM/s, where the highest experimentally observed flux was 12.9 µM/s. The flux predicted by ANN for new enzyme balances from the section "Flux Prediction Using ANN" was compared with the simulated flux for each enzyme (Figure 4.12). Figure 4.12 shows that the balances which were predicted with higher flux through GC-ANN were also estimated to have higher flux using the kinetic model. This validates the good quality of the kinetic model.

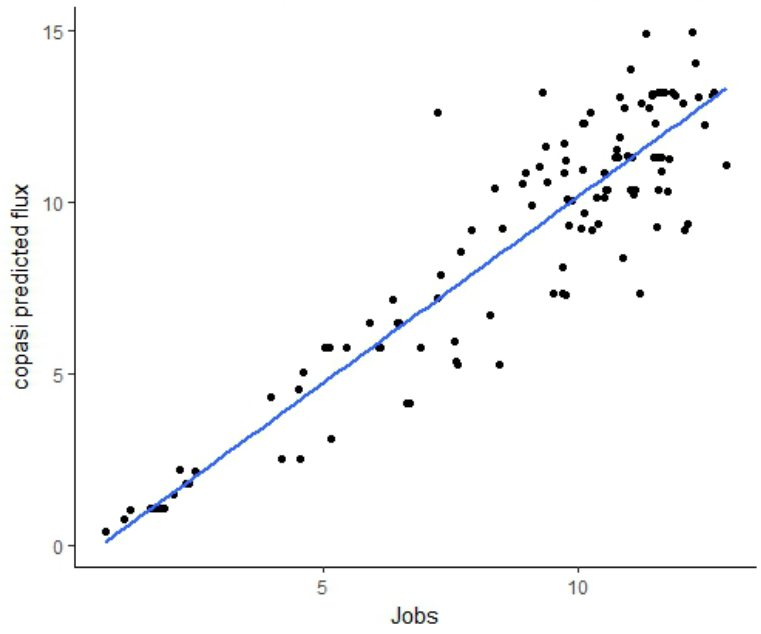


Figure 4.11: Relationship between experimental flux (JFievet) estimated by Fiévet *et al.* (Fiévet *et al.*, 2006) and COPASI (Hoops et al., 2006) estimated by the kinetic model.

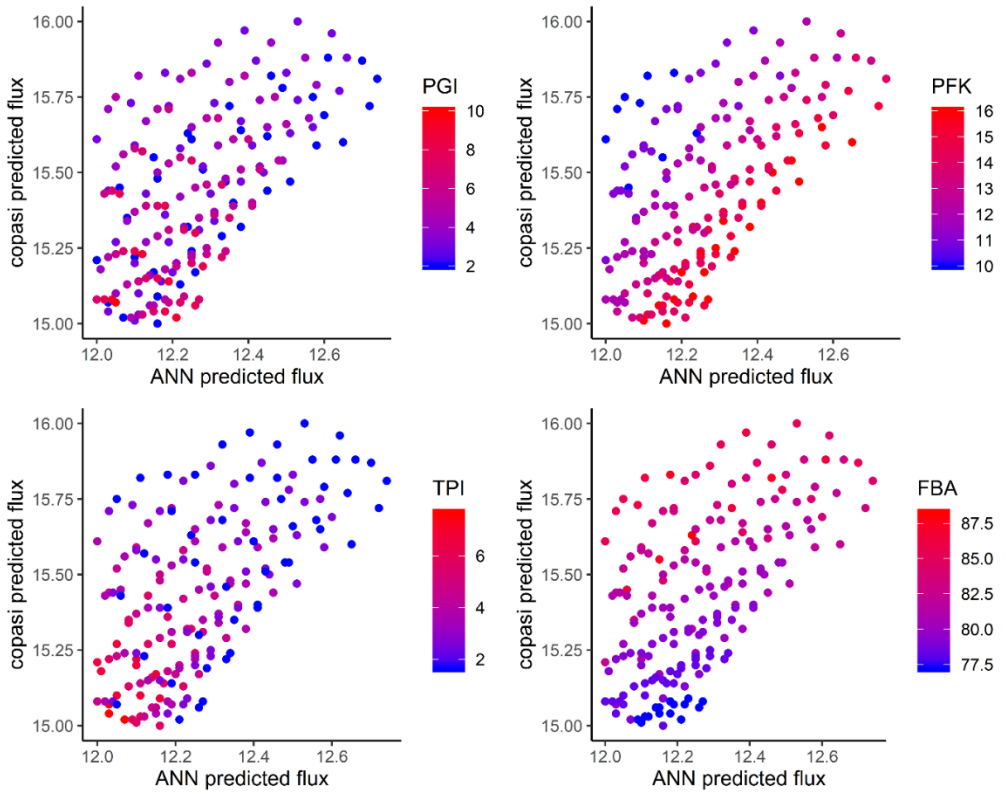


Figure 4.12: The relationship between flux values predicted by ANN vs COPASI for newly generated enzyme balances. The enzymes considered are: upper, left (PGI), right (PFK), lower left (TPI), right (FBA). The colour gradient from blue to red represents the particular enzyme concentration from low to high, respectively.

4.5.4.2 Experimental Validation of the Methodology

To validate GC-ANN approach to exploring the glass-ceiling (GC-ANN), the new enzyme balances generated were assayed *in vitro*. For the control experiment, 10 enzyme balances from previously used by Fiévet *et al.* (Fiévet *et al.*, 2006) was selected (Figure 4.13, Table 4.6). These selected balances were with a correlation R^2 of 0.99 and an RMSE of 0.17 between the predicted flux from kinetic model and the experimental flux assessed by Fiévet *et al.* (Fiévet *et al.*, 2006) (Figure 4.13). Figure 4.13 shows that balances selected for the control study are an appropriate choice. Two of these selected Fievet's balances were tested experimentally. The resulting fluxes for these two balances were 0.59 (±0.10) µM/sand 8.03 (±0.56) µM/s while Fievet *et al.* had determined 1.22 (±0.08) µM/s and 11.05 (±0.29) µM/s, respectively. Two of Fiévet's balances were tested experimentally (Table 4.5). The resulting fluxes for these two

balances were 0.59 (±0.10) µM/s and 8.03 (±0.56) µM/s while Fiévet *et al.* had determined 1.22 (±0.08) µM/s and 11.05 (±0.29) µM/s, respectively.

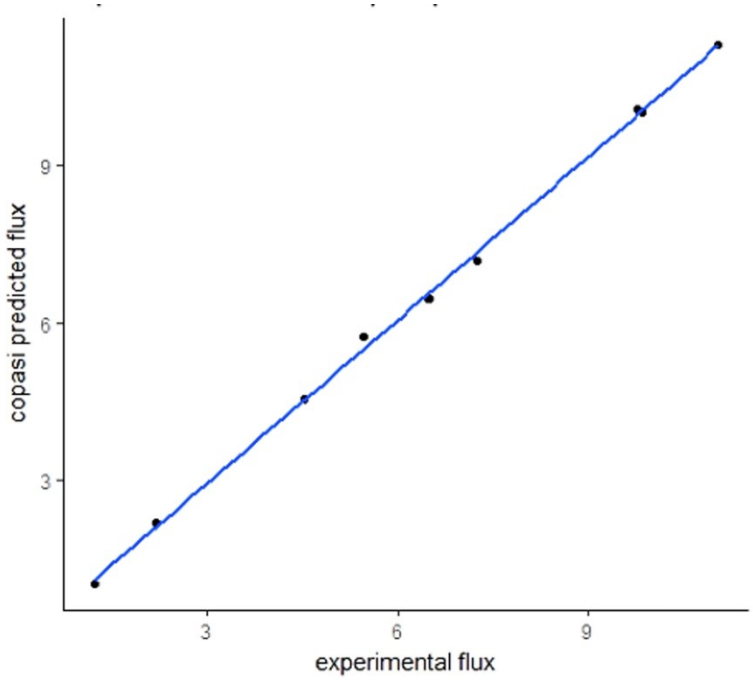


Figure 4.13: Correlation between Fiévet *et al.* (Fiévet *et al.*, 2006) experimental flux and Copasi predicted flux. The balances corresponding to these flux values are selected as the experimental control.

Table 4.5: Comparison of flux predicted between Fiévet *et al.* selected concentration (JFievet) and new estimation during current work (Jobs).

Index	**PGI (mg/l)**	**PFK (mg/l)**	**FBA (mg/l)**	**TPI (mg/l)**	**J_{obs} [µM s^{-1}]**	**J_{Fievet} [µM s^{-1}]**
1	70.0	5.0	5.0	21.9	0.59	1.22
10	19.19	17.0	56.4	9.67	8.03	11.05

31 new balances were selected from the GC-ANN approach (Figure 4.14; Table 4.6) for experimental validation. The flux values associated with the selected balances had a coefficient of determination R^2 of 0.44, between GC-ANN predictions and simulated flux. This low R^2 between ANN and Copasi prediction is due to the glass-ceiling effect: the underestimation of the flux due to the inability to obtain "out-of-the-box" values for the ANN was expected.

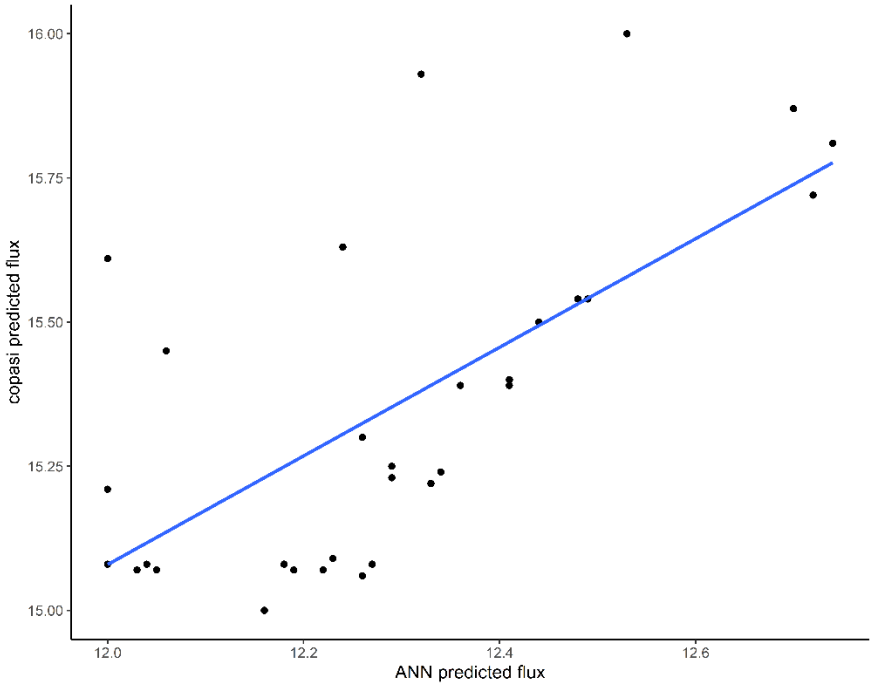


Figure 4.14: Comparison between GC-ANN predicted flux and simulated flux. The enzyme balances corresponding to these flux values are selected for experimental validation of the methodology.

Table 4.6: The enzyme concentrations (mg/l) predicted from ANN and in-silico modelling to have higher flux values. For the experimental validation, relative concentrations of enzymes obtained were used.

Index	**PGI mg/l**	**PFK mg/l**	**FBA mg/l**	**TPI mg/l**	**Index**	**PGI mg/l**	**PFK mg/l**	**FBA mg/l**	**TPI mg/l**
1	70	5	5	21.9	**22**	4	11	85.24	1.66
2	33	1	66.23	1.66	**23**	4	16	80.24	1.66
3	55	7.5	22.5	16.9	**24**	4	16	79.24	2.66
4	4.23	2.62	76.42	18.62	**25**	5	15	80.24	1.66
5	7.36	3.21	86.61	4.72	**26**	5	16	79.24	1.66
6	5.38	3.01	86.61	6.9	**27**	5	16	78.24	2.66
7	8.62	3.47	86.61	3.19	**28**	5	16	77.24	3.66
8	28.1	8.42	50.95	14.43	**29**	6	15	79.24	1.66
9	22.56	7.84	50.95	20.55	**30**	6	15	78.24	2.66
10	19.19	17	56.04	9.67	**31**	6	15	77.24	3.66

11	2	10	88.24	1.66	**32**	6	16	78.24	1.66
12	2	10	86.24	3.66	**33**	6	16	77.24	2.66
13	2	11	82.24	6.66	**34**	7	12	78.24	4.66
14	2	12	80.24	7.66	**35**	7	15	78.24	1.66
15	2	13	85.24	1.66	**36**	7	15	77.24	2.66
16	2	14	84.24	1.66	**37**	7	16	77.24	1.66
17	2	15	83.24	1.66	**38**	8	13	79.24	1.66
18	2	16	78.24	5.66	**39**	8	15	77.24	1.66
19	3	10	85.24	3.66	**40**	9	12	78.24	2.66
20	3	12	85.24	1.66	**41**	10	12	78.24	1.66
21	3	16	80.24	2.66					

4.5.4.2.1 Enzyme Assays for Measurement of Kinetic Parameters

HK activity was assessed using glucose-6-phosphate dehydrogenase (G6PDH) in a coupled reaction. The substrate glucose was converted to 6-phosphogluconate, the formation of NADPH was followed spectrophotometrically at 340 nm (Figure 4.15A).

the activities of PGI, PFK and FBA were assessed using a coupled NADH assay applied to the upper part of glycolysis (Figure 4.15B). To determine the activity of PGI, the assay was started with glucose-6-P (Figure 4.15B, reaction 1); for the measurement of the activities of PFK and FBA, fructose 6-P and fructose 1,6-bisP were used as the substrates (Figure 4.15B, reactions 2 and 3).

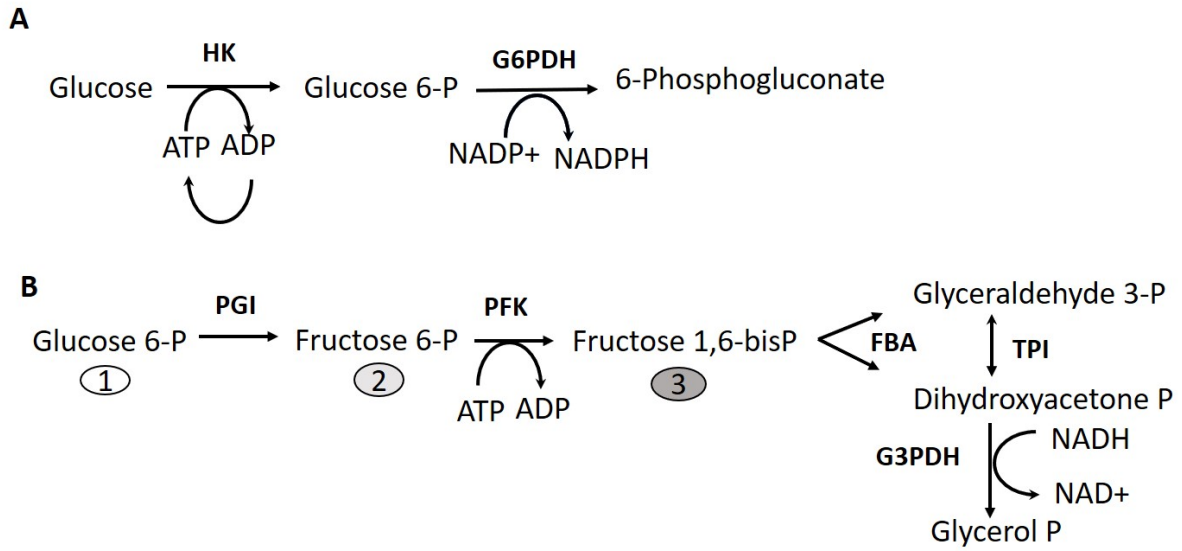


Figure 4.15: A) Coupled HK/G6PDH assay to assess the HK activity. (B) Coupled NADH assay to assess the activities of PGI, PFK and FBA. The individual reactions were started with substrates indicated by the numbers in circles.

All reactions were monitored by reading the absorbance of NADH at 340 nm and the initial rates were used to calculate the Michaelis constant K_m and the maximal velocity V_{max}. The kinetic parameters K_m and V_{max} for HK, PGI, PFK and FBA corresponded well to the values listed by the manufacturer (Sigma) or by the Enzyme Database Brenda (Table 4.7). Nevertheless, some enzymes, particularly HK and FBA, showed lower specific activity compared to the Sigma reference. The loss of activity could have occurred during delivery and/or storage of the enzymes or could be attributed to a different enzyme assay.

Table 4.7: Summary of the kinetic parameters of HK, PGI, PFK and FBA. The experimentally assessed values were deduced from Lineweaver-Burk and Eadie-Hofstee plots. Reference values for K_m and V_{max} from Brenda and Sigma's product data sheets are indicated, respectively.

		Reference Sigma	**This study**	**Reference Brenda**	**Lineweaver-Burk**			**Eadie-Hofstee**		
Enzyme	**Lot No.**	**sp. act. [U/mg]**	**sp. act. [U/mg]**	**K_m [mM]**	**K_m [mM]**	**V_{max} [U/ml]**	**k_{cat} s^{-1}**	**K_m [mM]**	**V_{max} [U/ml]**	**k_{cat} s^{-1}**
HK	SLBT5451	472	163	0.12-0.5 ("BRENDA - Information on EC 2.7.1.1 - hexokinase," n.d.)	0.28	225.5	299	0.30	248.7	330
PGI	SLBW8689	618	556	0.084-1.5 ("BRENDA - Information on EC 5.3.1.9 - glucose-6-phosphate isomerase," n.d.)	1.1	7409	1107	0.9	7685	1147
PFK	SLBW6641	72	73	0.023-0.15 ("BRENDA - Information on EC 2.7.1.11 - 6-phosphofructokinase," n.d.)	0.13	196	166	0.11	206	175
FBA	SLBR7752V	11.5	6.4	0.00084-2 ("BRENDA - Information on EC 4.1.2.13 - fructose-bisphosphate aldolase," n.d.)	0.14	19.6	17	0.12	18.7	16
	SLBV7445	12.4	10		n.d.	n.d.	n.d.	n.d.	n.d.	n.d.

4.5.4.2.2 Flux Determinations

The reaction mixtures for the measurements of the flux through the upper part of glycolysis were based on Fiévet *et al.* (Fiévet *et al.*, 2006) (Table 4.8). In contrast to Fiévet *et al.*, our mixtures were based on relative enzyme activities rather than enzyme concentrations. Calculations are explained in section Concentration Based on the Relative Activity.

Table 4.8: Comparison of ANN predicted flux (JANN in µM/s), simulated flux (J_{Copasi} in µM/s) and experimentally assessed flux (J_{Exp} in µM/s). The four enzymes PGI, PFK, FBA and TPI were used at the indicated concentrations for the experimental assessment of the flux with mean deviation (M.D) of triplicates.

Index	U/ml				µM/s			
	PGI	PFK	FBA	TPI	J_{ANN}	J_{Copasi}	J_{Exp}	M.D
11	2.74	0.7	3.71	24.39	12.24	15.63	15.7	2.5
12	2.74	0.7	3.62	53.77	12.06	15.45	16.3	2.7
13	2.74	0.77	3.45	97.84	12	15.21	12.1	4.2
14	2.74	0.84	3.37	112.53	12.03	15.07	16.6	0.1
15	2.74	0.91	3.58	24.39	12.7	15.87	13.9	3.9
16	2.74	0.98	3.54	24.39	12.74	15.81	18.3	1.2
17	2.74	1.05	3.50	24.39	12.72	15.72	17.1	0.2
18	2.74	1.12	3.29	83.15	12.16	15	20.1	0.3
19	4.11	0.7	3.58	53.77	12	15.61	14.4	0.1
20	4.11	0.84	3.58	24.39	12.53	16	15.8	0.2
21	4.11	1.12	3.37	39.08	12.44	15.5	20.6	0.2
22	5.48	0.77	3.58	24.39	12.32	15.93	15.4	0.2
23	5.48	1.12	3.37	24.39	12.49	15.54	16.1	2.3
24	5.48	1.12	3.33	39.08	12.36	15.39	19.3	0.6
25	6.85	1.05	3.37	24.39	12.48	15.54	18.5	0.6
26	6.85	1.12	3.33	24.39	12.41	15.4	17.8	0.1
27	6.85	1.12	3.29	39.08	12.29	15.25	16.3	0.3
28	6.85	1.12	3.24	53.77	12.18	15.08	19.7	2.5
29	8.22	1.05	3.33	24.39	12.41	15.39	17.8	1
30	8.22	1.05	3.29	39.08	12.29	15.23	19	0.6

31	8.22	1.05	3.24	53.77	12.19	15.07	21	0.6
32	8.22	1.12	3.29	24.39	12.34	15.24	15.6	3.1
33	8.22	1.12	3.24	39.08	12.23	15.09	17.8	2.2
34	9.59	0.84	3.29	68.46	12	15.08	17.1	0.7
35	9.59	1.05	3.29	24.39	12.33	15.22	17.7	1
36	9.59	1.05	3.24	39.08	12.22	15.07	18.8	1.8
37	9.59	1.12	3.24	24.39	12.27	15.08	20.4	0.6
38	10.96	0.91	3.33	24.39	12.26	15.3	15.9	0.9
39	10.96	1.05	3.24	24.39	12.26	15.06	17.9	0.8
40	12.33	0.84	3.29	39.08	12.04	15.08	15.8	0.9
41	13.7	0.84	3.29	24.39	12.05	15.07	13.6	2.4

Out of 41 selected balances, 31 newly predicted enzyme balances were tested experimentally to estimate flux. All 31 new enzyme balances experimentally tested were estimated with flux values greater than 12 µM/s (Table 4.8). Table 4.8 shows that 28 out of 31, *i.e.* 90.3%, have a value above15.0 µM/s, as expected according to the kinetic model. Moreover, 31 out of 31, *i.e.* 100%, have a value up to 12.0 µM/s, as expected according to our methodology.

4.5.5 Application: Selection of Cost-Efficient Enzyme Balances

For industrial-scale production, the selection of the best enzyme balances in terms of cost is essential. Therefore, the cost per µM of NADH consumed per second were estimated for all the enzyme balances generated (Figure 4.16) and for those selected balances from ANN prediction, which obey the enzyme concentration rule (flux greater than 12 µM/s), i.e., 335 balances from Flux Prediction Using ANN (Figure 4.17). The calculations are described in section Cost Calculation. The cost calculation for each reaction observed in the selection of enzymes could help to reduce cost. Figure 4.16 and Figure 4.17 show the variation in cost according to each balance and its flux and allow the selection of balances with higher flux at low cost.

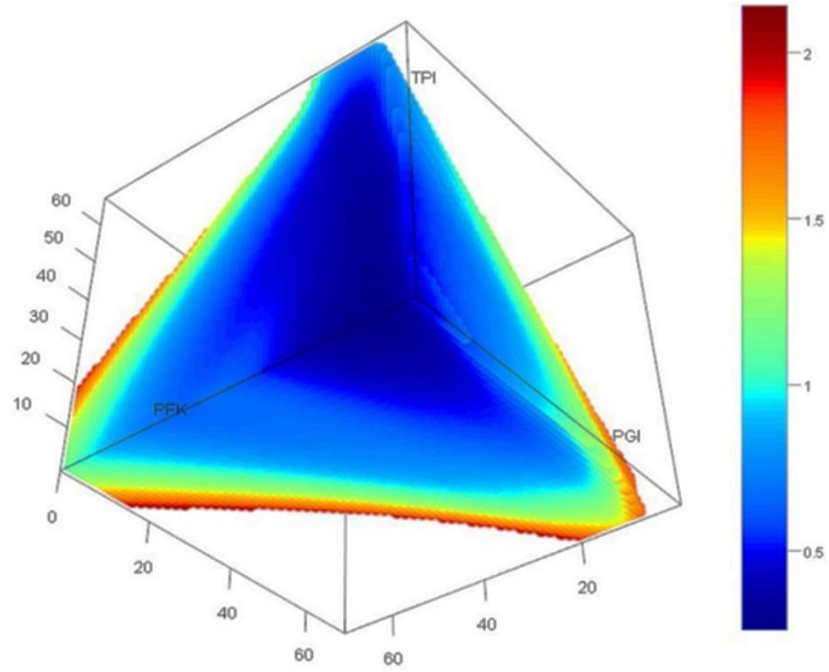


Figure 4.16: 3D-representation of the cost estimated for all the enzyme balances generated. The colour gradient is according to the cost required for each balance: blue is the lowest and red is the highest cost for a selected balances of the four enzymes PGI, PFK, FBA and TPI.

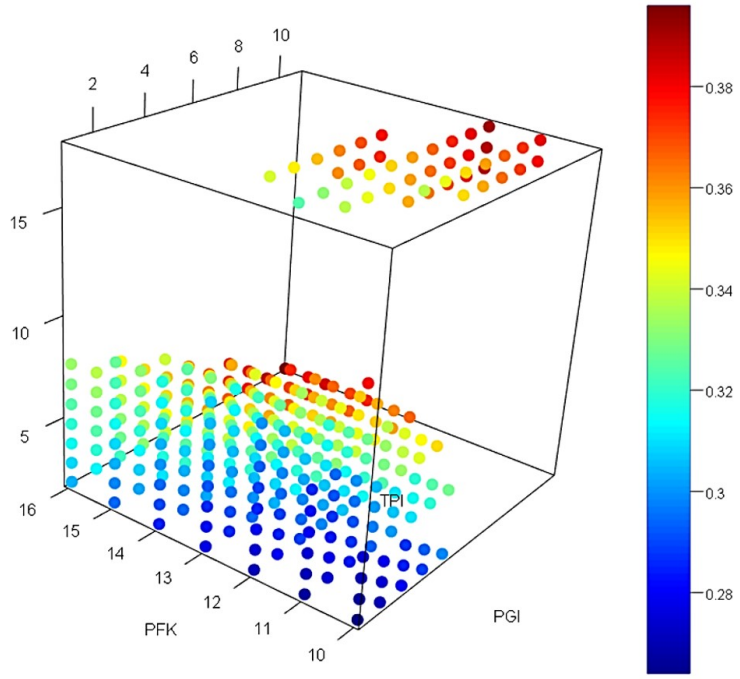


Figure 4.17: 3D-representation of the cost estimated for the enzyme concentration which obeys the rule obtained for higher flux values. The colour gradient is according to the cost required for each balance, blue is the lowest and red is the highest cost for a selected balance of the four enzymes PGI, PFK, FBA and TPI.

As an example: the enzyme balance (in mg/l) with PGI = 2, PFK = 12, FBA = 81.24 and TPI = 4.66 (index 13 in Annexe 3) could give a flux of 12.1 μM/s with a cost of 3.79 EUR.

4.6 Discussion

Traditionally, chemical molecules are synthesised by the chemical reaction of petroleum-based products. Due to the depletion of petroleum products, *in-vivo* biosynthesis has gained a lot of attention. Limitations of the cellular production system, such as low productivity, by-product formation and low host cell tolerance to toxins moved the focus towards development of cell-free systems. Compared to cell systems, cell-free systems have high productivity and high toxin tolerance (Lu, 2017). The selection of optimal enzyme balances for maximal productivity is a crucial step for industrial scale, cell-free production of biomolecules. The modelling of metabolic pathways helps to study and predict the behaviour of the biological system. Constraint-based methods facilitate the understanding of the system but do not provide information about the concentration of the individual metabolites. In contrast, kinetic models provide information about individual metabolite concentrations but require kinetic parameters of enzymes, which are tedious and expensive to determine (Bisswanger, 2014). Design of experiment (DOE) is a systematic approach to optimise the conditions for biomolecule production in the field of biotechnology (V. Kumar, Bhalla, & Rathore, 2014). In DOE, multiple variables are studied to find the correlation between the variables and the final outcome. The main objective of DOE is to reduce the number of experiments, time and cost; our study has the same objective. The benefit of GC-ANN is that the objective optimum can be "out-of-the-box" but will nevertheless be found without additional experiments.

4.6.1 GC-ANN Approach could be Used to Predict Out-of-the-Box Values

In this study, a new methodology, GC-ANN, to select the optimum enzyme balances for industrial biotechnology is devised. This approach aims to see beyond the "glass ceiling", using an artificial neural network and different statistical methods like PCA and data classification. The method was designed and validated for the upper part of glycolysis but could be applied to any other natural or reconstituted biosynthesis pathway.

The workflow of the methodology used in the upper part of glycolysis is summarised in Figure 4.5. In the first step, for selecting the optimum concentrations of the four relevant enzymes PGI,

PFK, FBA and TPI, a rule was devised for high flux values (supported by Figure 4.7-Figure 4.9). All the possible balances were generated using a step of 1 mg/l in terms of variation of each enzyme concentration. The balances newly generated in the present study have higher and lower limits than those in Fiévet *et al.* (Fiévet *et al.*, 2006). When these new enzyme balances were used to predict the flux through the upper glycolysis using ANN and the predicted fluxes were depicted in 3D representation (Figure 4.10), a zone (Figure 4.10, brown zone) with predicted flux >12 µM/s was observed. To explore this space in order to obtain even higher fluxes, the high-flux-rule was applied, *i.e.* 10 < PFK < 16; PGI < 11; TPI < 18; 59 < FBA (in mg/l) and 335 enzyme balances were scrutinized. The main idea behind our approach is based on the fact that *i).* ANN is known to be a good tool for predicting class and/or quantitative values inside the box (*i.e.* prediction close to training data), *ii)*. the brown region in Figure 4.10 contains values that are all very close to 12 µM/s (from 12 µM/s to 12.9 µM/s) because ANN is not good for extrapolation and new predictions remain inside the box; *iii)*. but we postulate that among these flux values some could be higher than predicted.

In the second step, to validate our hypothesis *in silico* and *in vitro* experiments were conducted:

4.6.2 *In Silico* Validation

Due to the availability of kinetic parameters, to avoid unnecessary expenses linked to *in vitro* assays:

- First, a kinetic model was built. Figure 4.11 shows good agreement (R2=0.84) between the fluxes predicted by the kinetic model and all the flux values experimentally assessed by Fievet *et al.* (Fiévet *et al.*, 2006). Then, 10 balances associated with experimental values between 0.74 to 12.9 µM/s of Fievet's data for the benchmark study were selected. Figure 4.12 excellent correlation with R^2 of 0.99 and an RMSE of 0.17 between the predicted flux from our kinetic model and the experimental flux assessed by Fievet *et al.* Taken together, these first results were a good validation of our kinetic model.

- Second, to validate our *in vitro* assay by reproducing the results obtained by Fiévet *et al.* (Fiévet *et al.*, 2006) *in vitro* experiments for the balances that had a good correlation between simulated and experimental flux were carried out. The experimentally determined fluxes using the balances selected from the Fievet data were lower than those previously determined by these authors (Table 4.5). Nevertheless, the fold-increase was comparable (approximately 9-fold, this study *vs* 13-fold, Fiévet *et al.* (Fiévet *et al.*, 2006)). The deviation of the absolute flux values could be attributed to experimental settings, i.e. NADH depletion assay in cuvettes at 390 nm (Fiévet *et al.*, 2006). *vs* in

96 well plates at 365 nm, this study; or differences in the assays performed to measure kinetic parameters of the individual enzymes.

Finally, as our kinetic model is validated, it was used to conduct the first verification, *in silico*, of our hypothesis. For 31 new balances selected according to the methodology described above section "Experimental Validation of the Methodology". Figure 4.14 shows how flux values predicted by the kinetic model fit with the simulated values. All the balances selected from the brown zone (Figure 4.10) are indeed superior to 12.0 µM/s. Moreover, the flux should be above 15.0 µM/s. So, this is a first, *in silico*, validation of our hypothesis, *i.e*. the ANN-based approach could be used to predict out-of-the-box values.

At this point, we have to keep in mind that this preliminary verification was conducted because the kinetic model was possible to establish, but this step is not mandatory in the proposed methodology. Indeed, the 31 balances were chosen first, based only on the outcome of GC-ANN methodology that combines ANN and different statistical methods like PCA and data classification.

4.6.3 *In Vitro* Validation

The 31 new enzyme balances were assessed experimentally. Table 4.8 proves our hypothesis: with careful selection of enzyme balances from the glass ceiling, it is possible to obtain higher flux values. For the 27 best enzyme balances, the improvement of flux ranges from 30% (observed flux: 15.4, original flux: 12) to 70% (observed flux: 21.0, original flux: 12). This clearly demonstrates that exploring the predicted values which hit the "glass ceiling" using the GC-ANN approach is a good way to select the optimum enzyme concentration.

Since artificial neural networks do not require much information regarding the experimental conditions and particularly in our case kinetic parameters hard to obtain, they are easy to apply in different fields of science. Our GC-ANN approach could be applied on any pathway provided the experimental data are available.

4.6.4 The Proposed Methodology is Cost-Efficient

From an industrial perspective, the production costs per quantity of product are very important. Choosing an enzyme balance that results in maximum flux at a very low cost per given quantity of product is essential. The ANN-based methodology makes it easy to estimate the total cost. The approximate price for each reaction is calculated using the details provided by the manufacturer, such as specific activity and units of enzyme in the sample. The approximate cost required for 1

µM of product formation per second through the pathway can be calculated. This will help to decide which is the most suitable enzyme balance for maximum flux in terms of cost minimisation, which is important for industrial-scale production. For example, to obtain a flux of 12.1 µM/s, the approximate cost should be 6.28 EUR, whereas the same flux value with a cheaper rate of 3.79EUR (~40%) could be achieved. Figure 4.16 clearly shows how costs vary. Details are provided in Annexe 3 and Annexe 4. Among the enzyme combinations selected for the validation of our methodology, PGI = 3, PFK= 16, FBA = 80.24, TPI= 2.66 (mg/l) have an estimated flux value of 20.6 µM/s with the lowest cost of 0.197 EUR per µM of NADH consumed per second using GC-ANN methodology for the selection of enzyme balances (Annexe 5). In contrast, the lowest price in Fiévet et al. (Fiévet et al., 2006) with the selected balance PGI = 7, PFK= 12, FBA = 66.23, TPI= 16.66 (mg/l) was 0.349 EUR per µM/s with an experimentally estimated flux value of 12.35 µM/s (Annexe 5). This method, therefore, makes it possible to identify the production costs of 1 µM of product from 0.197 € to 6.28 € to choose the best compromise between cost and speed of the reaction.

Lastly and interestingly, the validated kinetic model makes it possible to generate a huge amount of data so as to feed our ANN-based model with more flux values from the newly predicted enzyme balances. This should be explored in future studies.

4.7 Conclusion

The selection of enzymes is an important step in the production of biomolecules. Methods based on homology are widely used to select the best performing enzymes. In addition, the selection of optimum enzyme balances is also crucial. Most methods use kinetic information for concentration selection via modeling. However, the determination of kinetic parameters is not always easy; therefore, developing new methodologies for selecting the optimum enzyme balances is of great interest.

In this study, a new approach, GC-ANN, was developed which uses an artificial neural network along with different statistical methods (PCA and data classification) to select enzyme balances that improve the flux as well as the costs. The selected balances may not be the balances with the highest flux, but they will be among the best. This approach allows cost-efficient selection of enzyme balances using a small existing dataset, and it opens the door for rapid optimization of cell-free systems in an industrial environment.

Chapter 5 Kinetic Modelling of the Upper Part of Glycolysis.

5.1 Context

Glycolysis is one of the central metabolic pathways where glucose is converted into a series of intermediates forming pyruvic acid. Glycolysis reactions can be classified into two phases: the preparatory phase and pay off phase. In the preparatory phase, six-carbon glucose is converted into two, three-carbon sugar phosphates. ATP is consumed in this phase. In the pay off phase, glyceraldehyde-3-phosphate is converted into pyruvate with a series of the intermediate reactions. The energy-rich ATP and NADH are formed in this step. The pyruvate, end product of glycolysis, is the branch point for many biochemical syntheses. The glycolysis intermediates are involved in the synthesis of many chemical molecules.

Mathematical modelling of biological pathways is used in the optimisation of the process for chemical synthesis. The modelling helps to reduce the cost by the selection of optimum balances. The balances are defined by the combination of enzymes in the pathway. In this study, we focus on the preparatory part of glycolysis. Thus, the balances are the combination of enzymes phosphoglucoisomerase (PGI), phosphofructokinase (PFK) and fructose bisphosphate aldolase (FBA) and triosephosphate isomerase (TPI).

Different types of models have been developed to measure the flux through the metabolic pathway. Mainly, there are two kinds of models, i. the constraint-based model which uses constraints like mass balance, physicochemical constraints, *etc*, ii. kinetic model, which uses kinetic parameters of enzymes and helps in understanding the behaviour of the system. The kinetic model provides information about metabolite concentration. In Chapter 3 and Chapter 4 an artificial neural network model was developed to predict flux through the upper part of glycolysis. In both the studies, the ANN model was trained with 121 balances varied concentrations of enzyme PGI, PFK, FBA and TPI and corresponding flux values measured by *in vitro* experiments. The dataset consists of the flux value range from 0.78 μM/s to 12.9 μM/s, and as a training-based method, ANN cannot predict accurately beyond the range of training set. In Chapter 4, it was demonstrated that these newly predict values can be outside the training data and that our GC-ANN approach allows selecting the balances with higher flux. However, the ANN model was built on a small dataset of 121 balances and flux values. Increasing dataset by experiments is an expensive process, and choosing *in silico* method such as building a kinetic model of the pathway could be a good choice to reduce the cost and time required in generating new data. It could be a good choice, particularly when considering the availability of kinetic parameters measured in Chapter 4. In this study, the kinetic model of the upper part of glycolysis was built to estimate the flux through the pathway.

In the previous study described in Chapter 4, it was shown that while excellent predictions of flux values are obtained within the range of experimental values (0-12 μM/s) used in the training dataset, this model could predict a maximum flux around 15 μM/s. Nevertheless, experimental measurement provides flux values up to 20 units i.e. far from the maximal value predicted by the model. Thus, the objective of this study was to first optimise the kinetic model to replicate experimental conditions and then use the optimised model to generate a huge amount of data consisting of different enzyme balances and corresponding flux values for the upper part of glycolysis; this newly generated data would be used to train the ANN model to predict flux, which is expected to be higher than 12.9 μM/s. In this study, with the purpose of optimising the kinetic model of the upper part of glycolysis, parameter estimation (PE) for the kinetic parameters of the model was performed using ODE (reaction rate equation) (Hoops et al., 2006). The model parameters were estimated using the existing experimental data of measurement of flux, as NADH consumption by glycerol-3-phosphate dehydrogenase system.

5.2 Materials and Methods

5.2.1 Enzyme Assays for Measurement of Kinetic Parameters

Hexokinase (HK) activity was assessed using glucose-6-phosphate dehydrogenase (G6PDH) in a coupled reaction. The substrate glucose was converted to 6-phosphogluconate, the formation of NADPH was followed spectrophotometrically at 340 nm (Figure 5.1A).

The activities of phosphoglucoisomerase (PGI), phosphofructokinase (PFK) and fructose bisphosphate aldolase (FBA) was assessed using a coupled NADH assay applied to the upper part of glycolysis (Figure 5.1B). To determine the activity of PGI, the assay started with glucose-6-P (Figure 5.1B, reaction 1); for the measurement of the activities of PFK and FBA, fructose 6-P and fructose 1,6-bisP were used as the substrates (Figure 5.1B, reactions 2 and 3). The detailed method of kinetic parameter measurement is explained in section “Enzyme Assays for the Determination of Kinetic Parameters” of Chapter 4.

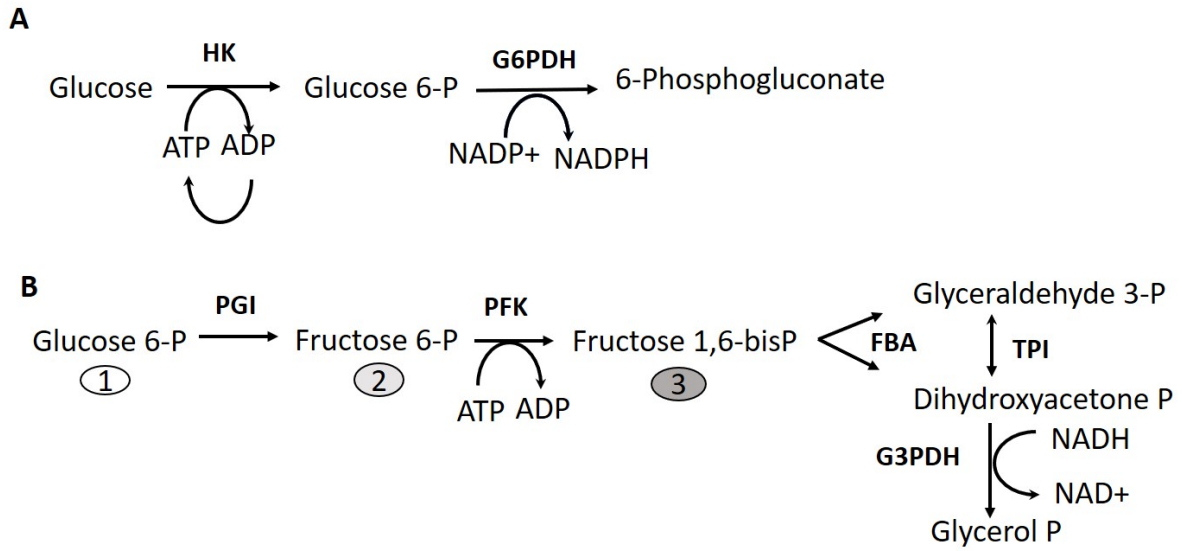


Figure 5.1: (A) Coupled HK/G6PDH assay to assess the HK activity. (B) Coupled NADH assay to assess the activities of PGI, PFK and FBA. The individual reactions were started with substrates indicated by the numbers in circles.

5.2.2 Reconstruction of *In Silico* Model

For the *in silico* reconstruction of the upper part of glycolysis, the kinetic model was built in CellDesigner (Funahashi *et al.*, 2008, 2003). The model consists of seven reactions. Creatine kinase was used for the regeneration of ATP, glycerol-3-phosphate dehydrogenase (G3PDH) was used for the measure the flux in terms of NADH consumption. The system consists of a constant concentration of HK, creatine kinase (CK), G3PDH, ATP. The kinetic equations and the parameters used in the kinetic model are given in Table 5.1. For the experimental measurement of flux, the concentration of PGI, PFK, TPI and FBA are varied (µM) as given in Table 5.3.

5.2.3 Experimental Flux Determination

To optimise the kinetic model, the experimental data of NADH consumption through the reconstructed *in vitro* system of the upper part of glycolysis are used. The experimental measurement of flux is explained in detail in "Experimental validation" and "Flux determinations" sections of Chapter 4. The data consist of triplicate measurement of NADH decay for 600 seconds for 31 experiments, with various concentrations of PGI, PFK, FBA and TPI. The plot of NADH consumption over time through the pathway is given as a figure in Annexe 6 and Annexe 7**.**

Table 5.1: Kinetic equations used at the beginning of optimisation of the upper part of glycolysis. G6P: Glucose6-phosphate; F6P: fructose-6-phosphate; FBP: fructose bisphosphate; DHAP: dihydroxyacetone phosphate. kcat: turnover number in s-1; Km: Michaelis-Menten Constant in µM; Ki: inhibition constant in µM and K_{eq}: equilibrium constant without units.

Reaction catalysed by	**Kinetic equation**	**Kinetic Parameter**
Hexokinase (HK)	$v = \frac{\text{kcatHk} * \text{HK} * \text{Glucose} * \text{ATP}}{((\text{Glucose} + \text{KmHKGlu}) * (\text{ATP} + \text{KmHKATP}))}$	kcatHK = 298.83; KmHKGlu = 280; KmHKATP = 100
Phosphoglucoisomerase (PGI)	$v = \frac{\text{kcatPgiF} * \text{PGI} * \frac{\text{G6P}}{\text{KmPGIG6P}} - \text{kcatPgiR} * \frac{\text{F6P}}{\text{KeqPGI} * \text{KmPGIF6P}}}{1 + \frac{\text{G6P}}{\text{KmPGIG6P}} + \frac{\text{F6P}}{\text{KmPGIF6P}}}$	kcatPgiF = 1107.367; kcatgiIR =3720; KmPGIG6P= 1100; KmPGIF6P= 4100; KeqPGI = 31
Phosphofructokinase (PFK)	$v = \frac{\text{kcatPfk} * \text{PFK} * \text{F6P}^{\text{nH}} * \text{ATP}}{(KmPfkF6P^{nH} + F6P^{nH}) * (KmPfkATP + ATP)}$	kcatPfk =166.075; KmPfkF6P = 130; KmPfkATP =120; nH = 1.1
Fructose bis-phosphate aldolase (FBA)	$v = \frac{\text{kcatFbaF} * \text{FBA} * \left(\frac{\text{FBP}}{\text{KmFbaFBP}}\right) - \text{kcatFbaR} * \text{FBA} * \left(\text{G3P} * \frac{\text{DHAP}}{\text{KmFbaG3P} * \text{KmFbaDHAP}}\right)}{\left(1 + \frac{\text{FBP}}{\text{KmFbaFBP}} + \frac{\text{G3P}}{\text{KmFbaG3P}} + \frac{\text{DHAP}}{\text{KmFbaDHAP}} + \text{FBP} * \frac{\text{G3P}}{\text{KmFbaFBP} * \text{KiFbaG3P}} + \text{G3P} * \frac{\text{DHAP}}{\text{KmFbaG3P} * \text{KmFbaDHAP}}\right)}$	kcatFbaF =16.789; kcatFbaR =720; KmFbaFBP = 140; KmFbaG3p =2000; KmFbaDHAP = 2400; KiFbaG3P= 10000
Triose-phosphate isomerase (TPI)	$v = \frac{\text{kcatTpiF} * \text{TPI} * \frac{\text{G3P}}{\text{KmTpiG3P}} - \text{kcatTpiR} * \text{TPI} * \frac{\text{DHAP}}{(\text{KmTpiDHAP} * \text{KeqTPI})}}{\left(1 + \frac{\text{DHAP}}{\text{KmTpiDHAP}} + \frac{\text{G3P}}{\text{KmTpiG3P}}\right)}$	kcatTPIF = 8486.67; kcatTpiR= 816.67; KmTpiDHAP= 1230; KmTpiG3P = 1270
Glycerol-3-phosphate dehydrogenase (G3PDH)	$v = \frac{\text{kcatG3PDH} * \text{G3DH} * \left(\frac{\text{DHAP}}{\text{KmG3dhDHAP}}\right) * \left(\frac{\text{NADH}}{\text{KmG3dhNADH}}\right)}{\left(1 + \frac{\text{DHAP}}{\text{KmG3dhDHAP}} + \frac{\text{Gly3P}}{\text{KmG3dhGly3P}}\right) * \left(1 + \frac{\text{NADH}}{\text{KmG3dhNADH}} + \frac{\text{NAD}}{\text{KmG3dhNAD}}\right)}$	kcatG3PDH= 189; KmG3dhDHAP =75; KmG3dhG3P = 909; KmG3dhNADH = 22; KmG3dhNAD = 83
Creatine kinase (CK)	$v = \frac{\text{kcatCK} * \text{CK} * \text{Phosphocreatine} * \text{ADP}}{\left(1 + \frac{\text{Phosphocreatine}}{\text{KmCkPCre}} + \frac{\text{Creatine}}{\text{KmCkCre}}\right) * \left(1 + \frac{\text{ADP}}{\text{KmCkADP}} + \frac{\text{ATP}}{\text{KmCkATP}}\right)}$	kcatCK = 148; KmCkPCre = 5000; KmCkCre =16000; KmCkADP = 800; KmCkATP = 500

5.2.4 Optimisation of Kinetic Model

The main objective of this study was to obtain the kinetic model of the upper part of glycolysis which could replicate the experimental condition. Therefore, to use the model to simulate flux with more enzyme balances without hands-on experiments. Further, to use these new balances and predicted flux for training the ANN model. Since the model was not optimised, kinetic parameter estimation was performed using the experimental data of NADH consumption over time, through the upper part of glycolysis.

5.2.4.1 Selection of Experimental Data

The experimental measurement for the flux determination was carried out in triplicates and final flux was calculated and the median deviation of the measurement of experimental triplicates. In 31 experimental measurements of NADH consumption over time, two groups were observed by visualisation inspection:

Group 1: Results of NADH consumption with a low deviation between triplicates. *i.e.* keeping all triplicate measurements (Indexes 14, 17-22, 24-27, 30, 31 and 37) (Annexe 6) and,

Group 2: Results of NADH consumption with low deviation between duplicates, *i.e.* by omitting the replicate which has a higher deviation. (Indexes 11-13, 15, 16, 23, 28, 29, 32-36, 38-41) (Annexe 7)

5.2.4.2 Parameter Estimation

Parameter estimation (PE) is the process of estimating the parameter in the model by mathematically fitting the simulated data to the measured dataset. This data could be from time-course analysis, steady-state or both. The goal is to minimise the objective function by scanning one or more parameters within the specified range. The objective function can be considered as the error between experimental and simulated data. The parameter estimation is computationally expensive. It is important to notice that if the optimisation problem is not well-posed (e.g. if the numbers of parameters to estimate is large regarding the numbers of targets) the existence and the uniqueness of the optimal solution cannot be guaranteed.

In this study, COPASI was used which is a tool which can handle both stochastic and deterministic simulation of pathways. The kinetic model consists of ordinary differential equations (ODE) (reaction rate equation). The main steps in the parameter estimation using COPASI is as follows:

1. Identify the model: From the data and the model, identify which model object is linked to the data.
2. Select the search algorithm: The search algorithms attempt to minimise the root mean square error (RMSE) (objective/cost function) between experimental and simulated data.
3. Select the parameter to be fitted: Select the parameters of unknown value (or to be estimated) of the model which should be fitted with the experimental data to minimise the objective function. These parameters will be estimated to find the minimised objective function.

Different algorithms available for parameter estimation task from COPASI are tested.

5.2.4.2.1 Glycerol-6-Phosphate Dehydrogenase Turnover Number (k_{cat})

The kinetic model was built, based on the experimentally measured k_{cat} for the enzymes HK, PGI, PFK, FBA and TPI, whereas the k_{cat} of G3PDH was taken from the literature. Since the k_{cat} was not measured in our experiments, estimations of k_{cat} for G3PDH was performed for only the Group 1 data were performed.

5.2.4.2.2 Iterative Estimation of Kinetic Parameters

The kinetic model contains seven reactions (Table 5.1), where the CK system was used to keep the ATP in the system constant. The enzyme HK was used at the constant concentration in the system to make sure the constant flux of glucose-6-phosphate. This requirement for the constant flux is used as the constraint in the model, *i.e.* the kinetic parameter, are chosen in this study to keep a constant flow of glucose-6-phosphate. Therefore, the other five i.e., PGI, PFK, FBA, TPI and G6PDH enzymes were chosen to estimate the kinetic parameters of the model. From five enzymes, 24 parameters *i.e.,* catalytic constant (k_{cat}) and the Michaelis-Menten constants (K_m) were needed to be estimated.

Firstly, estimating individual parameters would be computationally expensive and time-consuming. The model consists of 24 kinetic parameters from PGI, PFK, FBA, TPI and G6PDH and 31 experiments with varied enzyme concentration. If individual parameter must be estimated to fit with experimental data, at least 744 (i.e., 31 experiments x 24 parameters) parameter estimation analyses are needed to be performed in one iteration. More importantly, if the numbers of parameters to estimate is large, then it is often impossible to find a unique solution, and several sets of parameters could give the best. To reduce the number of experiments, the iterative approach of parameter estimation using the experimental triplicate

data between 60-120s was performed using COPASI (Hoops et al., 2006). By doing so, only 155 experiments for one iteration (31 case x 5 enzymes) was performed.

Secondly, to limit the search space for a parameter value, which reduces the calculation time, the values mentioned in the BRENDA database was used to limit the range (Table 5.2). The BRENDA database contains the curated enzyme kinetic parameters from the experiments reported in the literature. The parameter estimation was performed on both Group 1 and Group 2 experimental data.

Table 5.2: The Parameter range used for the parameter estimation for each enzyme. PGI: phosphoglucoisomerase, PFK: phosphofructokinase, FBA: fructose 1,6, biphosphate aldolase, TPI: triosephosphate isomerase, G6PDH: glycerol-6phosphate dehydrogenase, EC No: Enzyme commission number, F: forward catalytic constant, R: reverse catalytic constant.

Enzyme	EC No	K_m (mM)	k_{cat} (s^{-1})	Organism
PGI	5.3.1.9	G6P = 0.084 – 1.5; F6P = 0.11-0.307	F = 487-1410; R = 247.2-3720	yeast
PFK	2.7.1.11	F6P = 0.23-0.15; ATP = 0.07	F = 3.1-210	Bacilllus st
FBA	4.1.2.13	FBP = 0.00084-5; G3P = 0.3-1; DHAP = 0.4-2	F = 0.55 – 42.4; R= 10.28	Rabbit muscle
TPI	5.3.1.1	G3P =1.1-1.5; DHAP = 1.23 – 2.3	F = 4700-16700; R = 500	yeast
G3PDH	1.1.1.8	DHAP= 0.075-0.46; NADH = 0.0043-0.022; G3P= 0.19-0.909; NAD = 0.0044-0.38	F: 309	Rabbit muscle

From the upper part of glycolysis model, for 31 enzyme concentration conditions (Table 5.3) the kinetic parameters of each enzyme is estimated separately. After the estimation process, RMSE between the mean of the experimental triplicate concentrations of NADH and the concentration of NADH by obtained using the kinetic model with the newly estimated parameter were compared. The enzyme which has lower RMSE between experiment and estimated concentrations were updated to the model and repeated until all the enzyme parameters are newly estimated (referred as one cycle where all the parameters from five enzymes are estimated). The process is repeated up to five times. Then for parameters which deviated less across 31 conditions, at the end of the fifth cycle of parameter estimation, mean was calculated. The mean of the less deviated parameter is updated to the model, and highly deviated parameters are again estimated in five iterations, updating the parameter value with the newly estimated value.

The updated model after five cycles of estimation was utilised to simulate the other enzyme balances given in Table 5.3. The model was simulated for 120 seconds and the flux was calculated as the slope of NADH concentrations between 60 to 120 seconds like in experimental estimation. The variation of estimated kinetic parameters estimated after five cycles were compared.

Table 5.3: The concentration of enzymes (µM) used in the kinetic model of the upper part of glycolysis. The equivalent U/ml concentration was used in the experiments. PGI: phosphoglucoisomerase, PFK: phosphofructokinase, FBA: fructose 1,6, biphosphate aldolase, TPI: triosephosphate isomerase, JExp: experimental flux; MD: median deviation

Index	**U/ml**				**µM**				**JExp**	**M.D**
	PGI	**PFK**	**FBA**	**TPI**	**PGI**	**PFK**	**FBA**	**TPI**	**(µM/s)**	
11	2.74	0.7	3.71	24.39	0.080	0.282	9.451	0.096	15.7	2.5
12	2.74	0.7	3.62	53.77	0.080	0.282	9.237	0.211	16.3	2.7
13	2.74	0.77	3.45	97.84	0.080	0.310	8.809	0.384	12.1	4.2
14	2.74	0.84	3.37	112.53	0.080	0.338	8.595	0.442	16.6	0.1
15	2.74	0.91	3.58	24.39	0.080	0.367	9.130	0.096	13.9	3.9
16	2.74	0.98	3.54	24.39	0.080	0.395	9.023	0.096	18.3	1.2
17	2.74	1.05	3.50	24.39	0.080	0.423	8.916	0.096	17.1	0.2
18	2.74	1.12	3.29	83.15	0.080	0.451	8.380	0.327	20.1	0.3
19	4.11	0.7	3.58	53.77	0.121	0.282	9.130	0.211	14.4	0.1
20	4.11	0.84	3.58	24.39	0.121	0.338	9.130	0.096	15.8	0.2
21	4.11	1.12	3.37	39.08	0.121	0.451	8.595	0.154	20.6	0.2
22	5.48	0.77	3.58	24.39	0.161	0.310	9.130	0.096	15.4	0.2
23	5.48	1.12	3.37	24.39	0.161	0.451	8.595	0.096	16.1	2.3
24	5.48	1.12	3.33	39.08	0.161	0.451	8.487	0.154	19.3	0.6
25	6.85	1.05	3.37	24.39	0.201	0.423	8.595	0.096	18.5	0.6
26	6.85	1.12	3.33	24.39	0.201	0.451	8.487	0.096	17.8	0.1

27	6.85	1.12	3.29	39.08	0.201	0.451	8.380	0.154	16.3	0.3
28	6.85	1.12	3.24	53.77	0.201	0.451	8.273	0.211	19.7	2.5
29	8.22	1.05	3.33	24.39	0.241	0.423	8.487	0.096	17.8	1
30	8.22	1.05	3.29	39.08	0.241	0.423	8.380	0.154	19	0.6
31	8.22	1.05	3.24	53.77	0.241	0.423	8.273	0.211	21	0.6
32	8.22	1.12	3.29	24.39	0.241	0.451	8.380	0.096	15.6	3.1
33	8.22	1.12	3.24	39.08	0.241	0.451	8.273	0.154	17.8	2.2
34	9.59	0.84	3.29	68.46	0.281	0.338	8.380	0.269	17.1	0.7
35	9.59	1.05	3.29	24.39	0.281	0.423	8.380	0.096	17.7	1
36	9.59	1.05	3.24	39.08	0.281	0.423	8.273	0.154	18.8	1.8
37	9.59	1.12	3.24	24.39	0.281	0.451	8.273	0.096	20.4	0.6
38	10.96	0.91	3.33	24.39	0.322	0.367	8.487	0.096	15.9	0.9
39	10.96	1.05	3.24	24.39	0.322	0.423	8.273	0.096	17.9	0.8
40	12.33	0.84	3.29	39.08	0.362	0.338	8.380	0.154	15.8	0.9
41	13.7	0.84	3.29	24.39	0.402	0.338	8.380	0.096	13.6	2.4

These enzyme concentrations are originally based on Table 4.6 converted into U/ml and μM based on the calculations described in Chapter 3: "Concentration Based on the Relative Activity". Therefore index starts from 11.

5.2.4.2.3 Summary of the Followed Methodology

The methodology followed is summarised in Figure 5.2 is as follows:

1. From the initial upper part of glycolysis model, the parameters were estimated for each enzyme using the experimental measurement of NADH concentration (step 1).
2. The RMSE between the experiment and the model with the new estimated parameter were calculated (step 2).
3. The parameters of enzymes with low RMSE from step 2 were updated to the model, and for the remaining enzymes, step 1 and 2 were performed until all five enzyme parameters were estimated. This referred to as one iteration(step 3).

4. The estimation of all five-enzymes (*i.e.*, PGI, PFK, FBA, TPI, G3PDH) kinetic parameters (both k_{cat} and K_m) were performed iteratively, five times and referred to as one cycle(step 4).

Steps 1-4 were performed for all 31 enzyme combinations outlined in indices 11-41 in Table 5.3.

5. The mean of the less deviated parameters across 31 experiments was updated to the model, and highly deviated parameters were estimated for five iterations using the mean updated model.
6. After estimating and updating the model with the highly deviated parameters, the model was simulated for 31 concentrations of enzymes from Table 5.3. The RMSE between the simulated flux and experimental flux was calculated.

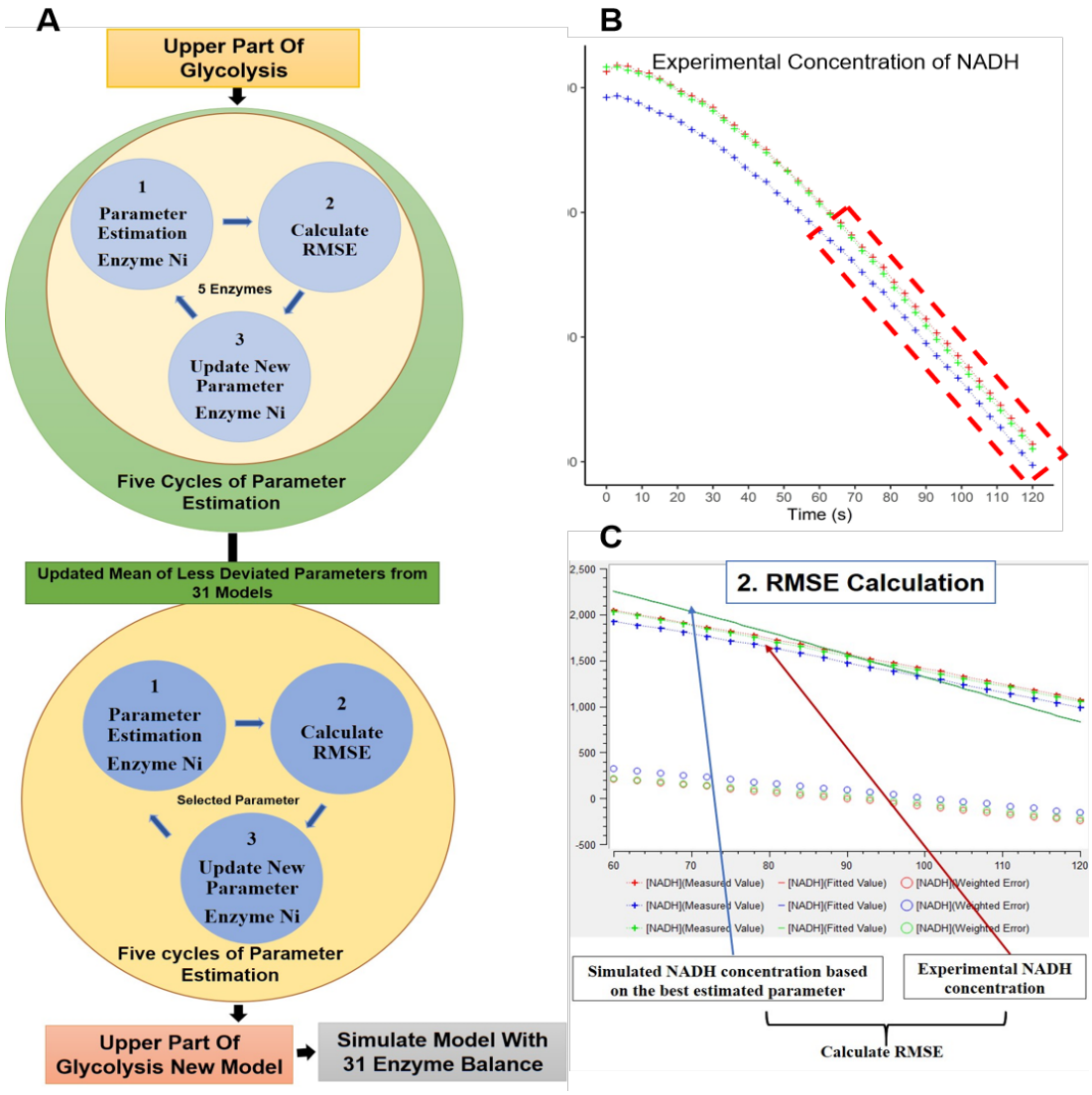


Figure 5.2: (A) Summary of the methodology followed for the parameter estimation. Enzyme *ni* represents the kinetic parameters of PGI, PFK, FBA, TPI and G3PDH iteratively (i=1 to 5). (B:) Experimental measurement of NADH between 0 to 120 seconds. Between 60-120 seconds concentration are considered for Parameter estimation. (C) calculation of RMSE between experimental and fitted value after the parameter estimation.

The above methodology was followed for all 31 enzyme combinations outlined in indices 11-41 in Table 5.3 individually.

5.2.4.3 Ranking of Simulated Flux

The rank correlation of data measures the ordinal association between the ranking of two different variables or different ranking of same variables and the rank correlation coefficient measures the degree of similarity between the two rankings and can be used to weigh the significance of relation. Kendall Tau and Spearman coefficient are two of the most popular rank correlations. These two methods measure the association between variables considering the rank of the data and not the value itself. The data is ranked by putting the variables in order and ranked.

Kendall Tau calculation is based on the concordant and discordant pairs. The rank order between a pair of two variables is said concordant when they have a similar ranking and discordant if the ranking is different.

Kendall Tau coefficient between the two random variables is defined as Equation 5.1

Equation 5.1: Calculation of the Kendall Tau coefficient. Where, τ: Kendall Tau coefficient. *n:* number of observations

$$\tau = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{n(n-1)/2}$$

Spearman Coefficient calculation is based on the deviation between the two rankings and usually have larger values than the Kendall tau.

Equation 5.2: Calculation of the Spearman rank correlation coefficient. Where, ρ= Spearman rank correlation; d_i = the difference between the ranks of corresponding variables; n= number of observations

$$\rho = 1 - \frac{6\sum {d_i}^2}{n(n^2-1)}$$

The correlation coefficients take the value between -1 and +1. The Kendall Tau and Spearman Coefficients between experimental and model-simulated flux is calculated.

5.3 Result and Discussion

The kinetic model was built to reconstruct *in vitro* upper part of glycolysis. The model consists of newly estimated parameters for the forward direction, the reversible reaction and the co-substrates parameters are taken from the literature. The model was simulated with varied concentration of enzymes (Table 5.3) for 120 seconds, the flux was calculated as the slope of NADH consumption between 60-120 seconds. The experimental vs simulated flux from the kinetic model were compared (Figure 5.3). The model has an RMSE of 5.147 between all the experimental flux and simulated flux.

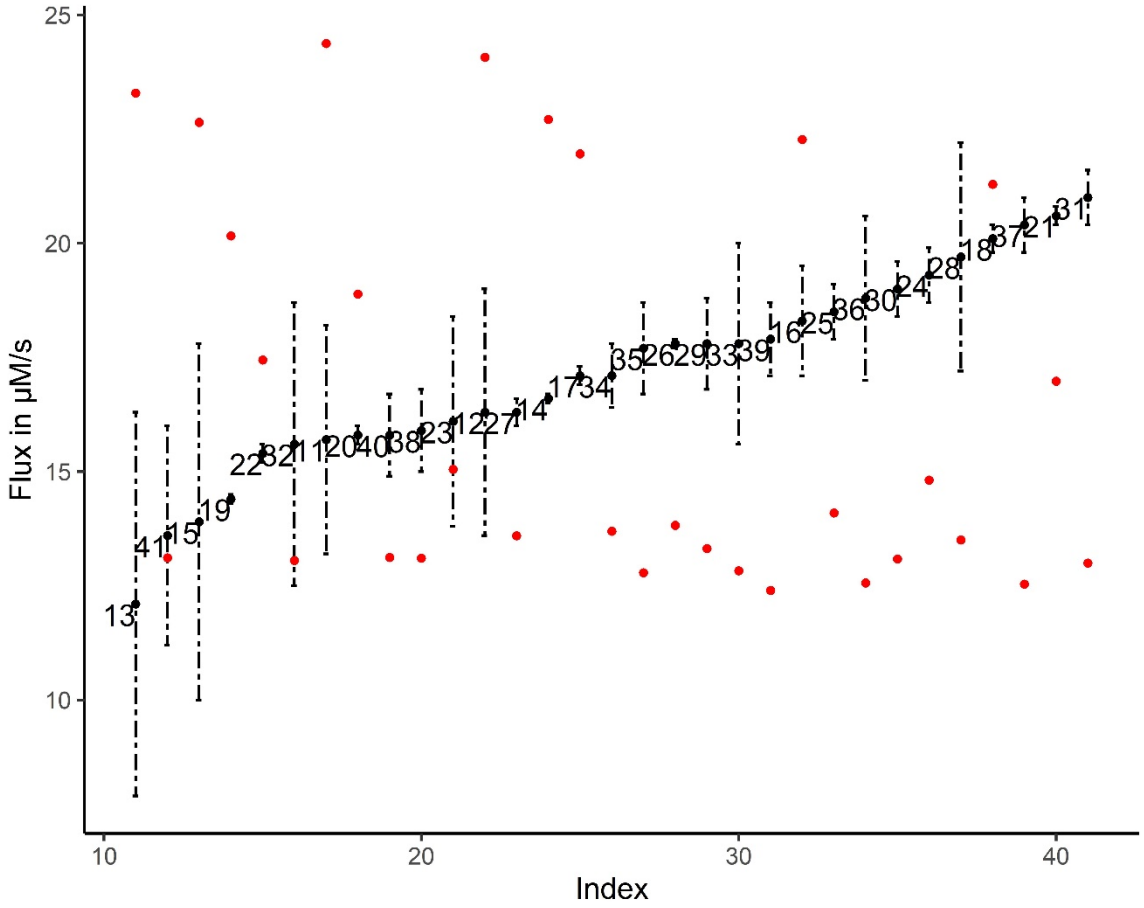


Figure 5.3: The experimental flux vs simulated flux of the upper part of glycolysis with varied concentration of PGI, PFK, FBA and TPI. The black filled circle represents the experimental flux with

standard deviation in black dashed line, red filled circles represents the simulated flux from the kinetic model. X-axis: indexed according to sorted experimental flux in ascending order.

Figure 5.3 shows that the original model is not optimised to replicate the experimental condition and the parameters need to be fine-tuned so that this kinetic model could thus be used for generating more data of enzyme balance, and for predicting flux to train ANN model. Hence the kinetic parameters of the model were optimised to replicate the experimental flux.

5.3.1 Optimisation of Kinetic Model

The enzymes HK and CK did not directly contribute to the flux variation. The HK was important to keep the constant flux of Glucose-6P while the CK system was important to keep the ATP in the system regenerated. Therefore, these enzymes parameters were not taken into account for the parameter optimisation.

5.3.1.1 k_{cat} Estimation Of Glycerol-6-Phosphate Dehydrogenase

In this study, k_{cat} of all the enzymes were measured, except for the G3PDH. Hence, initially, the parameter estimation was performed only for optimising k_{cat} of G3PDH. It was observed that by only optimising the k_{cat} value, a plateau in the final flux measurement for other conditions in Figure 5.4. Figure 5.4 proves that by only estimating the k_{cat} of G3PDH, the model cannot be optimised to replicate the experimental system.

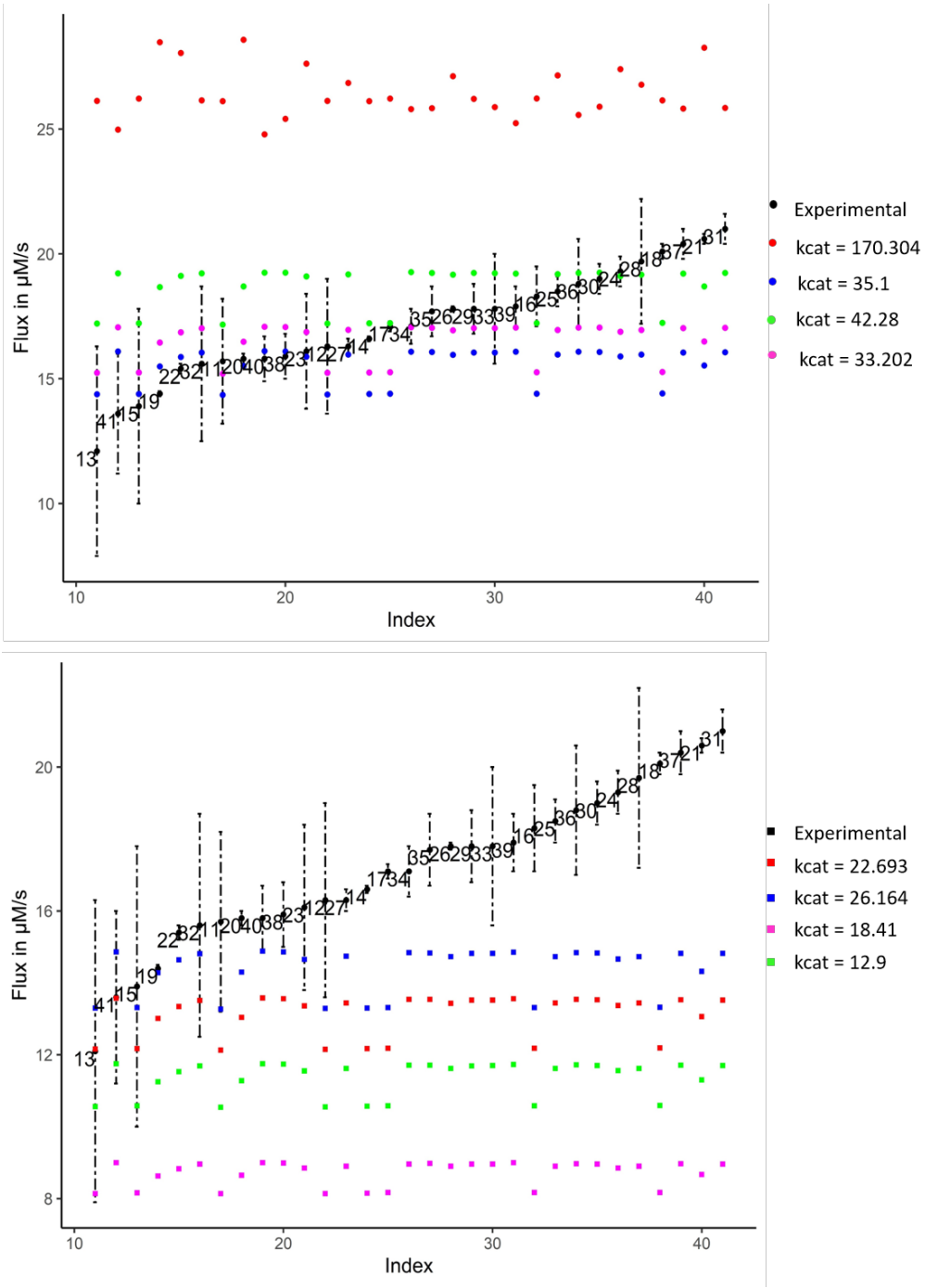


Figure 5.4: Effect of newly estimated k_{cat} parameters from glycerol-3-phosphate dehydrogenase (top and bottom). X-axis: indexed according to sorted experimental flux in ascending order. k_{cat} values are given in s^{-1}.

5.3.1.2 Iterative Estimation Of Parameters

To measure the flux through the pathway, the NADH consumption between 60-120 seconds was considered in the experiments the same as in Fiévet *et al.* (Fiévet *et al.,* 2006). Hence, the parameter estimation was carried out using data between 60-120 seconds. The parameter

estimation was performed iteratively for five enzymes *i.e.* PGI, PFK, FBA, TPI and G3PDH, for 31 datasets up to five cycles. It was observed that, at the fifth cycle of parameter estimation, there was not much improvement in the fitting of the model to experimental data, therefore estimation was stopped at the fifth cycle. The variation of RMSE across different iteration of parameter estimation is represented Annexe 8 and Annexe 9.

The efficiency of the new model was measured as the RMSE between the experimental measurement for 31 datasets with that of simulated flux using the updated parameter. The best model obtained after five cycles of estimation has an RMSE of 1.93, between the experimentally measured flux (total of 31 datasets) and the simulated flux. The RMSE, Kendall tau coefficient and Spearman coefficient were calculated for the model which measures the association between the two random variables. The Kendall tau and Spearman coefficients are used to measure the relationship between simulated and experimental flux. If the simulated and experimental flux is positively correlated then the coefficient values will be positive 1 for Kendall tau and spearman coefficient and if they are negatively correlated, then then the Kendall Tau and Spearman coefficient will be negative 1. The indexes with positive Kendall tau coefficient i.e., positive correlation between experimental and model-simulated flux is given in Table 5.4, and RMSE values less than 2, are given in Table 5.5. The kinetic parameters, yielding positive Kendall Tau, with the RMSE between experimental and simulated flux less than 2 is given in Annexe 10.

Table 5.4: The Kendall tau and Spearman coefficient between the best model obtained after the five cycles of parameter estimation and mean of experimental flux.

Index	Kendall Tau	Spearman Coefficient
Index37	0.42209	0.5861
Index29	0.41776	0.58025
Index30	0.38745	0.55079
Index39	0.25758	0.35226
Index38	0.18832	0.2413
Index40	0.17966	0.25361
Index41	0.17533	0.24533
Index35	0.11039	0.15737

Table 5.5: The model with RMSE less than 2 between the best model after five cycles of parameter estimation and the mean of experimental measurement for all 31 enzyme combinations outlined in indices 11-41 in Table 5.3.

Index	RMSE
Index37	1.97001
Index29	1.97653
Index30	1.93239
Index39	1.98293

The kinetic parameters corresponding to the model with the lowest RMSE after five cycles of parameter estimation (index 30:Table 5.5) are provided in

Table 5.6. All the kinetic parameters estimated during the iterative estimation are found to be within the biological range.

Table 5.6: Kinetic parameter obtained for the index 30 based model after five cycles of iterative parameter estimation. These kinetic parameters in the model yield an RMSE of 1.93 with mean of the experimental triplicates for 31 varied concentrations of enzymes PGI, PFK, FBA and TPI. The k_{cat} values are given in s^{-1}, K_m values in µM and equilibrium constant K_{eq} have no units.

Kinetic parameters	Index30	Kinetic parameters	Index30
KeqPGI	1.000	**kcatPfk**	27.784

kcatFbaR	0.000	**KmPGIG6P**	84.000
KeqTPI	0.021	**KmFbaFBP**	3.978
KmG3dhNADH	4.301	**kcatPgiR**	543.518
kcatFbaF	1.541	**kcatPgiF**	1398.289
KiFbaG3P	10000.000	**KmFbaDHAP**	1238.556
kcatTpiF	16700.000	**KmFbaG3P**	2500.000
kcatG3dhG3PDH	256.548	**KmTpiDHAP**	1521.178
KmPGIF6P	306.506	**KmTpiG3P**	1370.018
KmG3dhDHAP	119.781	**kcatTpiR**	942.491
KmPfkATP	76.389	**KmG3dhGly3P**	768.741
KmPfkf6p	143.616	**KmG3dhNAD**	380.000

The comparison of experimental flux vs simulated flux using the newly estimated parameter is given in (Figure 5.5) confirms that after estimating the kinetic parameters for all the enzymes in five iterations provided a better fit with the experimental flux.

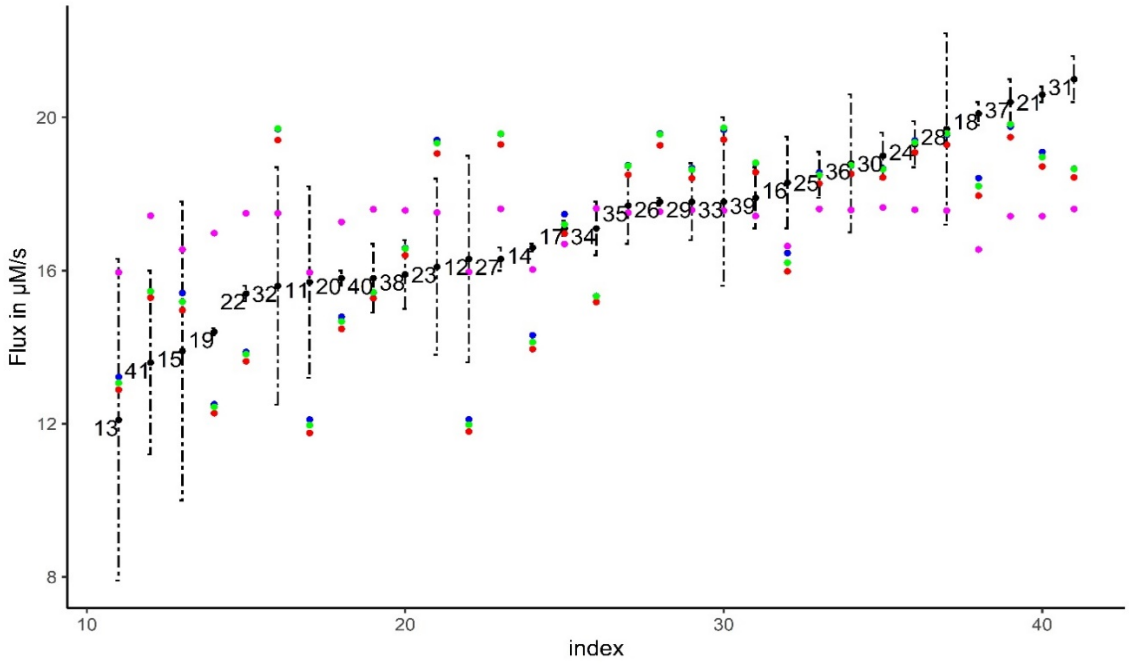


Figure 5.5: The experimental flux vs estimated flux from the model with after 5 iterations of parameter estimation. The black dots represent experimental flux with standard deviation, red, blue, green, magenta dots represent the estimated flux after 5 cycles of parameter estimation from index 29, 30, 37 and 39 respectively. X-axis: indexed according to sorted experimental flux in ascending order.

5.3.1.3 Selective Parameters Estimation

The parameters between 31 datasets are compared after five cycles of estimation (Figure 5.6). Figure 5.6 shows that some of the parameters have deviated less (less than 60%) within 31 experiments and some parameters deviate more (more than 60%) compared the initial model. The deviation is calculated in percentage as Equation 5.3 and the percentage of deviation is given in Annexe 11.

Equation 5.3: The percentage deviation of kinetic parameters during the iterative parameter estimation across 31 experimental conditions.

$$\text{Deviation parameter } (\%) = \frac{\text{Average of 31 experiments}}{\text{Initial value of the model}} * 100$$

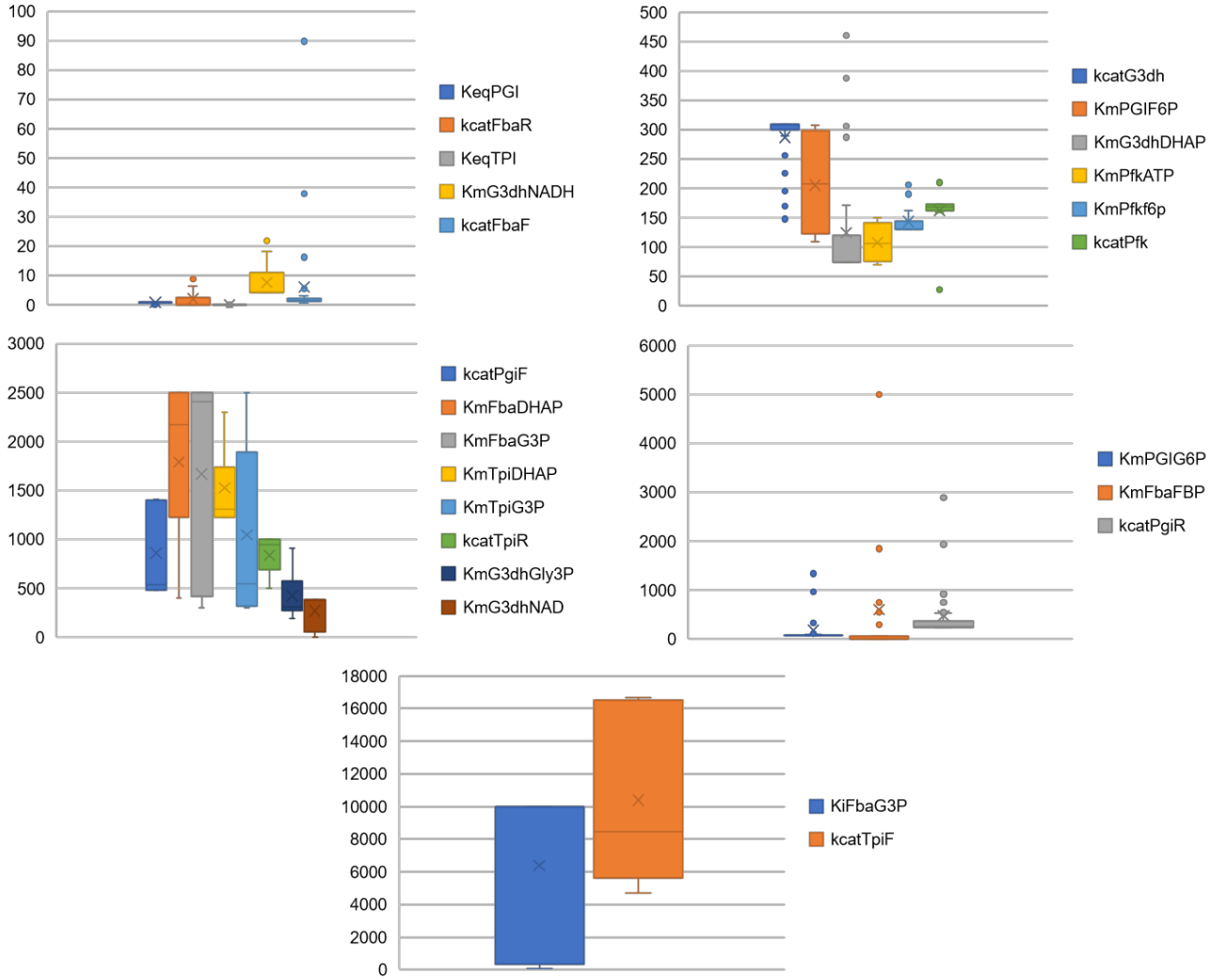


Figure 5.6: Summary variation of kinetic parameter parameters of PGI, PFK, FBA, TPI and G3PDH after the 5 cycles of parameter estimation from 31 datasets.

The hypothesis is- the parameter which contributes highly to the final flux variation, deviates less and those parameters, which has limited control over the final flux, deviate more. To test our hypothesis and improve the model accuracy further, only parameters which were deviated most (more than 60% from initial value) in the five cycles of PE in 31 datasets were chosen for further analysis (Figure 5.7 and Table 5.7). These parameters will be referred to as selective parameters.

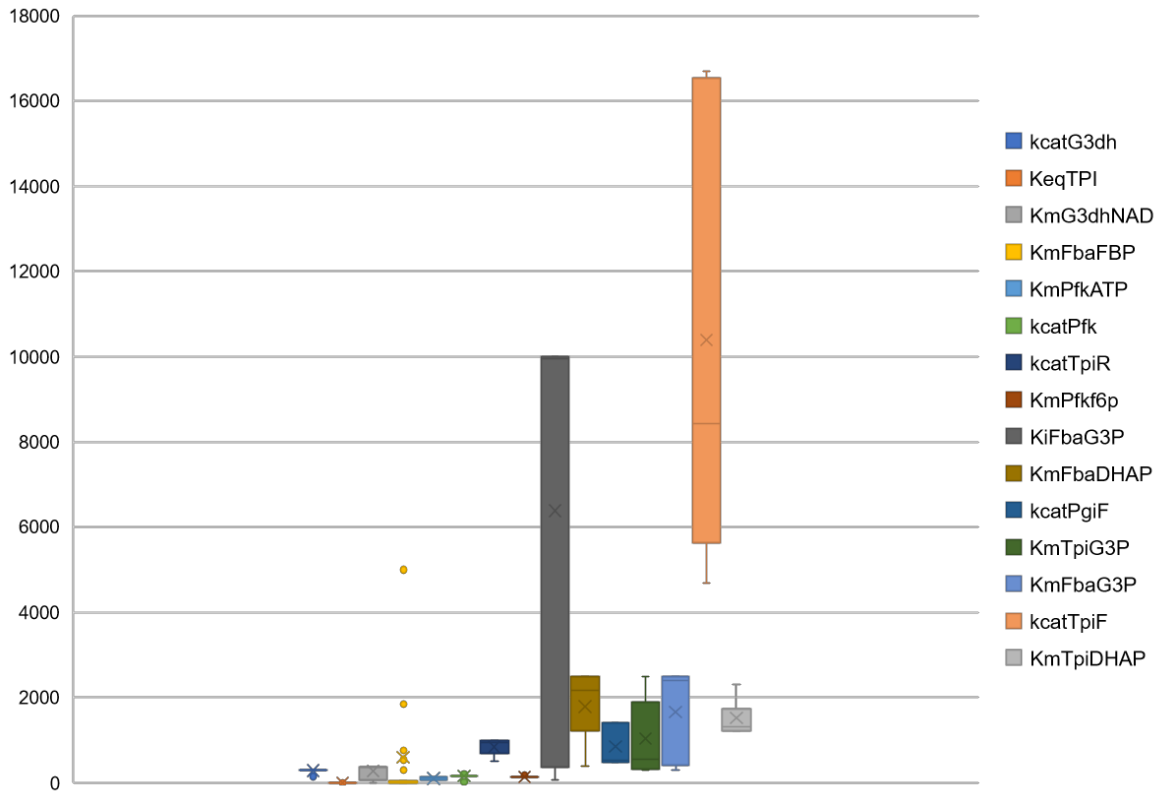


Figure 5.7: The selected parameters after five cycles of estimations from 31 datasets with higher deviation, for further analysis.

Table 5.7: Kinetic parameters that deviated most in between all 31 enzyme combinations outlined in indices 11-41 in Table 5.3 selected for further analysis.

Parameter	**Parameter**
kcatPgiF	KeqTPI
KmPfkf6p	KmTpiDHAP
kcatPfk	KmTpiG3P

KmPfkATP	kcatTpiF
KiFbaG3P	kcatTpiR
KmFbaDHAP	KmG3dhNAD
KmFbaFBP	kcatG3dh
KmFbaG3P	

Part of our hypothesis is that the parameter which deviated less has more control over the final flux. Thus, the model was updated with the mean of the less deviated parameters across 31 experiments to take account for different flux values as listed in Table 5.8. The selective kinetic parameters were estimated again in five cycles.

Table 5.8: The mean of the parameters with low variation (less than 60%) used in the updated model for further analysis. The units of K_m are µM and k_{cat} is s^{-1}.

Parameter	**Mean**	**Parameter**	**Mean**
KmPGIG6P	181.4721	kcatFbaR	2.039088
kcatPgiR	481.9601	KmG3dhGly3P	426.182
KmPGIF6P	205.7683	KmG3dhNADH	7.67841
KeqPGI	0.81634	KmG3dhDHAP	124.8282
kcatFbaF	6.193805		

The selective parameter estimation performed using the model containing enzyme concentrations of index 26 (from Table 5.3) yielded an RMSE of 1.91, Kendall Tau and Spearman coefficient of 0.292 and 0.403 respectively. However, Figure 5.8 revealed that, though RMSE between experimental flux and simulated flux, after the selective PE process, the model was inefficient in simulating flux across the different condition.

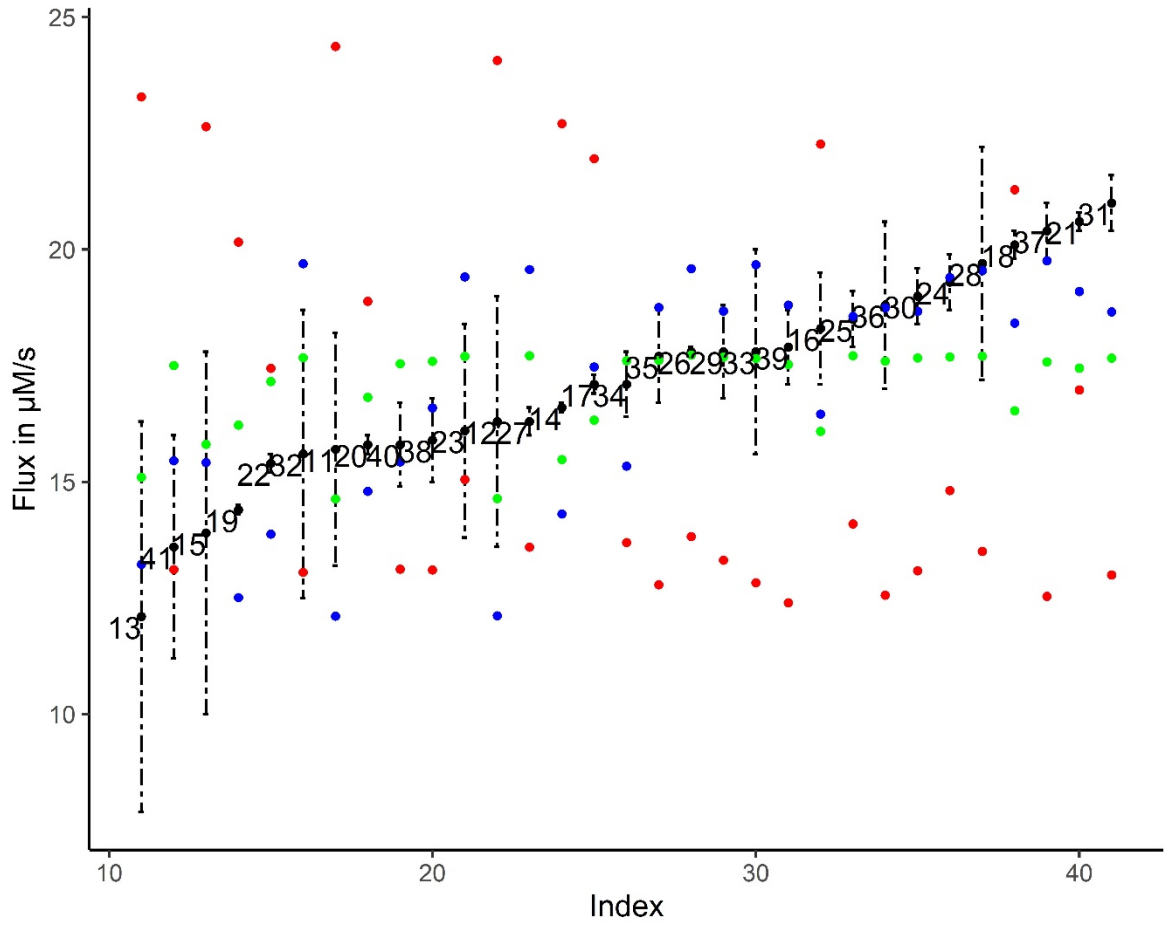


Figure 5.8: The comparison of best models obtained by iterative parameter estimation (index 30) and selective parameter estimation (index 26). Black circles represent experimental flux with standard deviation, red circles represent the flux with the original model before parameter estimation, blue circles represent index 30 based simulation after iterative estimation and green circles represent index 26 based selective simulation. X-axis: indexed according to sorted experimental flux in ascending order.

The highly deviating parameters by iterative PE and after five cycles of selective PE approaches were compared (Figure 5.9). The RMSE, Kendell Tau and Spearman coefficients were computed for all the 31 enzyme combinations (outlined in indices 11-41 Table 5.3)PE (Annexe 12) and only the positive values (i.e., the experimental and model-simulated flux across 31 conditions are positively correlated) are given in Table 5.9.

.

Table 5.9: RMSE, Kendell tau and Spearman coefficient for selective parameter estimation.

Index	RMSE	Kendall Tau	Spearman Coefficient	Index	RMSE	Kendall Tau	Spearman Coefficient
Index13	7.490	0.340	0.483	**Index31**	3.022	0.219	0.400
Index15	6.206	0.353	0.495	**Index32**	2.042	0.318	0.439
Index20	2.174	0.024	0.018	**Index33**	2.150	0.314	0.435
Index22	2.350	0.058	0.087	**Index34**	2.062	0.374	0.508
Index24	2.413	0.297	0.415	**Index35**	4.407	0.331	0.470
Index25	2.084	0.236	0.343	**Index36**	3.233	0.348	0.492
Index26	1.912	0.292	0.403	**Index37**	2.544	0.201	0.295
Index27	2.500	0.331	0.472	**Index38**	3.956	0.318	0.462
Index28	4.458	0.370	0.518	**Index39**	3.196	0.180	0.266
Index29	2.153	0.249	0.346	**Index40**	8.183	0.227	0.328
Index30	2.131	0.219	0.309	**Index41**	13.623	0.197	0.286

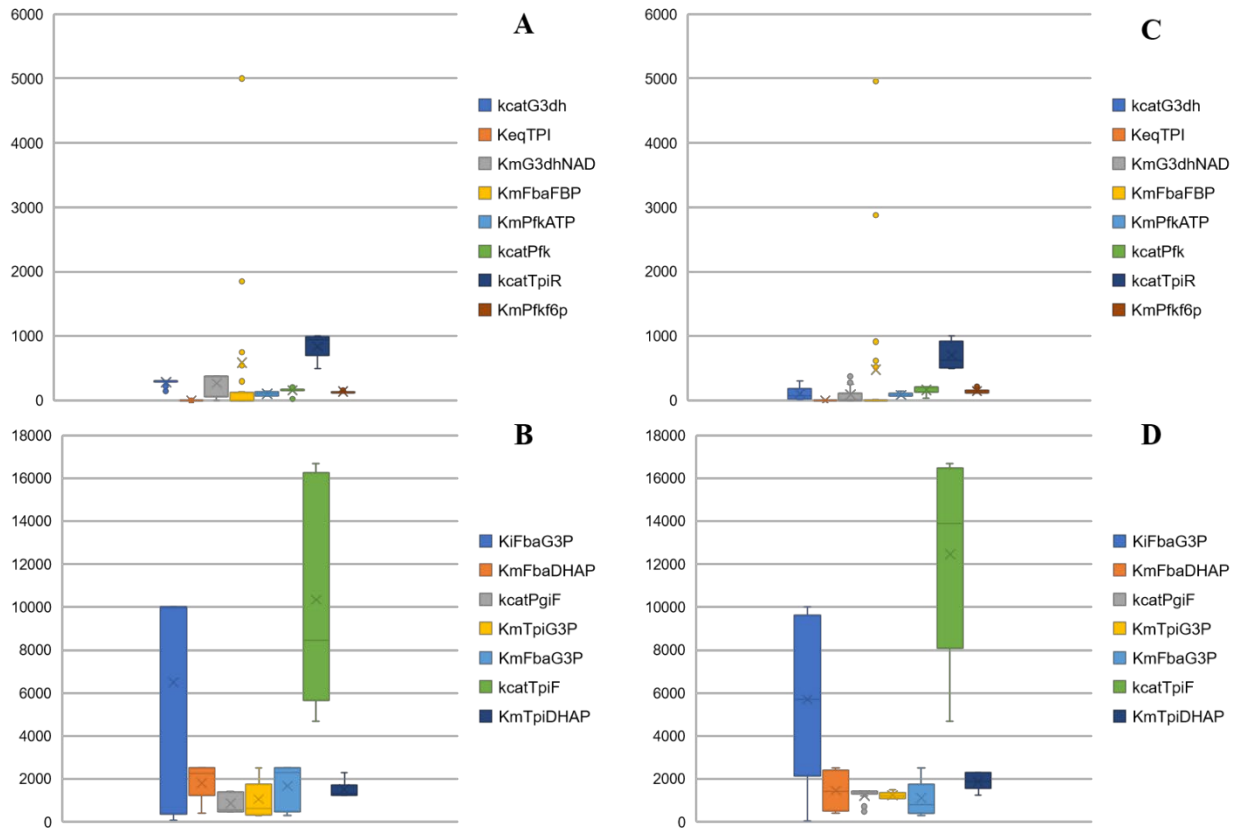


Figure 5.9: (A, B) The estimated highly deviated parameters after five cycles of iterative parameter estimation. (C, D) The parameter estimation of highly deviated parameters after updating the model with the mean of the less deviated parameter from iterative parameter estimation.

From Figure 5.9 and Table 5.10, it is evident that, by estimating the highly deviated parameter from the iterative PE approach, even though the range of variation varied, it did not improve the quality of the simulation in terms of RMSE, Kendall Tau and Spearman coefficients. In other words, the RMSE increased and Kendall tau and Spearman coefficient values decreased. In Figure 5.10 it is clear that, after PE using selective parameter approach, the flux predicted by kinetic models is not successful in simulating experimental flux accurately.

Table 5.10: The comparison of RMSE, Kendall tau and Spearman coefficients for the model with positive coefficient after the five iterative estimations, and after five cycles of estimation selective parameters.

After Iterative Approach				**After Selective Approach**		
Index	**RMSE**	**Kendall Tau**	**Spearman Coefficient**	**RMSE**	**Kendall Tau**	**Spearman Coefficient**
Index37	1.970005	0.422087	0.586099	2.543507	0.201303	0.295168
Index29	1.976526	0.417758	0.580248	2.153019	0.248923	0.346414
Index30	1.932388	0.387454	0.550792	2.131409	0.218619	0.308887
Index39	1.98293	0.257581	0.352265	3.195822	0.179657	0.266115
Index38	2.796894	0.188316	0.241299	3.955999	0.318189	0.462423
Index40	10.76548	0.179658	0.253606	8.182841	0.227278	0.327852
Index41	14.86231	0.175328	0.245335	13.62289	0.196974	0.286291
Index35	2.115905	0.110392	0.157369	4.406505	0.331176	0.470493

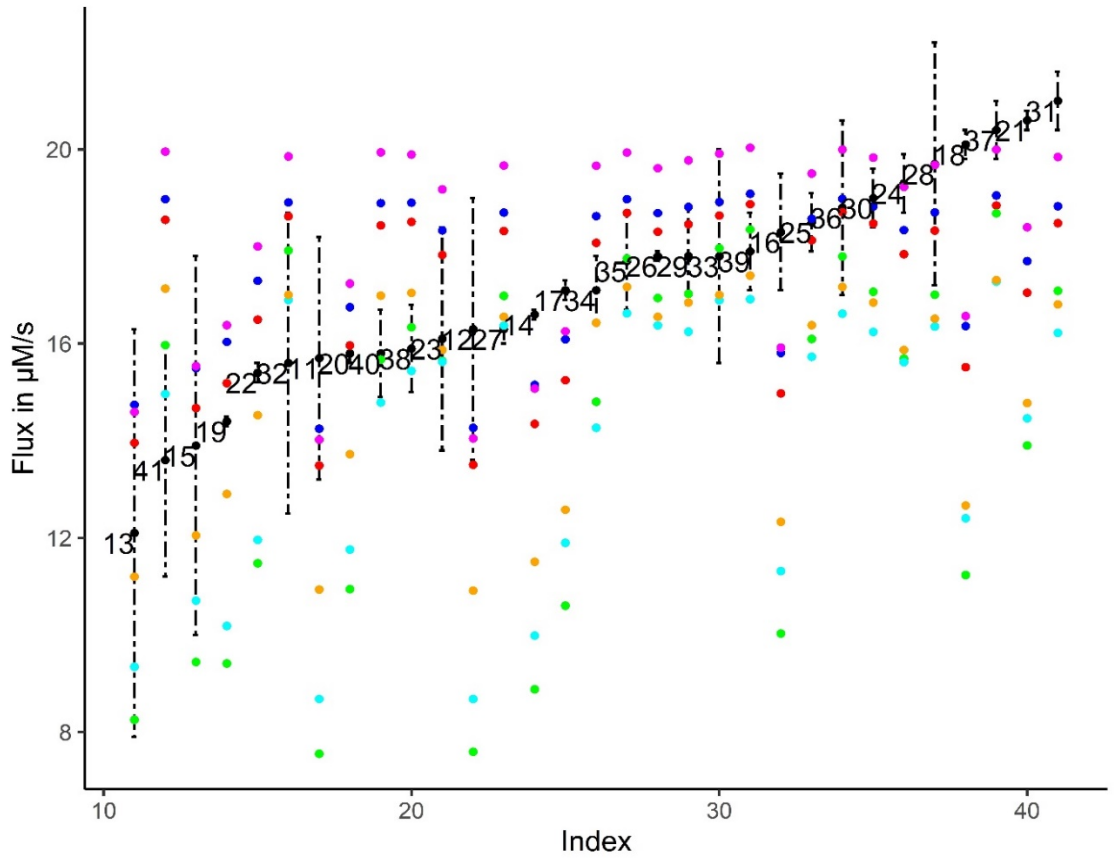


Figure 5.10: Comparison of experimental flux vs the flux simulated by the kinetic model after five iterations of parameter estimation using the mean of the selected parameter. Black circles represent experimental flux with standard deviation, red, blue, green, magenta, cyan and orange circles represent the flux simulated by the kinetic model with mean of parameter and newly estimated parameter for index 29, 30, 35, 37, 38 and 39 respectively. X-axis: indexed according to sorted experimental flux in ascending order.

Previously, the ANN model was built to predict the flux through the upper part of glycolysis (Chapter 3). The data used for training the ANN model has a flux range of 0.79 µM/s to 12.9 µM/s. The best ANN model has an RMSE of 0.8 and R^2 of 0.9 between experimental flux and ANN predicted flux with 13 hidden units in a single layer using the logistic activation function. For the experimental enzyme balances used in this study (Table 5.2), if the flux is predicted using the ANN model from Chapter 3, an RMSE of 4.98 and R^2 of 0.675 will be obtained. Such result was expected, indeed, since ANN models are known to not perform well outside the training dataset, whereas the kinetic model which is not a training-based model has no such limitation. Therefore, in this study, the kinetic model was built using experimentally measured kinetic parameters. The main objective of the present study was to build the kinetic model of the upper part of glycolysis which can simulate the various range of flux values to generate

more data for the ANN model. So, the ANN model could be extended beyond the range of 0.79 µM/s to 12.9 µM/s. The first kinetic model built in this study was simulated with experimental flux values from 12.1 to 21 µM/s and the maximum outcome was 24.3 µM/s. From Figure 5.3, it is shown that the kinetic model is not optimised to replicate the experimental system. After the extensive investigation to optimise the parameters of the initial kinetic model, the analysis showed that it is not always easy to build the kinetic model to replicate the experimental conditions. Indeed, regarding the large number of parameters to be estimated, there is not a unique optimal solution to the optimisation problem. By performing different parameter estimations, the model was finally optimised to simulate the flux with an RMSE of 1.91. Nonetheless, this model quality is insufficient to be used for generating data to improve and extend the ANN model presented in a previous Chapter 3-Chapter 4. Nevertheless, the final kinetic model obtained with parameter estimation can be used to check if the balances found in the glass-ceiling of the ANN (Chapter 4) are potential high flux or not. However, flux value simulated by the model may not have a good agreement with experiments.

5.4 Conclusion

The parameter estimation is the process of trying to find the best parameter of the model using the experimental data. In this study, parameter estimation was performed for the upper part of glycolysis using 31 experiments with different enzyme concentrations. By the iterative approach, the kinetic parameters of PGI, PFK, FBA, TPI and G3PDH are estimated. The final best model obtained after the different steps of parameter estimation has an RMSE of 1.91 between the experimental flux and the model simulated flux. This study showed that even when the experimental kinetic parameters are available, and assuming that they are exact, it is not always easy to fit the model to experimental data to predict the behaviour and replicate the system. If there are too many parameters to be estimated, the process of estimation is computationally expensive, and besides, the existence and the uniqueness of an optimal solution are not guaranty. Even though the model obtained after series of parameter estimation, had a better efficiency in terms of RMSE (1.91) to predict flux, this would not be enough to meet our objective which was i. to use the kinetic model to predict accurately flux for different enzyme concentration, ii. then, utilise these newly predicted fluxes, associated enzyme concentrations, for training the ANN model and extend its range of application in terms of flux prediction. Nevertheless, the final kinetic model obtained in this study is good enough as a

guide to check the trend in terms of high flux: this could help during industrial process optimisation.

Chapter 6 The State of Art in the Malic Acid Synthesis

6.1 Introduction

Malic acid (MA) is a four-carbon dicarboxylic acid that is an intermediate in the Tricarboxylic acid (TCA) cycle. It is one of the major acid components in fruits and plants (Battat, Peleg, Bercovitz, Rokem, & Goldberg, 1991). The US department of energy classified malic acid as one of the top 12 building block chemicals. Malic acid has been widely used as an acidulant and taste enhancer/modified in the food and beverage industry and has also found its application in metal cleaning, textile industries, water treatment and fabric dying (Cheng, Zhou, Lin, Wei, & Yang, 2017; Wei, Cheng, Lin, Zhou, & Yang, 2017; K. Zhang, Zhang, & Yang, 2013). The polymer of β-L-malic acid (poly malic acid) is used as the biodegradable plastic (Cheng et al., 2017; Chi et al., 2016; Y. K. Wang, Chi, Zhou, Liu, & Chi, 2015). L-malic acid is used as an amino acid precursor for the treatment of hyperammonemia and liver dysfunction (Rosenberg, Miková, & Krištofíková, 1999).

In this chapter, I review different methods, pathways, microorganisms and substrates used for the malate synthesis.

6.2 Methods of Malic Acid Synthesis

Malic acid contains two optical isomers, i.e., D-, L- and the mixture of DL- isomers. Different types of acid show different properties. The D- isomer is produced synthetically, whereas L-isomer is observed naturally. L- isomer is an intermediate of cell metabolism. MA is produced by three methods:

6.2.1 Chemical Synthesis

Currently, the major portion of malic acid is synthesised by chemical synthesis. The petroleum-derived maleic anhydride, resulting from the oxidation of benzene, n-butane, or fumarate through chemical hydration, produces a mixture of DL-malic acid. This method requires high temperature and pressure which make the process difficult and cost-inefficient. The depletion of petroleum sources and environmental concerns led to finding alternative methods for production (Goldberg, Rokem, & Pines, 2006).

6.2.2 Enzymatic Conversion

Pure L-malic acid is produced by enzymatic conversion of fumarate using immobilised or isolated fumarase. The substrate inhibition and temperature sensitivity of fumarase make it difficult to industrial-scale production (Giorno, Drioli, Carvoli, Cassano, & Donato, 2001; Kimura, Kawabata, & Sato, 1986).

6.2.3 Fermentation

In 1924, malic acid is identified as a fermentation product by Dakin (Dakin, 1924). From then, many species are identified, characterised for the production of malic acid from various feedstocks such as lignocellulosic biomass (corn, sugarcane), xylose, *etc* (Mondala, 2015; Wei et al., 2017; Xia, Xu, Hu, & Liu, 201s6; T. Zambanini et al., 2016; Zou, Wang, Tu, Zan, & Wu, 2015). The sustainable, eco-friendly feedstock usage and the drawbacks of other methods encouraged developing microbial fermentation for malate production.

6.3 Biosynthesis Pathways for Malic Acid Synthesis

The L-malic acid is identified as one of the intermediates in biological pathways including bacteria, fungi and C4-plants (Ludwig, 2016; Neufeid, Peleg, Rokem, Pinest, & Goldberg, 1998). Four major pathways are recognised (Figure 6.1) for the biosynthesis of L-malic acid (Zelle *et al.*, 2008). All four pathways share the common precursor pyruvate. The pyruvate is synthesised from glucose through the central carbon metabolism pathway, glycolysis.

6.3.1 Oxaloacetate Reduction Pathway

The oxaloacetate reduction is a cytosolic pathway, also known as the reductive tricarboxylic acid (rTCA) pathway. First, pyruvate is gets carboxylated to oxaloacetate (OAA) by pyruvate carboxylase (PYC) followed by the reduction of OAA to malic acid in the cytosol by malic enzymes (malate dehydrogenase, MDH). The OAA reduction pathway is an ATP neutral pathway, that produces 2 mols of malate per mol of glucose with fixation of 2 mol of CO_2 per mol of glucose. The rTCA pathway is the most economic and widely studied in many organisms for the synthesis of malic acid.

In 1983, Osmani suggested the existence of a cytosolic pathway in *Aspergillus flavus* (Stephen A. Osmani and Michael C. Scrutton, 1983) and identified that PYC is important for malic acid

synthesis (Peleg, Stieglitz, & Goldberg, 1988) which is found only in the cytosol in *A. flavus* and *Aspergillus niger* found both in cytosol and mitochondria. Peleg *et al.* in 1988 found higher malate dehydrogenase (MDH) activity than fumarase during malic acid production. Later in 1989 from ^{13}C, NMR study proved that malate is majorly synthesised from the rTCA pathway in *A. flavus*. Currently, this pathway is introduced into many microorganisms to produce malic acid (Peleg, Barak, Scrutton, & Goldberg, 1989).

6.3.2 Tricarboxylic Acid Cycle (TCA cycle)

TCA is the second most studied pathway for malic acid synthesis. Oxaloacetic acid and acetyl CoA produced from pyruvate is catalysed to citric acid in mitochondria followed by several oxidation reactions to form malate. This pathway produces a maximum of 1 mol malate/mol of glucose with the release of 2 CO_2 per mol of glucose (Zelle *et al.*, 2008).

The studies showed that, only deleting the mitochondrial MDH which reverse catalyse the malate to oxaloacetate, is not enough for improved malate production *via* the TCA cycle (Trichez *et al.*, 2018). Along with these enzyme deletions, over-expression of malate insensitive phosphoenolpyruvate carboxylase (Ppc_{K620S}), and inactivation of the acetate pathway improves the malate synthesis (Trichez *et al.*, 2018).

6.3.3 Glyoxylate Cyclic Pathway

The glyoxylate cyclic pathway produces 1 mol of malate per mol of glucose (Zelle *et al.*, 2008). The enzyme from glyoxylate pathway iso-citrate lyase (ICL) and enzyme from TCA, isocitrate dehydrogenase (IDH) are competing for the same substrate isocitrate. This pathway is repressed by the high concentration of glucose, which makes it not a good fit for malic acid synthesis (Dai *et al.*, 2018).

6.3.4 Glyoxylate Noncyclic Pathway

The glyoxylate non-cyclic pathway involves the same enzymes as that of the cyclic pathway. The non-cyclic glyoxylate cycle results in a theoretical maximum yield of 1.33 mol/ mol of glucose, because of the replenishing of oxaloacetic acid by pyruvate carboxylation.

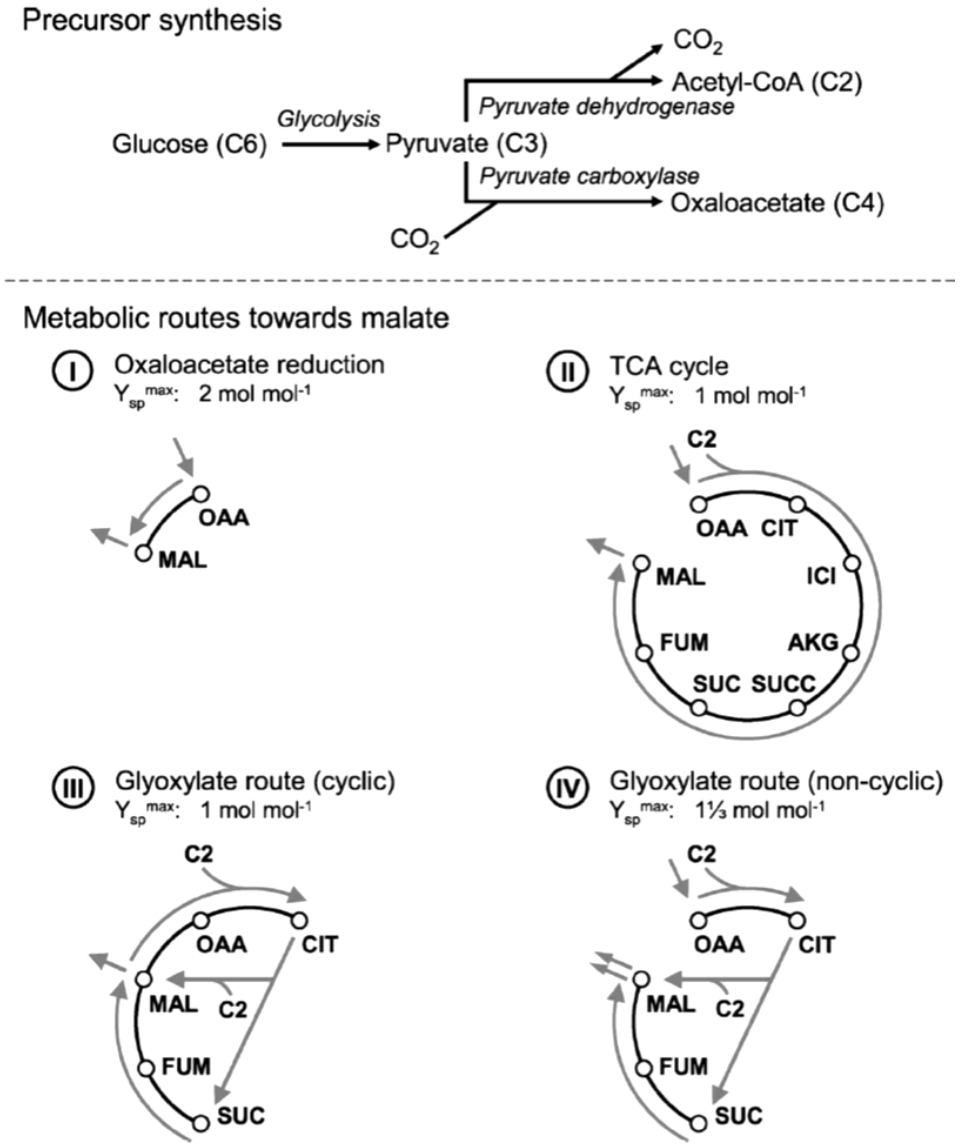


Figure 6.1: Four possible biosynthesis pathways for the production of malic acid synthesis as described in Zelle *et al.* (Zelle *et al.*, 2008) i) oxaloacetate reduction pathway, ii) tricarboxylic acid pathway, iii) glyoxylate pathway, iv) glyoxylate noncyclic pathway. The precursor oxaloacetate is produced from pyruvate *via* glycolysis.

6.3.5 Secretion of Malic Acid

Accumulation of high concentration of malic acid is toxic to cells, and need to be transported across the plasma membrane. Many dicarboxylic acid transporters are identified in various microorganisms (Day & Hanson, 1977; Manuela & Leao, 1990; Osothslip & Subden, 1986; Zoglowek, Krömer, & Heldt, 1988). The transporter from *Schizosaccharomyces pombe* has shown higher activity in malate transport and heterologous expression resulted in a higher yield of malate (Osothslip & Subden, 1986).

6.4 Microorganisms for the Production of Malic Acid

Traditionally, malic acid was extracted from apple juice which comprises 0.4-0.7% of malate and extraction from eggshells requires high energy consumptions and heavy pollution (Lin *et al.*, 2012). Many microorganisms have been identified and developed for malic acid synthesis. Microbial fermentations have advantages of using different raw materials such as glycerol (West, 2015a), soybean (Cheng *et al.*, 2017) or xylose (Z. Li, Hong, Da, Li, & Stephanopoulos, 2018), along with the major substrate glucose.

Aspergillus* species*: *Aspergillus* species has advantages of utilising various sugars such as glucose, xylose (Begum & Alimon, 2011; Cardoso Duarte & Costa-Ferreira, 1994; Prathumpai *et al.*, 2003) and, is used for the production of many organic acids, like malic acid with high yields and production rates (Table 6.1).

In *A. flavus,* the malic acid is synthesised *via* reductive carboxylic acid cycle and a maximum 76% of theoretical yield was obtained (Peleg et al., 1988). *A. flavus* accumulates large amounts of malic acid when cultivated on media containing a high concentration of glucose, limited nitrogen and $CaCO_3$ which supplies CO2 for the pyruvate carboxylase. However, the aflatoxin production disqualifies the food-grade chemical production from *A. flavus* (Battat et al., 1991; Geiser, Pitt, & Taylor, 1998).

The other *Aspergillus* species such as *A. oryzae* or *A. niger* are proved to be "Generally Recognized as Safe" for industrial-scale production and usage of products in food and pharmaceuticals (Abe, Gomi, Hasegawa, & Machida, 2006; Andersen, Nielsen, & Nielsen, 2008). Brown *et al.* showed that by over-expressing cytosolic malate dehydrogenase, pyruvate carboxylase and dicarboxylate transporter, the higher malic acid yield of 1.38 mol/mol could be obtained in *Aspergillus oryzae* (Knuf *et al.*, 2013). However, the yield, titre and productivity from other species are lower compared to *A. flavus*.

The availability of annotated genome sequences helps to develop gene-editing strategies and methods. However, *Aspergillus* species have disadvantages such as lack of self-replicating vectors, and low transformation efficiencies unlike *Escherichia coli* or *Saccharomyces cerevisiae*. The filamentous growth of fungus *Aspergillus* decreases the oxygen supply to the fermentation system (Klement & Büchs, 2013), which makes it difficult for product formation and industrial-scale production.

Table 6.1: Comparison of malic acid production in various Aspergillus species.

Substrate	Organism	Culture condition				Pathway	Engineering strain	Results			Reference
		pH	Temperature in °C	Time (h)	Operating mode			Malic acid (g/l)	Yield (mol mol−1)	Productivity (g/l/h)	
Glucose 98 g/l	*Aspergillus flavus*	-	-	6-8days	-	rTCA	-	36	-	-	(Peleg et al., 1988)
Glucose 120 g/l	*A. flavus*	-	32	190h	Stirred fermenter	TCA	-	113	-	0.59	(Battat *et al.*, 1991)
Glucose 100 g/l	*Aspergillus oryzae* NRRL 3488	6.9	34	164	-	rTCA	Overexpression of C4T318, mdh3, and PYC	154	1.38	0.94	(Knuf *et al.*, 2013)
					-		Overexpression of C4T318 transporter	122	1.17	0.74	
Thin stillage	*A. niger* ATCC 9142	-	-	-	-	-	-		0.8 g/g	-	(West, 2011)
Glycerol	*Aspergillus niger* ATCC 12846	-	25	192	-	-	-		-	-	(West, 2015b)
Syngas* 15.84 g/l	*Aspergillus oryzae*	5.9	37	-	Stirred tank reactor	-	-	4.34	-	0.27 g/g	(Oswald, Dörsam, Veith, Zwick, & Neumann, 2016)
Glucose 110 g/l	*Aspergillus oryzae*	-	-	-	3L- fed-batch culture	rTCA	Overexpression of PYC, MDH, PFK and heterologous expression of, malate permease, overexpression of PEP carboxykinase, PEP carboxylase	165	-	1.38	(J. Liu et al., 2017)

*Synthetic gas (syngas) is a mixture of hydrogen and carbon monoxide which is produced during the gasification of biomass and waste streams (Hammerschmidt *et al.*, 2011; Rokni, 2015)

Escherichia coli: *E. coli* has proved to be an excellent platform for the production of many chemicals using metabolic engineering. In *E. coli,* rTCA and TCA pathways are studied for the malic acid synthesis. Moon *et al.,* using the metabolic flux analysis, showed that by incorporating Phosphoenolpyruvate (PEP) carboxykinase (pckA) into *E. coli*, it can achieve high malic acid production (Soo Yun Moon; Soon Ho Hong; Tae Yong Kim; Sang Yup Lee, 2008). Prevention of acetic acid formation, using pta mutant alone, achieved 1.42 g/l of malic acid. Whereas, by inserting pckA from *Mannheimia succiniciproducens,* a production of 9.25 g/l of MA is achieved (Table 6.2).

Saccharomyces cerevisiae: In *S. cerevisiae,* malic acid is produced mainly by the rTCA pathway in the cytosol, which converts pyruvic acid to oxaloacetate. Overexpression of the pyruvate carboxylase, malate dehydrogenase and heterologous expression of malate transporter from the *S. pombe,* increased the yield of malate in *S. cerevisiae* (Table 6.3).

Other microorganisms: Other than the classical producers of malate, many other organisms are used for the malic acid synthesis (Table 6.4).

The vitamin-auxotroph of *Torulopsis glabrata* is a well-established microorganism, used for the industrial production of pyruvate. By overexpressing pyruvate carboxylase, malate dehydrogenase and the transporter SpMAE1, accumulation of 8.5 g/l malate concentration was achieved in the engineered strain (Xiulai Chen *et al.*, 2013). *Thermobifida fusca* is aerobic, moderately thermophilic, filamentous soil bacterium with high activity and thermostability of cellulose with a wide pH range. Using *T. fusca,* 62.76 g/l of malate was obtained from cellulose (Deng, Mao, & Zhang, 2016). *Zygosaccharomyces rouxii* uses the medium with 300 g/l of glucose and glutamic acid to produce 75 g/l of malate (Taing & Taing, 2007). *Ustilago trichophora* TZ1: The yeast-like growing smut fungus was used to produce the malic acid from glycerol. *Pichia pastoris* is a methylotrophic yeast which has a better efficient heterologous gene expression, physical robustness and well-established fermentation. The integration of the pyruvate carboxylase gene (pyc), the cytoplasmic malate dehydrogenase gene (mdh1) into the chromosomal DNA of *P. pastoris* GS115 produced the malic acid at 42.28 g/l.

Table 6.2: Comparison of Malic acid synthesis in Escherichia coli.

Substrate	Organism	culture condition				Engineering strain	Results			Reference
		pH	Temperature in °C	time (h)	Operating mode		Malic acid (g/l)	Yield (mol mol−1)	Productivity (g liter−1 h−1)	
Glucose	*E coli K-12*	7	37	-	batch culture	deletion of MDH, ME; overexpression of PEP carboxylase + (inactivation of acetate pathway or NADH-insensitive CS mutant)	-	0.82	-	(Trichez *et al.*, 2018)
Glucose	*E coli*	6.7	37	12h	5L bioreactor (aerobic)	PEP carboxykinase pckA from *M. succiniciproducens;* pta mutant	9.25	0.56	0.74	(Jantama *et al.*, 2008)
Glucose	*E coli (XZ658)*	-	-	-	-	E. coli ATCC 8739 (ldhA ackA adhE pflB)	163 mM	1	-	(X. Zhang, Wang, Shanmugam, & Ingram, 2011)
		-	-	72	-		34	1.42	0.47	
Glucose	*E. coli SGJS115*	-	37	9	flask	PEP carboxylase	-	9.90%	-	(Park, Chang, Jin, Pack, & Lee, 2013)
Glucose	*E coli*	5.5	30	48	-	synthetic scaffold between pyc and sfcA	5.72	-	-	(Somasundaram, Eom, & Hong, 2018)
		-	-	-	-		30.2			
Glucose	*E. coli strain WGS-10*	-	-	12h	aerobic	pta mutant, PEP carboxykinase from *M. succiniproducens*	9.25	0.75	0.74	(Soo Yun Moon; Soon Ho Hong; Tae Yong Kim; Sang Yup Lee, 2008)
Xylose	*E. coli*	-	37	72h	shake flask cultivations; aerobic	maeA, maeB, mdh, fumA, fumC, and fumB were knocked out; overexpressing dte from *Pseudomonas cichorii* and *E. coli*: fucK, fucA, aldA, glcDEF, glcB, and katE	5.90	0.80 g/g	-	(Z.-J. Li *et al.* 2018)

Table 6.3: Comparison of engineered Saccharomyces cerevisiae for the production of malic acid.

Substrate	organism	culture condition				pathway	Engineering strain	Results			Reference
		pH	Temperature in °C	time (h)	Operating mode			Malic acid (g litre-1)	Yield (mol mol−1)	Productivity (g liter−1 h−1)	
Glucose 189 g/l	*Saccharomyces cerevisiae*	~8	30	-	calcium carbonate-buffered shake flask cultures.	-	overexpression of pyruvate carboxylase, high-level expression of cytosolic mdh, heterologous expression of malate transporter	59	0.42	-	(Zelle *et al.*, 2008)
Glucose 100 g/l (556 mM)	*Saccharomyces cerevisiae*	5	30	82h	batch culture	-	-	268 mM	0.48	-	(Zelle, Hulster, Kloezen, Pronk, & Maris, 2010)
Glucose 20 g/l	*Saccharomyces cerevisiae*	-	-	48	shake flask fermentation	rTCA	pyc: *A. flavus*; mdh: *R. oryzae;* *S*pma: *S. pombe;*	30.25	0.3 g/g	-	(Xiulai Chen et al., 2017)

Table 6.4: Alternative organisms used for malic acid synthesis.

Substrate	organism	culture condition				pathway	Engineering strain	Results			Reference
		pH	Temperature in °C	time (h)	Operating mode			Malic acid (g litre-1)	Yield (mol mol−1)	Productivity (g liter−1 h−1)	
Glucose 60 g/l	*Torulopsis glabrata*	-	30	60	flask	-	pyc, mdh, transported overexpression	8.5	0.17 g/g	-	(Xiulai Chen *et al.*, 2013)
Cellulose 100 g/l	*Thermobifida fusca*	-	-		-	rTCA	-	62.76	-	-	(Deng et al., 2016)
dry corn stover 50 g/l (18.45 g/l glucan and 9.65 g/l xylan).		-	-		-	-	-	21.47	-	-	
Glycerol 250 g/l	*Ustilago trichophora TZ1*	-	30	-	shake flask			196	0.82 g/g	0.39	(Zambanini, Sarikaya, et al., 2016)
Glycerol 250 g/l	*Ustilago trichophora TZ1*	6.5	-	-	-	-	-	195	-	1.94	(Zambanini, Kleineberg, et al., 2016)
Glucose 10%	Pichia pastoris		-	-	-	-	pyc, mdh, retarded fumarase	42.28	-	-	(T. Zhang, Ge, Li Deng, Tan, & Wang, 2015)
Glucose 10% (w/v) 100g/l	*Monascus araneosus AI-W9087*	6	37	5 days	-	-	-	28	-	-	(Lumyong & Tomita, 1993)
hydrolysate sugar (from corn) 150 g/l	*Aureobasidium pullulans*		-		repeated batch	-	-	38.6	0.3 g/g	0.4	(Zou *et al.*, 2015)

6.5 Other Techniques Used in The Malic Acid Production

Along with the traditional gene knockouts, heterologous expression, other techniques are used in malic acid synthesis such as photochemical, electrical and enzymatic conversion (Amao & Ishikawa, 2007; Inoue, Yamachika, & Yoneyama, 1992; H. Zheng, Ohno, Nakamori, & Suye, 2009). All these methods have not yet matured for industrial-scale production.

Recently, one-pot biosynthesis of malic acid had attracted research interest (T. Shi, Liu, & Zhang, 2019; Xiaoting Ye, Honda, Morimoto, Okano, & Ohtake, 2013). The enzymes from the thermophilic enzymes, which are stable at high temperatures, were used to design the pathway for malic acid synthesis. Using glucose as the substrate, 60% molar yield was obtained by Ye *et al.* (Xiaoting Ye *et al.*, 2013), and with maltodextrin as substrate, Shi *et al.* achieved 95.3% of theoretical yield using the ATP balanced synthetic pathway (T. Shi *et al.*, 2019).

6.6 Sources for the Malic Acid Synthesis

A variety of substrates are used to produce malate from simple glucose to complex lignocellulosic biomass. The following are the few examples of the majorly used substrates:

Glucose*:* The glucose is used as the source for the malic acid synthesis in the majority of studies (Table 6.1-Table 6.4). Most of the natural producers and engineered strains are based on the glucose-based pathways. The theoretical yield of malate production from glucose is 1-2 mol/ mol glucose depending on the pathways used for the production, as explained in the "Biosynthesis Pathways for Malic Acid Synthesis*"* sections. But the industrial-scale production of malic acid from glucose is expensive compared to other feedstocks (Xia *et al.*, 2016).

Glycerol: Recent increase in attention for biodiesel, accompanied by the production of around 19 million tons per year of crude glycerol (Anand & Saxena, 2012). Different production processes using glycerol as a precursor have been reported for the production of 1,3-propanediol, polyhydroxyalkanoates, lipids, succinate, citrate or erythritol. One advantage of the microbial conversion of glycerol to C4 dicarboxylic acids, such as malate or succinate, is the possibility of CO2 fixation through the action of pyruvate Carboxylase. The theoretical yield malate from glycerol is 1 mol malate/ mol glucose.

Thin stillage: Thin stillage is the by-product from grain-based ethanol production. It contains a high percentage of glycerol (West, 2011). The ability of malic acid production using thin stillage was studied in *A. niger* strains and *A. flavus* ATCC 13697 (West, 2011). *A. niger* strains produce higher malic acid using thin stillage compared to *A. flavus* using glucose. Lower Malic acid production could be due to slower utilisation of source in *A. flavus.*

Xylose: Xylose is the second most abundant sugar in lignocellulosic biomass (Nieves, Panyon, & Wang, 2015). A novel pathway for malic acid synthesis was engineered using *E. coli* to utilise xylose as the substrate. The pathway was constructed by overexpressing D-tagatose 3-epimerase, L-fuculokinase, L-fuculose-phosphate aldolase, aldehyde dehydrogenase and, malate consuming enzymes, *i.e.* malic enzymes, malate dehydrogenase and fumarate reductase, were knocked out. The hydrogen peroxide (H_2O_2), which is toxic to the cell was produced during the synthesis of malate from xylose (Demple, Halbrook, & Linn, 1983; Imlay, Chin, & Linnt, 1986). The H_2O_2 activity was decomposed by overexpressing catalase which in turn improved the cell growth and malate production up to 90% of maximum theoretical yield (Z. Li *et al.*, 2018).

Corn: The corncob consists of 35% of hemicellulose. The concentrated 150g/l corncob hydrolysate contains approximately 96 g/l glucose, 54 g/l xylose, 0.14 g/l furfural, 2.34 g/l, HMF, 0.1 g/l formic acid and 1.8 g/l acetic acid (Zou *et al.*, 2015). Using 110 g/l sugar equivalent of corncob hydrolysate, 36.24±0.65 g/l of malic acid was produced (Zou *et al.*, 2015)

6.7 Conclusion

Malic acid is a four-carbon dicarboxylic acid used mainly as an acidulant in the food and beverage industry. Currently, malic acid is mainly synthesised by the chemical process, using the petroleum-derived substrate. The environmental concern and depletion in non-renewable resources encouraged to find an alternative eco-friendly method of synthesis. Many organisms are identified, characterised and engineered for the malic acid synthesis. However, industrial-scale production using malic acid has not yet been achieved. Recently, malic acid was produced *via* a synthetic pathway in a cell-free system using thermophilic enzymes from maltodextrin. This system was able to produce 95% of the maximum theoretical yield of malate. This proves that malic acid can be synthesised by a cell-free system and, by optimising the pathway, can achieve an already high MA production even if not full. Hence in this thesis, an attempt was made to model malic acid synthesis via cell-free synthetic pathway. The main objective is to develop a model which can be used as a plugin to test different substrates and experimental condition.

Chapter 7 Modelling of the Cell-Free System for Synthesis of Malic Acid.

7.1 Context

Malic acid (MA) is a C4-dicarboxylic acid, mainly used as an acidulant in the food and beverage industries. Malate is an intermediate of Krebs cycle which is a deprotonated form of malic acid. Currently, malic acid is produced for commercial use from petroleum-based maleic anhydride. Due to the depletion of petroleum products and environmental concerns encouraged scientists towards synthesising chemicals through the biological system. In Chapter 2, malic acid synthesis using different microorganisms (natural and/or engineered), sources, and techniques developed are discussed in detail. Even if many studies have been performed to optimise the microorganisms for producing malic acid, achieving theoretical maximum yield still remain a challenging goal. Brown *et al.*, by over-expressing pyruvate carboxylase, malate dehydrogenase and native C4-dicarboxylate transporter in *Aspergillus oryzae* NRRL 3488, achieved 1.39 mol malate per mol of glucose (Brown *et al.*, 2013). Zhang *et al.* produced 1.42 mol of malate/mol of glucose by adopting two-stage bioprocess for *Escherichia coli*(X. Zhang et al., 2011). This is the highest malate yield achieved so far by using microorganisms, however, it is far from the maximum theoretical yield of 2 mol of malate/ mol of glucose.

Metabolic engineering and synthetic biology are one of the promising approaches for sustainable production of chemicals by cell-based systems. The emergence of the field synthetic biology led to the development of many technologies including biosensors (Kotula *et al.*, 2014; Siciliano *et al.*, 2018), combinatorial transcriptional regulation (Du, Yuan, Si, Lian, & Zhao, 2012) *etc,* in the recent decade. The advantage of using cell-based systems is self-reproduction. Nevertheless, the engineered organisms carry the risk contaminating and affecting humans. Another challenge of cell-based synthetic biology is the requirement of laborious design and testing. The cell-free synthetic biology helps in tackling these risk where the pathway is built outside the living cell. Cell-free systems (CFS) are easy to manipulate, monitor and optimise. CFS adds the advantages of i. high yield by eliminating side reactions, and, ii. high tolerance to toxic intermediates and products (Carlson, Gan, Hodgman, & Jewett, 2012; Dudley, Nash, & Jewett, 2019; Rollin, Tam, & Zhang, 2013). The cell-free systems are emerging as powerful systems for biomanufacturing of proteins, bio-commodities and value-added chemicals (Carlson et al., 2012; Dondapati, Pietruschka, Thoring, Wüstenhagen, & Kubick, 2019; Dudley et al., 2015, 2019; Kojima, Uchiya, Manshio, & Masuda, 2020; Taniguchi, Okano, & Honda, 2017; Ward, Chatzivasileiou, & Stephanopoulos, 2019; Y. H.Percival Zhang, 2010).

Using thermostable enzymes is important for the industrial-scale biomanufacturing because these enzymes eliminate enzyme contaminants from other microbes and prolong the lifetime of enzymes. Recently, enzymes from thermophilic archaea have been discovered and used in cell-free systems. Lately, CFS using enzymes hyperthermophilic archaea were used for the synthesis of malic acid (T. Shi *et al.*, 2019; Xiaoting Ye *et al.*, 2013). Ye *et al.* used glucose as a substrate and achieved 60% of the theoretical maximum yield (Xiaoting Ye *et al.*, 2013) whereas Shi *et al.* achieved 95.3% from maltodextrin (T. Shi et al., 2019). The designed pathway fix 2 mol of CO_2 per mol of glucose, but the enzyme production using *Escherichia coli* produces CO_2. Therefore the total fixation of CO_2 will be less than the theoretical value of 2 mol CO_2 per mol of glucose. Moreover, the temperature of 50 °C could lead to the degradation of intermediates and hence less malate yield. In this study, the ATP-balanced pathway designed by Shi *et al.* for malate production was modelled. The goal of this work is to obtain the optimised model which can be used as a plugin for different substrates such as lignocellulose.

7.2 Materials

7.2.1 Experimental System for the Malic Acid Synthesis

Shi *et al.* designed three pathways for synthesis malate from maltodextrin using hyperthermophilic enzymes in a cell-free system (T. Shi *et al.*, 2019). The pathways design are described in the next section. The basic architecture of the pathway includes 13 enzymes from the hyperthermophilic archaea which adds an advantage of high thermostability. The designed pathways fix 2 mol of carbon per mol of glucose, which is the theoretical maximum.

7.2.1.1 Design of Artificial Synthetic Pathway

The pathway designed for malate production includes three parts:

1. Production of glyceraldehyde-3-phosphate: In the first step, the glyceraldehyde-3-phosphate (G3P) is formed from starch (maltodextrin) by a six enzymes cascade: alpha-glucan phosphorylase (alphaGP), phosphoglucomutase (PGM), ATP-dependent-6-phosphofructokinase (PFK), fructose bisphosphate aldolase (ALD) and triosephosphate isomerase.

2. Production of 3-phosphoglycerate: 3-phosphoglycerate (3PG) is formed from glyceraldehyde-3-phosphate. This step can be achieved in three ways:

   2.a. Classical glycolysis route: This route is catalysed by glyceraldehyde 3-phosphate dehydrogenase (GAPDH) which catalyses the conversion of G3P to 1,3-biphosphoglecerate, and followed by synthesis of 3-phosphoglycerate by phosphoglycerate kinase (PGK). PGK can generate two ATP per glucose molecule.

   2.b. Without ATP generation: This route is catalysed by non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN) which catalyses the conversion of G3P to 3PG without the generation of ATP which is observed in hyperthermophilic archaea.

   2.c. Combination of the above two ways.

3. Production of malate: The 3-PG was converted into 2-phosphoglycerate by phosphoglycerate mutase (PGAM) followed by the formation of phosphoenolpyruvate (PEP) by enolase (ENO). The phosphoenolpyruvate carboxylase (PEPC) was used to convert PEP to oxaloacetate with the fixation of 2 mol of CO_2 per mol of glucose. The malate dehydrogenase (MDH) converts oxaloacetate (OAA) to malate using NADH.

When Part 1 and Part 3 are combined with part 2.a, the entire pathway generates 1 ATP per glucose and the pathway was called ATP excess pathway. When Part 1 and 3 are combined with 2.b., the pathway was called ATP deficit and when part 1 and 3 are combined with 2.a. and 2.b. it was called ATP balanced pathway (Figure 7.1). All the enzymes used in the pathway for synthesis of malate are given in Table 7.1.

7.2.1.2 Cell-Free Synthesis of Malate

Shi *et al.* tested the three designed pathways for malate production (described in section "Design of Artificial Synthetic Pathway"), using 1U/ml of 13-enzymes from hyperthermophilic archaea (Table 7.1) (T. Shi *et al.*, 2019). The ATP-balanced pathway produced the highest malate compared to ATP-excess and ATP-deficit pathways. When the enzyme cocktail increased to 15 U/ml, malate yield of 0.986 mol/mol of glucose was obtained (50% of theoretical yield) (T. Shi *et al.*, 2019). In the experimental system, there was no rate-limiting step observed for the designed pathway.

27.5 mM glucose equivalent maltodextrin was used for the malate synthesis by Shi *et al.* (T. Shi *et al.*, 2019). To enhance the starch utilisation, the maltodextrin was treated with

isoamylase (IA) which produce linear amylodextrin. The isoamylase treatment increased malate yield up to 70.2% of the theoretical yield. 4-gluconotransferase (4-GT) was added at 12 hours to generate more glucose-1-phosphate (G1P) from maltotriose and maltose resulting in an increase of malate to 90.4%. The polyphosphate glucokinase (PPGK) along with 5mM polyphosphate was added at hour 24, to utilise the residual glucose formed by 4GT. After 48 h, 95.3% of the theoretical maximum yield (52.4 mM) was observed (T. Shi *et al.*, 2019).

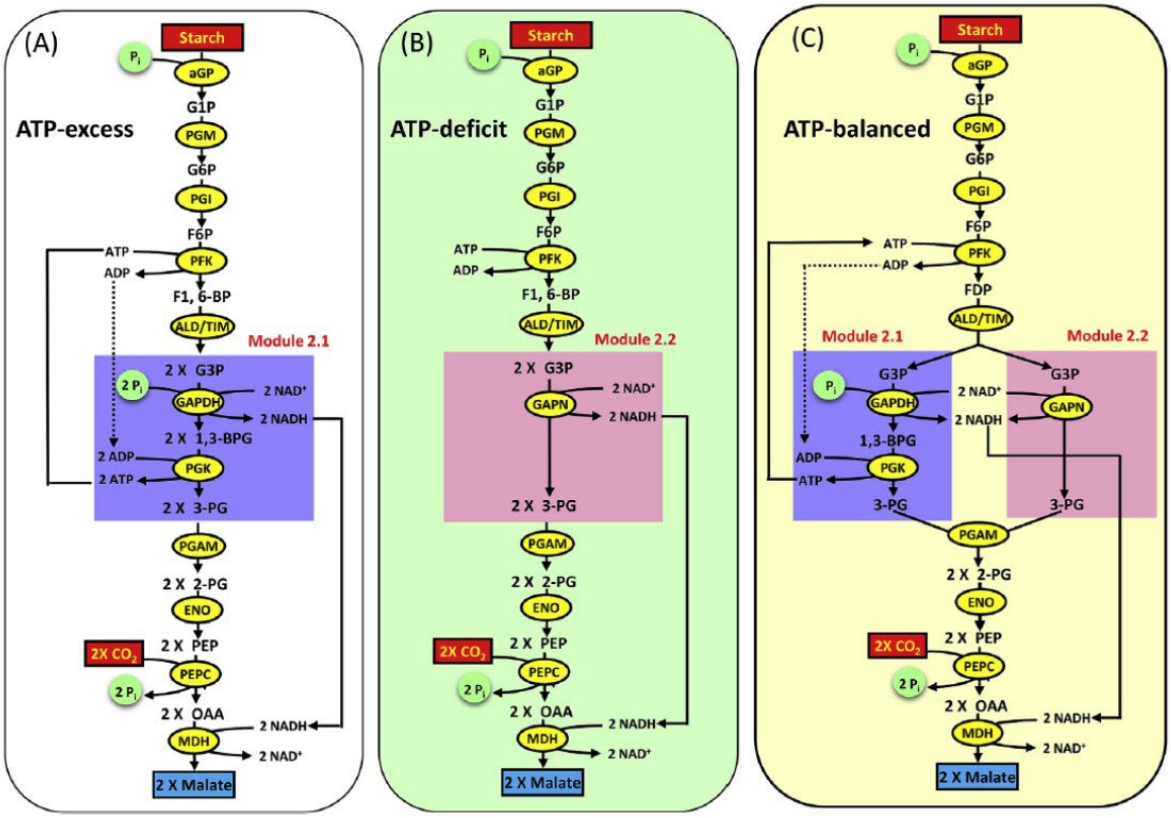


Figure 7.1: The schematic representation of malate synthesis pathway designed by Shi *et al.* (T. Shi *et al.*, 2019). ATP excess pathway (A), ATP-deficit pathway (B) and ATP- balanced pathway (C). The enzymes used are alpha-glucan phosphorylase (αGP), phosphoglucomutase (PGM), 6-phosphate isomerase (PGI), ATP-dependent 6- phosphofructokinase (PFK), fructose-bisphosphate aldolase (ALD), triosephosphate isomerase (TIM), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), phosphoglycerate kinase (PGK), non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN), cofactor-independent phosphoglycerate mutase (PGAM), enolase (ENO), phosphoenolpyruvate carboxylase (PEPC), malate dehydrogenase (MDH). The metabolites are glucose 1-phosphate (G1P), glucose 6-phosphate (G6P), fructose 6-phosphate (F6P), fructose 1,6-diphosphate (F1,6-BP), glyceraldehyde 3-phosphate (G3P), 1,3-diphosphoglycerate (1,3-BPG), 3-phosphoglycerate (3-PG), 2-phosphoglycerate (2-PG), phosphoenolpyruvate (PEP), oxaloacetate (OAA), and inorganic phosphate (Pi).

7.3 Methodology Followed for *In Silico* Modelling of Malate Synthesis

In this study, the malate synthesis pathway was modelled to identify the important regulators and to optimise the pathway. The computational model was built as described in further sections.

7.3.1 Finding Homologous Enzymes

For all enzymes used in the experimental cell-free systems (Table 7.1), kinetic parameters were not available in the literature. Therefore, in this study, the homologous enzymes were selected from which kinetic parameters were available to build the model. To find the homolog of an enzyme for which kinetic information was not available, phylogenetic analysis was performed, using software MEGA X (Molecular analysis Genetic Analysis) (S. Kumar, Stecher, Li, Knyaz, & Tamura, 2018)

Table 7.1: The enzymes used in the cell-free system of malic acid synthesis from hyperthermophiles.

Hyperthermophilic enzymes used in this study.	**Enzyme Abb.**	**Source**	**EC**	**Reaction**
Alpha-glucan phosphorylase	αGP	*Thermotoga maritima* MSB8	2.4.1.1	(1,4-alpha-D-glucosyl) n + phosphate=(1,4-alpha-D-glucosyl) n-1 + alpha-D-glucose 1-phosphate
Phosphoglucomutase	PGM	*Thermococcus kodakarensis* KOD1	5.4.2.2	D-Glucose 1-phosphate=D-Glucose 6-phosphate
Phosphoglucoisomerase	PGI	*Thermus thermophilus* HB27	5.3.1.9	D-Glucose 6-phosphate=D-fructose 6-phosphate
6-phosphofructokinase	PFK	*T. thermophilus* HB8	2.7.1.11	ATP + D-fructose 6-phosphate = ADP + D-fructose 1,6-bisphosphate
Fructose-bisphosphate aldolase	ALD	*T. thermophilus* HB27	4.1.2.13	D-fructose 1,6-bisphosphate = glycerone phosphate + D-glyceraldehyde 3-phosphate
Triosephosphate isomerase	TIM	*T. thermophilus* HB27	5.3.1.1	D-Glyceraldehyde 3-phosphate=glycerone phosphate
Glyceraldehyde-3-phosphate dehydrogenase	GAPDH	*T. maritima* MSB8	1.2.1.12	D-glyceraldehyde 3-phosphate + phosphate + NAD+ = 3-phospho-D-glyceroyl phosphate + NADH + H+
Phosphoglycerate kinase	PGK	*T. thermophilus* HB27	2.7.2.3	ADP + 3-phospho-D-glyceroyl phosphate = ATP + 3-phospho-D-glycerate
Non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase	GAPN	*T. kodakarensis* KOD1	1.2.1.90	D-Glyceraldehyde 3-phosphate + NAD+ + H2O = 3-phospho-D-glycerate + NADH
Cofactor-independent phosphoglycerate mutase	PGAM	*Pyrococcus horikoshii* OT3	5.4.2.12	2-phospho-D-glycerate=3-phospho-D-glycerate
Enolase	ENO	*T. thermophilus* HB27	4.2.1.11	2-phospho-D-glycerate = phosphoenolpyruvate + H2O
Phosphoenolpyruvate carboxylase	PEPC	*T. thermophilus* HB27	4.1.1.31	Phosphoenolpyruvate + CO2 + H2O =Orthophosphate + oxaloacetate
Malate dehydrogenase	MDH	*Archaeoglobus fulgidus*	1.1.1.37	oxaloacetate + NADH = malate + NAD+

The phylogenetic analysis was performed in three steps:

1. Collection of sequences from alternative enzymes: The kinetic parameters for all the enzymes used in the experimental systems by Shi *et al.* were not available in the literature. For those enzymes, where the kinetic parameters were available in the BRENDA database, amino acid sequences were collected. The BRENDA Database is a comprehensive manually curated database of enzyme kinetic data from experiments and literature.
2. Multiple sequence analysis (MSA): MSA was performed for each enzymes using the sequences collected from step 1 using MEGA with MUSCLE algorithm (Edgar, 2004a, 2004b). MUltiple Sequence Comparison by Log-Expectation (MUSCLE) is .a multiple

sequence alignment algorithm which is a combination of local and global alignment proved to perform better than global alignment algorithm ClustalW (Edgar, 2004a, 2004b). The software MEGA X (S. Kumar *et al.*, 2018) default values are used for the sequence alignment with algorithm MUSCLE.

3. Building phylogenetic tree: Once the MSA is built, the phylogenetic tree was constructed using default values of the algorithm maximum likelihood using a bootstrap value of 500 with the Jones-Taylor-Thornton (JTT) model of substitution matrix. Bootstrapping increases the confidence of the phylogenetic tree. The bootstrapping values indicate how many times out of 500 the same branch was observed during the repetition of phylogenetic reconstruction on a re-sampled set of data.

The kinetic parameters from closely related organisms were collected.

7.3.2 Malic Acid Synthesis Model

In this study, the kinetic model for the malate synthesis *via* ATP-balanced pathway was built using CellDesigner (Funahashi *et al.*, 2008, 2003). The k_{cat} values are calculated using the specific activities and molecular weight mentioned as in Shi *et al.*2019 (T. Shi *et al.*, 2019). The K_m values are taken from the BRENDA database for homologous enzymes found through phylogenetic analysis. The model consists of 15 enzymes with mM equivalent of 15U/ml of each enzyme. 27.5 mM glucose equivalent maltodextrin, 2mM ATP, 2mM NAD+, 5mM phosphate and 100mM of HCO3- are used in the model. The enzymes 4GT, PPGK and 5 mM polyphosphate were added according to the experimental set up at hour 12 (add 4GT) and hour 24 (add PPGK and polyphosphate).

The model data was stored in SBML (Systems Biology) format Markup Language, Hucka *et al.* (Hucka et al., 2003) which is a standard for representing networks biochemicals. The SBML format uses an XML architecture (eXtensible Markup Language) adapted to contain all information related to metabolites and biochemical reactions. The format adopted is compatible with most bioinformatics software used for modelling and analysis of biochemical systems.

7.3.3 Estimation of Kinetic Parameters

The SBML model was imported in COPASI to perform further analysis. The model was optimised using the experimental malate concentration at different time points of 48h fermentation measured by Shi *et al.* (T. Shi *et al.*, 2019). In the experimental system, there was no rate-limiting step was observed by Shi *et al.* (T. Shi *et al.*, 2019). Nevertheless, in the kinetic

model, it was observed that GAPDH, GAPN and PGK appear as potential regulators for malate concentration. Indeed, the enzyme GAPN and PGK are involved in the regeneration of NADH and ATP in the system, and GAPN and GAPDH share the substrate which is Glyceraldehyde-3-phosphate (Figure 7.1). So, keeping high and constant concentrations of NADH and ATP should allow higher production of malate at the end of the cascade, and high efficiency of GAPN and GAPDH should increase the concentration of the intermediate Glyceraldehyde-3-phosphate. Therefore, GAPDH, GAPN and PGK parameters were selected to be optimised. To limit the search space, the range between parameter divided by 10 and parameter and multiplied by 10 were used for the estimation. Limiting the search space helps in reducing the time required for finding optimised parameter. The parameter estimation was performed in COPASI using the genetic algorithm, with the default 2000 generations of population size 20. The algorithm attempts to minimise the sum of squares of variation between experimental data and simulated data.

The parameter estimation is performed in the following steps

1. Only Isoamylase treated maltodextrin as substrate: In the first step, the experimental measurement of malate concentration using Isoamyalse treated maltodextrin without 4GT and PPGK was used. This will prevent the effect of 4GT and PPGK on the final malate concentration. The k_{cat} and K_m of GAPDH, GAPN and PGK were fitted with experimentally measured malate concentration. The kinetic model contains 13 enzymes as given in Table 7.1. Three algorithms i.e, Genetic algorithm (evolution-based), Particle swarm and Hooke-Jeeves algorithm (geometry-based) were tested each with three iterations while updating the parameters of the model from the previous iteration.
2. Update kinetic model: The newly estimated parameters were updated to the model.
3. Estimate 4GT and PPGK added model: For the kinetic model obtained after step 2, the 4GT and PPGK were added at the 12th hour and the 24th hour to enhance the starch utilisation. The data from 4GT, PPGK added experimental system was used for the parameter estimation. The parameters of PFK, GAPDH, GAPN, PGK and MDH were selected for the estimation as explained in further sections. The kinetic model was updated with new estimated parameters.

The methodology followed for the optimisation of the kinetic model is summarised in Figure 7.2.

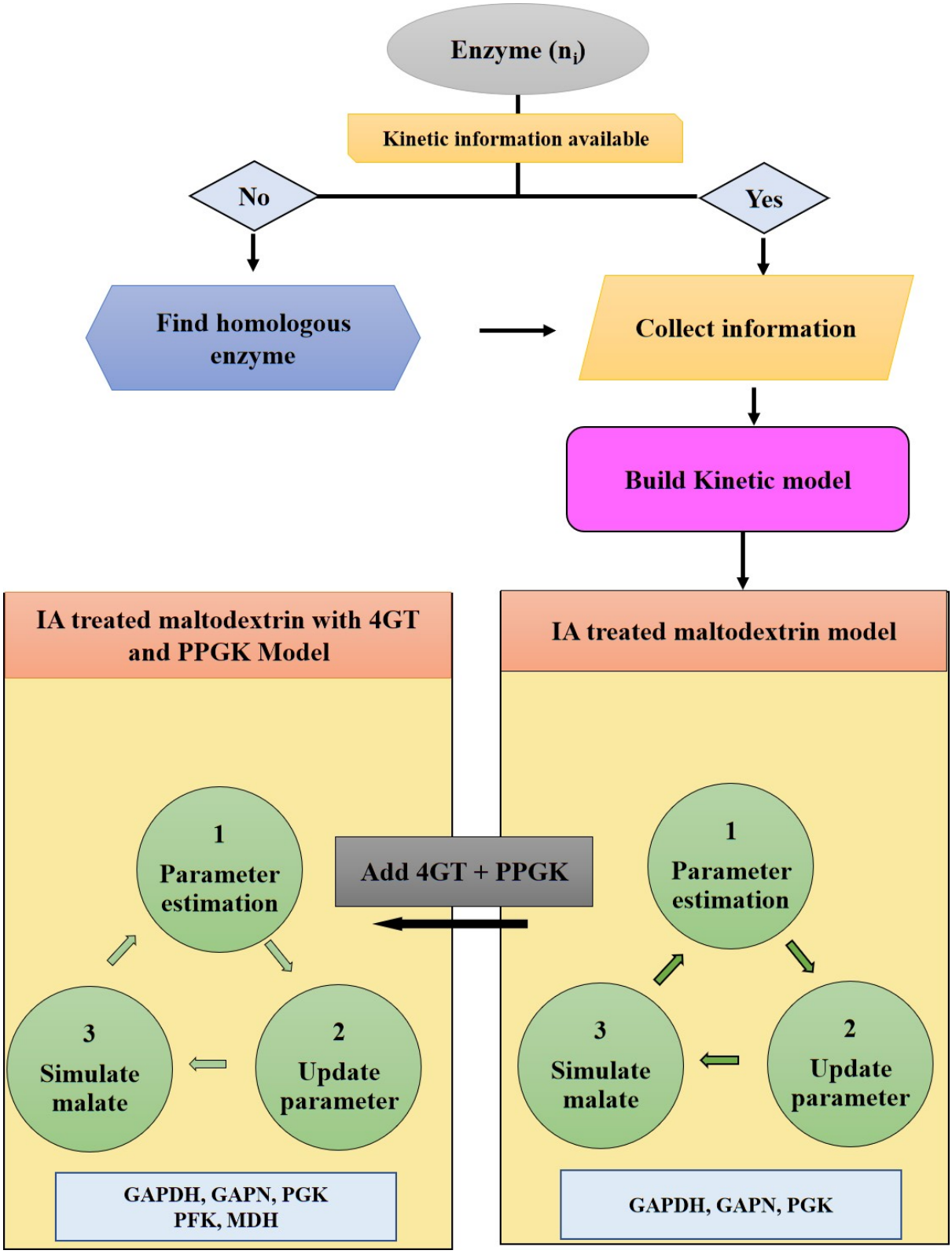


Figure 7.2: The methodology followed for the optimisation of kinetic model for the synthesis of malate.

7.4 Results and Discussion

Shi *et al.* experimentally proved that out of three pathway designed i.e, ATP-excess, ATP-Deficit, and ATP-balanced for malate synthesis, ATP-balanced pathway produced more malate (2.58 mM) compare to the other two i.e. ATP-excess (0.60 mM), ATP-deficit (1.47 mM). Therefore, ATP-balanced pathway was selected to build the computational model.

7.4.1 Homologous Enzymes

To build the kinetic model of the system, it is important to know the kinetic parameters of the enzymes involved in the pathway. For the all hyperthermophilic enzymes used in the study, kinetic parameters were not available. Therefore, to choose the alternative enzyme, phylogenetic analysis was performed. The phylogenetic analysis helps to find the evolutionary relationships between the organisms. Hence, in this study, the alternative enzymes were identified for the enzymes for which the kinetic data was not available. The alternative enzyme parameters were used as the starting point to overcome the problem of lacking data.

The phylogenetic tree (Figure 7.3) showed that all the hyperthermophilic enzymes clustered together. From Figure 7.3, the close homolog of *Pyrococcus horikishii* is *P.* furiosus for which kinetic parameters are available. The result from phylogenetic analysis for all the enzymes are summarised in Table 7.2. The Phylogenetic tree for other enzymes are given in Annexe figures Annexe 13 to Annexe 18.

Figure 7.3: The phylogenetic tree for the enzyme cofactor-independent phosphoglycerate mutase (PGAM). The phylogeny was constructed using organisms from which enzyme parameters are available in the BRENDA database.

According to Table 7.2 (and Annexe 14), for the enzyme Phosphoglucoisomerase (PGI) from *Thermus thermophilus HB27, the* closest homolog found are *Geobacillus stearothermophilus*, *Hungateiclostridium thermocellum* and *Methanocaldococcus jannaschii*. The *G. stearothermophilus* and *H. thermocellum* are thermophilic bacteria whereas *M. jannaschii* is an archaeal species. This could be because of the horizontal transfer of genes between archaea and thermophilic bacteria (Nelson *et al.*, 1999; Rudolph, Hansen, & Schönheit, 2004; P. Wang, Wang, Guo, Huang, & Zhu, 2020). However, further detail study is required to support this conclusion.

Table 7.2: The alternative enzymes found through phylogenetic analysis. In parameter availability, Yes: refers to the availability of kinetic parameters and No refers to lack of kinetic data from the enzymes used in the experimental study. Re No: reaction number; EC: Enzyme commission number.

Re. No	**Hyperthermophile enzymes used in the experiment**	**Enzyme Abb.**	**Source**	**EC**	**Parameter availability**	**Parameters available from homologous enzyme**
1	Alpha-glucan phosphorylase	αGP	*Thermotoga maritima MSB8*	2.4.1.1	No	*Pyrococcus furiosus*
2	Phosphoglucomutase	PGM	*Thermococcus kodakarensis KOD1*	5.4.2.2	Yes	-
3	Phosphoglucoisomerase	PGI	*Thermus thermophilus HB27*	5.3.1.9	No	*Geobacillus stearothermophilus Methanocaldococcus jannaschii Hungateiclostridium thermocellum*
4	6-phosphofructokinase	PFK	*T. thermophilus HB8*	2.7.1.11	Yes	-
5	Fructose-bisphosphate aldolase	ALD	*T. thermophilus HB27*	4.1.2.13	No	*Thermus aquaticus*
6	Triosephosphate isomerase	TIM	*T. thermophilus HB27*	5.3.1.1	No	*Mycobacterium tuberculosis*
7	Glyceraldehyde-3-phosphate dehydrogenase	GAPDH	*T. maritima MSB8*	1.2.1.12	No	*Thermus thermophilus HB27*
8	Phosphoglycerate kinase	PGK	*T. thermophilus HB27*	2.7.2.3	Yes	-
9	Non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase	GAPN	*T. kodakarensis KOD1*	1.2.1.90	No	*Thermoproteus tenax*

10	Cofactor-independent phosphoglycerate mutase	PGAM	*Pyrococcus horikoshii OT3*	5.4.2.12	No	*Pyrococcus furiosus*
11	Enolase	ENO	*T. thermophilus HB27*	4.2.1.11	No	*Chloroflexus aurantiacus* *Brucella abortus*
12	Phosphoenolpyruvate carboxylase	PEPC	*T. thermophilus HB27*	4.1.1.31	No	*Thermosynechococcus elongatus* *Nostoc sp. 7120* *Synechocystis* sp. PCC 6803
13	Malate dehydrogenase	MDH	*Archaeoglobus fulgidus*	1.1.1.37	Yes	-

7.4.2 Construction of Kinetic Model

After finding the alternative enzymes through phylogenetic analysis, kinetic parameters were collected from the BRENDA database and literature Table 7.3.

Table 7.3: Kinetic parameters used in the ATP-balanced pathway for malic acid synthesis. * indicates the kinetic parameters are taken from the enzymes used in the original study by Shi *et al.* (Shi *et al.*, 2019). Re No: reaction number; EC: enzyme commission number; k_{cat}: catalytic constant or turn-over number in s^{-1}; K_m: Michaelis-Menten constant in mM; K_a: association constant in mM; K_{eq}: Equilibrium constant; K_i: inhibition constant in mM.

re No	Enzyme Abb.	Source	EC no	k_{cat} calculated (s-1)	K_m (mM)	K_{eq} (mM)	Parameter from other sources
1	αGP	*Pyrococcus furiosus*	2.4.1.1	6.2	Pi:30; maltotrose: 70	0.3	-
2	PGM	*Thermococcus kodakarensis KOD1* *	5.4.2.2	5	G1P:3.0	20	-
3	PGI	*Methanocaldococcus jannaschii*	5.3.1.9	40.83	G6P: 1; F6P: 0.04	0.361	-
4	PFK	*T. thermophilus HB8**	2.7.1.11	2.5	-	641	KaF6P: 0.027; KaMgATP: 0.006m PFKn = 4; KiATP =
5	ALD	*Thermus aquaticus*	4.1.2.13	8.16	FBP: 0.305	3.2 × 10-4	KmG3P: 0.052; KmDHAP: 0.171 mM from *S. solfataricus*

6	TIM	*Mycobacterium tuberculosis*	5.3.1.1	181.25	DHAP = 0.0025; G3P= 0.084	0.108	-
7	GAPDH	*Thermus thermophilus HB27*	1.2.1.12	47.67	g3p: 0.3; nad: 0.1	0.0765	KmPhosphate: 8.3 frm G. stearothermophilus
8	PGK	*T. thermophilus HB27*	2.7.2.3	96.63		1.8 × 10^3	Km3PG: 0.54; KmADP: 0.085; KmBPG: 5.6; KmATP: 9.7 from *S. solfataricus*
9	GAPN	*Thermoproteus tenax*	1.2.1.90	1.73	G3P: 0.02; NAD:3.1	-	-
10	PGAM	*Pyrococcus furiosus*	5.4.2.12	28	3PG:0.49	0.185	Km2PG: 0.2 from *A. fulgidus*
11	ENO	*Chloroflexus aurantiacus*	4.2.1.11	80	2PG: 0.03	5.19	-
12	PEPC	*Synechocystis sp. PCC 6803*	4.1.1.31	10	PEP: 0.3; HCO3-: 0.8;	4.0 × 10^6	-
13	MDH	*Archaeoglobus fulgidus*	1.1.1.37	35	OAA: 0.043; NADH:0.024	1.3 × 10^5	KmMal: 0.095; KmNAD: 0.14 taken from *T. Thermophilus*
14	4GT	*Thermococus litoralis**	2.4.1.25	33.83	Maltohep: 0.46	-	
15	PPGK	*Thermobifda fusca* *	2.7.1.63	32.3	Glucose: 0.8; Ppi: 0.1	-	

The kinetic model was built for the ATP-balanced pathway designed by Shi *et al.* (T. Shi *et al.*, 2019). The schema of the kinetic model from Celldesigner is given in Figure 7.4, which consists of 15 reactions. The kinetic equations used for each reaction are given in Table 7.4 and kinetic parameters are given in Table 7.3.

Table 7.4: Kinetic equation used for modelling the malic acid synthesis via ATP-balanced pathway in this study.v: rate of reaction; k_{cat}: catalytic constant or turn-over number in s^{-1}; K_m: Michaelis-Menten constant in mM; K_a: association constant in mM; K_{eq}: Equilibrium constant; K_i: inhibition constant in mM.

re No	Enzyme Name	Enzyme Abb.	Kinetic Equation
1	Alpha-glucan phosphorylase	αGP	$v = \frac{\text{kcataGP} * (\text{aGP}) * \left(\frac{\text{maltodextrin}}{\text{aGPKmGlyc}}\right) * \frac{\text{Pho}}{\text{aGPKmPi}}}{\left(1 + \frac{\text{maltodextrin}}{\text{aGPKmGlyc}}\right) * \left(1 + \frac{\text{Pho}}{\text{aGPKmPi}}\right)}$
2	Phosphoglucomutase	PGM	$v = \frac{\text{kcatPGM} * (\text{PGM}) * \frac{\text{G1P}}{\text{PGMKmG1P}}}{\text{PGMKmG1P} + \text{G1P}}$
3	Phosphoglucoisomerase	PGI	$v = \frac{\text{kcatPGI} * (\text{PGI}) * (\text{G6P} - \frac{\text{F6P}}{\text{PGIKeq}})}{\text{PGIKmG6P} * \left(1 + \frac{\text{F6P}}{\text{PGIKmF6P}}\right) + \text{G6P}}$
4	6-phosphofructokinase	PFK	$v = \frac{\text{kcatPFK} * (\text{PFK}) * \left(\frac{\text{F6P}}{\text{PFKKaF6P}}\right) * \left(\frac{\text{ATP}}{\text{PFKKaATP}}\right)}{\left(1 + \frac{\text{F6P}}{\text{PFKKaF6P}} + \left(\frac{\text{PEP}}{\text{PFKKiPEP}}\right)^{\text{PFKn}}\right) * \left(1 + \frac{\text{ATP}}{\text{PFKKaATP}}\right)}$
5	Fructose-bisphosphate aldolase	ALD	$v = \frac{\text{kcatFBA} * (\text{ALD}) * (\text{FBP} - \text{DHAP} * \frac{\text{G3P}}{\text{FBAKeq}})}{\text{FBAKmFBP} * \left(1 + \frac{\text{DHAP}}{\text{FBAKmDHAP}} + \frac{\text{G3P}}{\text{FBAKmG3P}}\right) + \text{FBP}}$
6	Triosephosphate isomerase	TPI	$v = \frac{\text{kcatTPI} * (\text{TPI}) * (\text{DHAP} - \frac{\text{G3P}}{\text{KeqTPI}})}{\text{TPIKmDHAP} * \left(1 + \frac{\text{G3P}}{\text{TPIKmG3P}}\right) + \text{DHAP}}$
7	Glyceraldehyde-3-phosphate dehydrogenase	GAPDH	$v = \frac{\text{kcatGAPDH} * (\text{GAPDH} * \text{default}) * \left(\frac{\text{G3P}}{\text{GAPDHKmG3P}}\right) * \left(\frac{\text{NAD}}{\text{GAPDHKmNAD}}\right) * \left(\frac{\text{Pho}}{\text{GAPDHKmPho}}\right)}{\left(1 + \frac{\text{G3P}}{\text{GAPDHKmG3P}}\right) * \left(1 + \frac{\text{NAD}}{\text{GAPDHKmNAD}}\right) * \left(1 + \frac{\text{Pho}}{\text{GAPDHKmPho}}\right)}$

8	Phosphoglycerate kinase	PGK	$v = \frac{\text{kcatPGK} * (\text{PGK}) * \left(\frac{\text{BPG}}{\text{PGKKmBPG}}\right) * \left(\frac{\text{ADP}}{\text{PGKKmADP}}\right)}{\left(1 + \frac{\text{BPG}}{\text{PGKKmBPG}}\right) * \left(1 + \frac{\text{ADP}}{\text{PGKKmADP}}\right)}$
9	Non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase	GAPN	$v = \frac{\text{kcatGAPN} * (\text{GAPN} * \text{default}) * \left(\frac{\text{G3P}}{\text{GAPNKmG3P}}\right) * \left(\frac{\text{NAD}}{\text{GAPNKmNAD}}\right)}{\left(1 + \frac{\text{G3P}}{\text{GAPNKmG3P}}\right) * \left(1 + \frac{\text{NAD}}{\text{GAPNKmNAD}}\right)}$
10	Cofactor-independent phosphoglycerate mutase	PGAM	$v = \frac{\text{kcatPGAM} * (\text{PGAM}) * \left(\text{P3G} - \frac{\text{P2G}}{\text{PGAMKeq}}\right)}{\text{PGAMKm3PG} * \left(1 + \frac{\text{P2G}}{\text{PGAMKm2PG}}\right) + \text{P3G}}$
11	Enolase	ENO	$v = \frac{\text{kcatENO} * (\text{ENO}) * \frac{\text{P2G}}{\text{ENOKm2PG}}}{1 + \frac{\text{P2G}}{\text{ENOKm2PG}}}$
12	Phosphoenolpyruvate carboxylase	PEPC	$v = \frac{\text{kcatPEPC} * (\text{PEPC}) * \left(\frac{\text{PEP}}{\text{PEPCKmPEP}}\right) * \left(\frac{\text{CO2}}{\text{PEPCKmHCO3}}\right)}{\left(1 + \frac{\text{PEP}}{\text{PEPCKmPEP}}\right) * \left(1 + \frac{\text{CO2}}{\text{PEPCKmHCO3}}\right)}$
13	Malate dehydrogenase	MDH	$v = \frac{\text{kcatMDH} * (\text{MDH} * \text{default}) * \left(\frac{\text{OAA}}{\text{MDHKmOAA}}\right) * \left(\frac{\text{NADH}}{\text{MDHKmNADH}}\right)}{\left(1 + \frac{\text{OAA}}{\text{MDHKmOAA}} + \frac{\text{mal}}{\text{MDHKmMal}}\right) * \left(1 + \frac{\text{NADH}}{\text{MDHKmNADH}} + \frac{\text{NAD}}{\text{MDHKmNAD}}\right)}$
14	4-glucano transferase	4GT	$v = \frac{\text{kcatGT} * \text{GT} * \text{aGlyco}_{\text{n1}}}{\text{GTKmGlycn1} + \text{aGlyco}_{\text{n1}}}$
15	Polyphosphate glucokinase	PPGK	$v = \frac{\text{kcatPPGK} * \text{PPGK} * \left(\frac{\text{Gluc}}{\text{PPGKKmGlu}}\right) * \left(\frac{\text{polyP}}{\text{PPGKKmPolyP}}\right)}{\left(1 + \frac{\text{Gluc}}{\text{PPGKKmGlu}}\right) * \left(1 + \frac{\text{polyP}}{\text{PPGKKmPolyP}}\right)}$

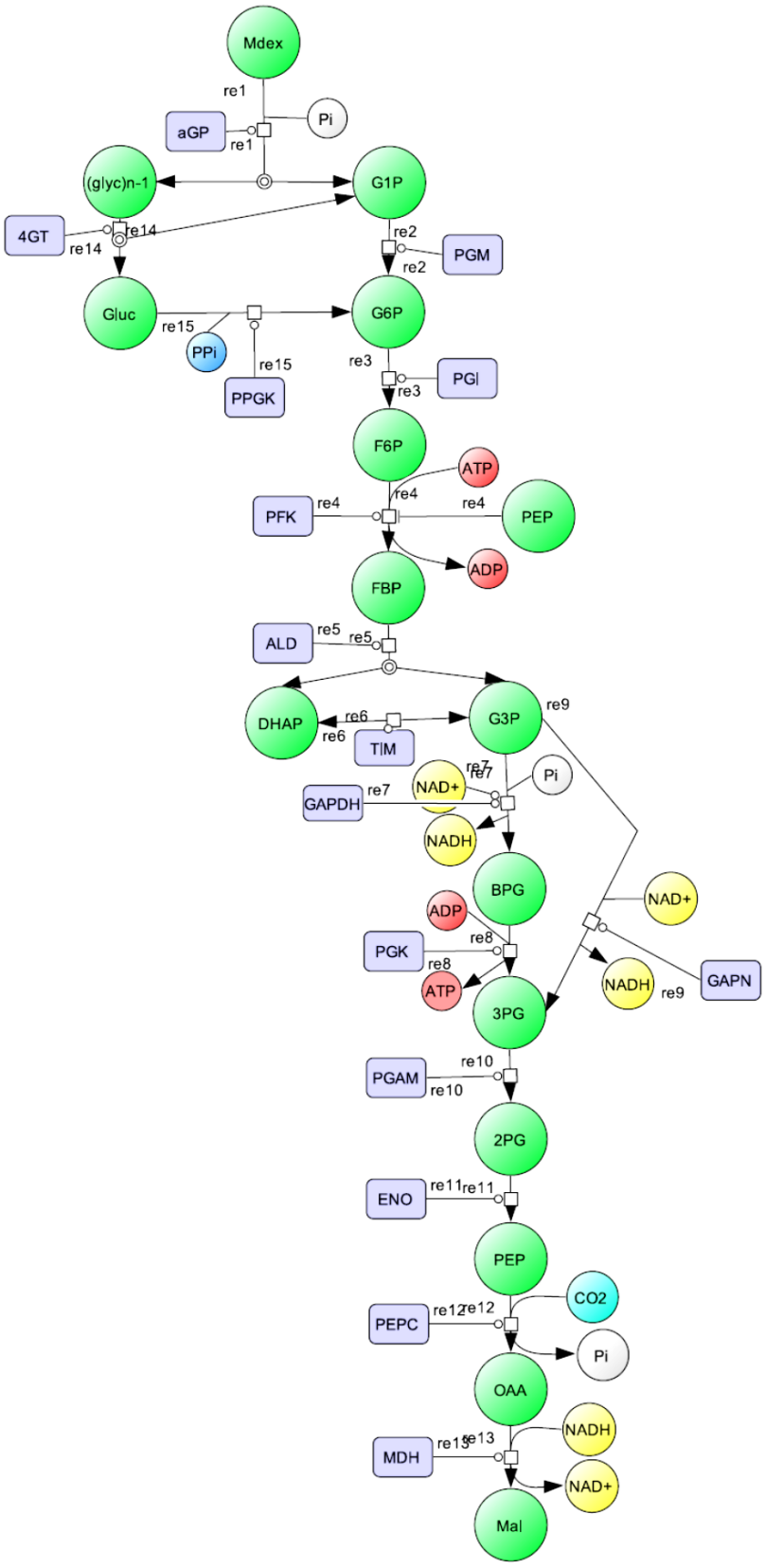


Figure 7.4: Schema of the kinetic model built for the synthesis of malate. The enzymes used are alpha-glucan phosphorylase (aGP), phosphoglucomutase (PGM), 6-phosphate isomerase (PGI), ATP-

dependent 6- phosphofructokinase (PFK), fructose-bisphosphate aldolase (ALD), triosephosphate isomerase (TIM), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), phosphoglycerate kinase (PGK), non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN), cofactor-independent phosphoglycerate mutase (PGAM), enolase (ENO), phosphoenolpyruvate carboxylase (PEPC), malate dehydrogenase (MDH), 4Gluconotransferase (4GT), Polyphosphate glucokinase (PPGK). The metabolites are glucose 1-phosphate (G1P), glucose 6-phosphate (G6P), fructose 6-phosphate (F6P), fructose 1,6-diphosphate (FBP), glyceraldehyde 3-phosphate (G3P), 1,3-diphosphoglycerate (BPG), 3-phosphoglycerate (3-PG), 2-phosphoglycerate (2-PG), phosphoenolpyruvate (PEP), oxaloacetate (OAA), malate (Mal), inorganic phosphate (Pi), polyphosphate (PPi), (1,4-alpha-D-glucosyl)n-1 ((glyc)n-1), Glucose (Gluc)

7.4.3 Optimisation of Kinetic Model

The model, with isoamylase treated maltodextrin without 4GT and PPGK (named model-1), was simulated for 48 hours as in the experimental system. Figure 7.5 shows that the maximum malate produced by the kinetic model was 5 mM and the intermediates of the pathway glucose-6-phosphate (G6P) and fructose-6-phosphate (F6P) were 3mM and 8mM respectively. This indicates that not all the intermediates are converting to the final product, and this could be due to a lack of ATP in the system.

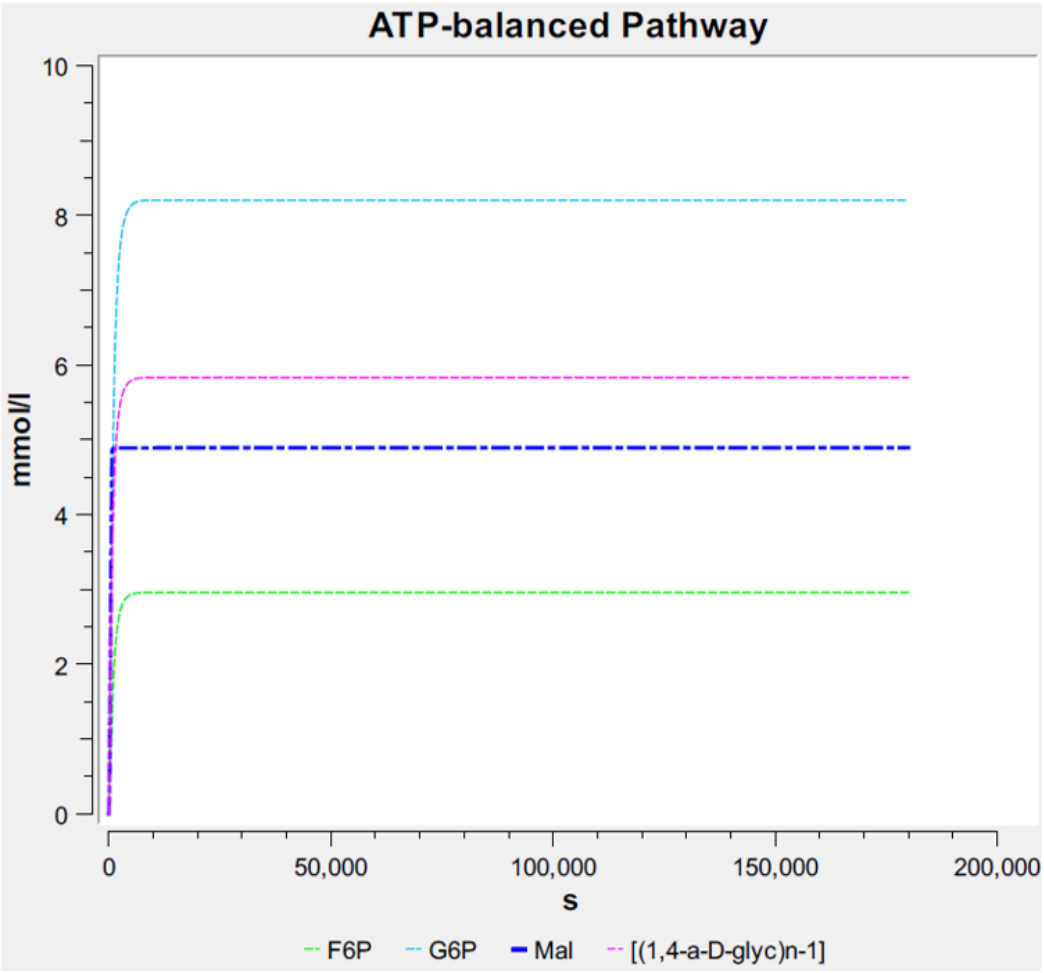


Figure 7.5: The simulated concentration of malate and intermediates via ATP-balanced pathway.[(1,4-a-D-glyc)n-1]: (1,4-alpha-D-glycosyl)n-1; G6P: Glucose 6-Phosphate; F6P: Fructose-6-Phosphate; Mal: Malate.

To check if ATP concentration in the model is limiting the final malate production, the concentration of ATP was varied from 2mM, 5mM to 10 mM. As the concentration of ATP increases in the model, the final malate concentration increases to 4.89, 12.176 and 24.74 mM respectively (Figure 7.6). The decrease in the intermediate concentration (G6P and F6P) is observed too (G6P was 8.20, 8.94, 5.05 mM and F6P was 2.96, 3.22 and 1.8 mM at ATP concentration of 2mM, 5mM and 10mM respectively). This indicates that the ATP concentration was limiting the malate production, and thus by regenerating ATP, the enzymatic system should produce more malate.

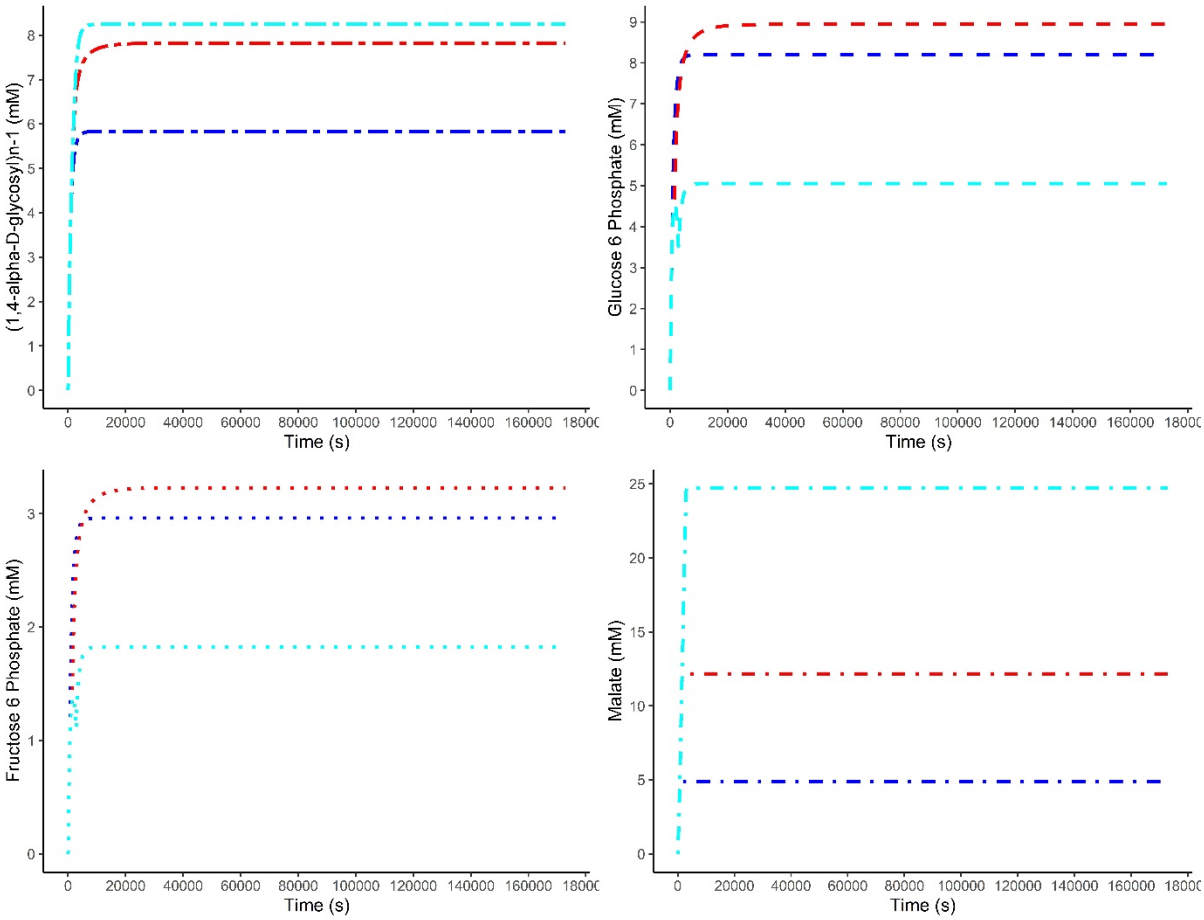


Figure 7.6: The effect of ATP concentration on different intermediates and final malate production in the ATP-balanced pathway for malate synthesis. The line colour represents as follows: Blue: 2mM ATP; red: 5mM ATP: and cyan: 10mM ATP in the model.

The low concentration of malate (5mM) at an ATP concentration of 2mM (Figure 7.5) is much lesser than the experimental condition at the end of 48 hours, which was 52.4mM. This low concentration of malate simulated by the model indicates that the model is not optimised. Therefore, parameter estimation was performed to fit the model to experimental measurement of the malate concentration throughout 48 hours of fermentation. The experimental data points from Shi *et al.* (Annexe 19) was extracted using WebPlotDegitizer (https://automeris.io/WebPlotDigitizer/). The extracted points divided into two groups: test and validation dataset (Figure 7.7).

Two sets of experiments are considered for the parameter estimation, based on the enzymes used for the complete utilisation of maltodextrin. In the first step, the malate only produced by isoamylase treated maltodextrin was utilised (without 4GT and PPGK addition) for the parameter estimation. The kinetic model has only 13 enzymes and parameter estimation was

performed using three different algorithms i.e, Genetic algorithm (GA, evolution-based), Particle swarm (PS) and Hooke-Jeeves algorithm (HJ, geometry-based). For each algorithm, parameter estimation was performed in three iterations while updating the parameters from the previous iteration. The objective function value, root mean square (RMS) between experimental and model-simulated malate concentration between experimental data set and validation set, was chosen to select the best performing algorithm out of three studied algorithms. The GA performed better than the PS and HJ (Annexe 20**).**

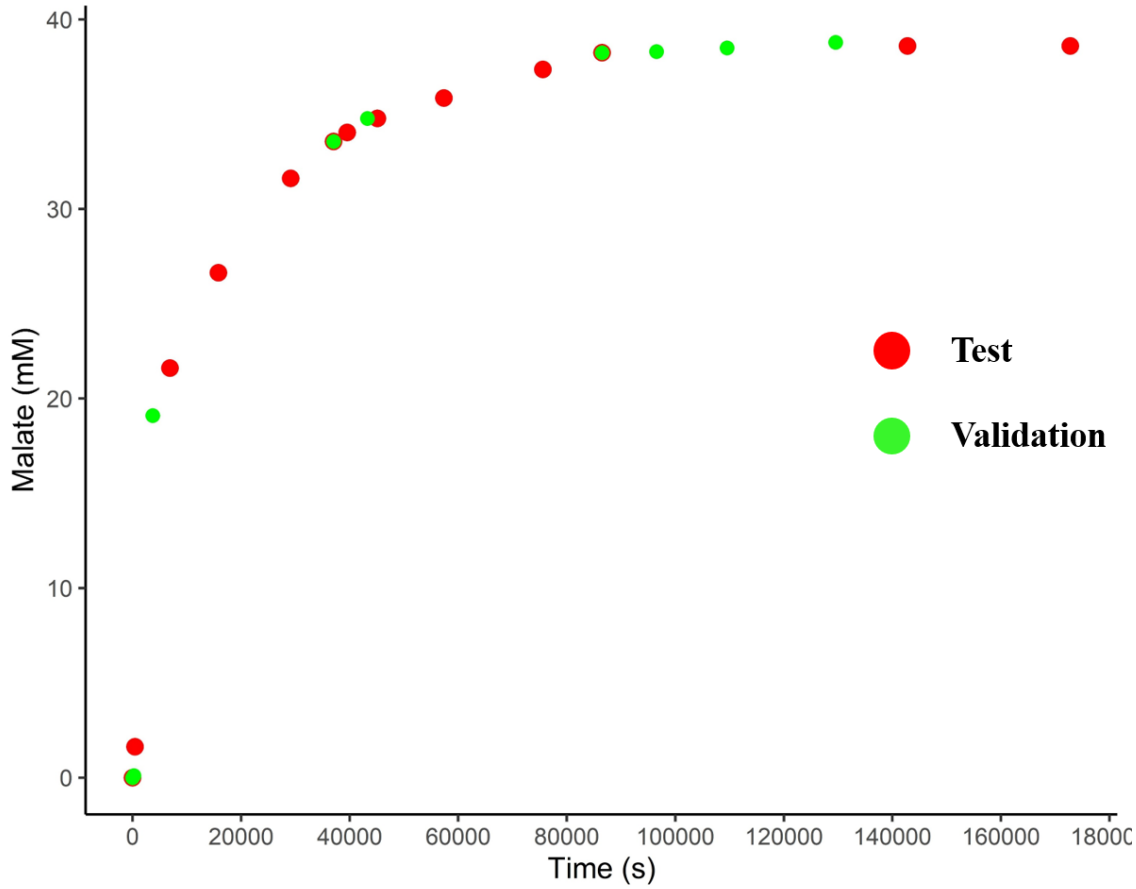


Figure 7.7: The malate concentration used for the parameter estimation of the kinetic model. The data points are extracted from Shi *et.al.* for malate synthesis from isoamylase treated maltodextrin.

The newly estimated parameter from the genetic algorithm (Table 7.5) after three iterations were updated to the model-1 and the model was simulated for 48 h. Figure 7.8 indicates that by optimising the GAPD, GAPN and PGK kinetic parameters (Table 7.5), a higher malate concentration can be obtained. Thus, as expected, by optimising these three enzymes ATP was

regenerated at a higher concentration. The ADP formed at the PFK step now will be regenerated at PGK step. The PGK is the enzyme in the pathway which convert ADP to ATP which is important for complete conversion of F6P to FBP. The GAPN and GAPDH share the same substrate, by optimising these two enzymes, it helps in balancing the flow of substrate and provides enough substrate to the enzyme PGK. The results observed in Figure 7.6 also proves that ATP concentration is important in the kinetic model to obtained higher malate.

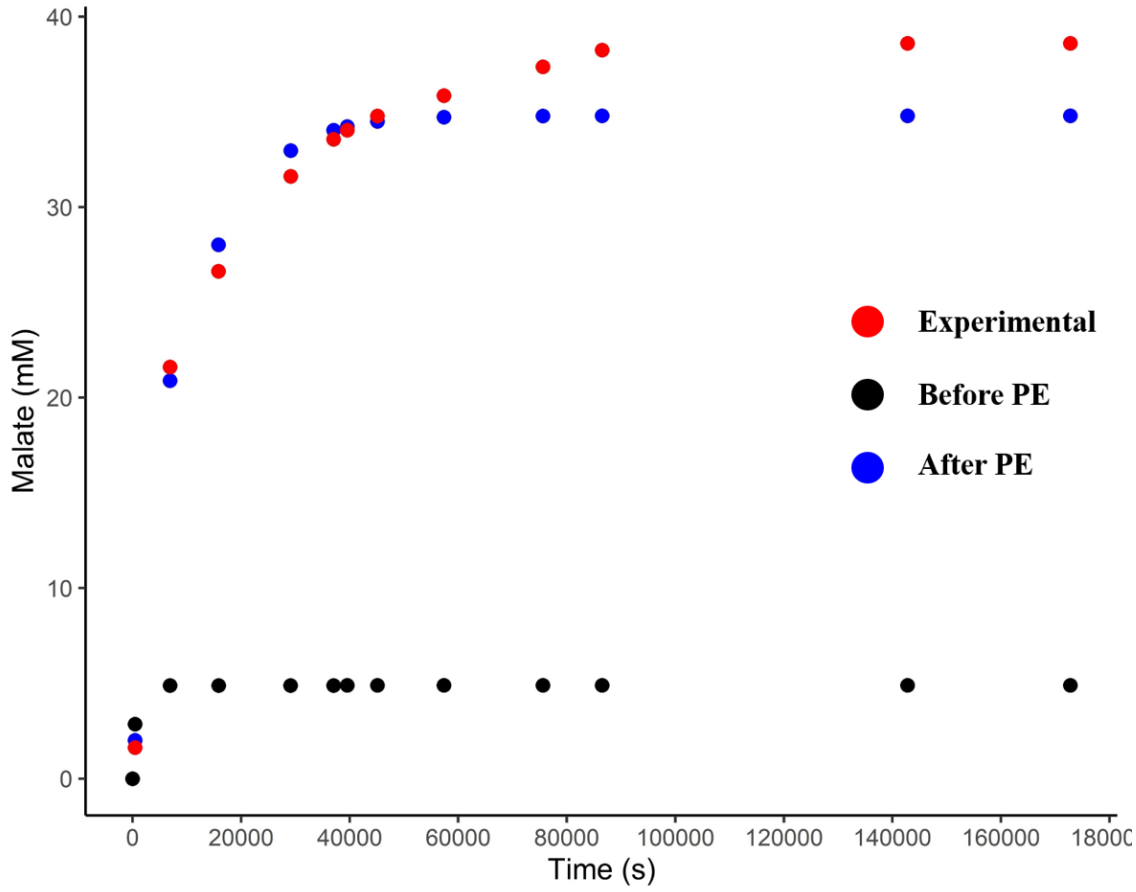


Figure 7.8: Comparision of malate concentration between experimental system and kinetic model before and after the parameter estimation. The malate concentration if isoamylase treated maltodextrin from Shi *et al.* was used for parameter estimation.

Table 7.5: The kinetic parameters used in the model before and after estimating the parameters (GAPDH, GAPN PGK parameters) using the IA-treated model. K_m: Michaelis-Menten constant in mM and k_{cat}: turnover number or catalytic constant in s^{-1}. The parameters were estimated using data from Shi *et al.* (T. Shi *et al.*, 2019) in COPASI with genetic algorithm.

Kinetic parameters used in the original model			**Kinetic parameters estimated from IA treated model**		
Reaction	**Kinetic parameter**	**Original value**	**Reaction**	**Kinetic parameter**	**New value**
re7	GAPDHKmG3P	0.3	**re7**	GAPDHKmG3P	1.54304
re7	GAPDHKmNAD	0.1	**re7**	GAPDHKmNAD	1
re7	GAPDHKmPho	8.3	**re7**	GAPDHKmPho	4.18037
re7	kcatGAPDH	47.67	**re7**	kcatGAPDH	377.954
re8	PGKKmADP	0.085	**re8**	PGKKmADP	0.011195
re8	PGKKmBPG	5.6	**re8**	PGKKmBPG	2.19344
re8	kcatPGK	96.63	**re8**	kcatPGK	32.1124
re9	GAPNKmG3P	0.02	**re9**	GAPNKmG3P	0.2
re9	GAPNKmNAD	3.1	**re9**	GAPNKmNAD	0.31
re9	kcatGAPN	1.73	**re9**	kcatGAPN	0.173

In the second step, to enhance the starch utilisation, 4-gluconotransferase (4-GT) and polyphosphate glucokinase (PPGK) were added to the optimised model-1 (referred to as model-2). The model-2 was simulated to observe the malate production with the addition of two enzymes. Figure 7.9 represents the malate synthesis through the pathway (model-2) which was approximately 45 mM, and the intermediates of the pathway fructose 1,6-bisphosphate (~4.5 mM) and 1,3-biphosphoglycerate (~2 mM). These intermediate concentrations indicate that not all the substrate is getting converted to final product in the model.

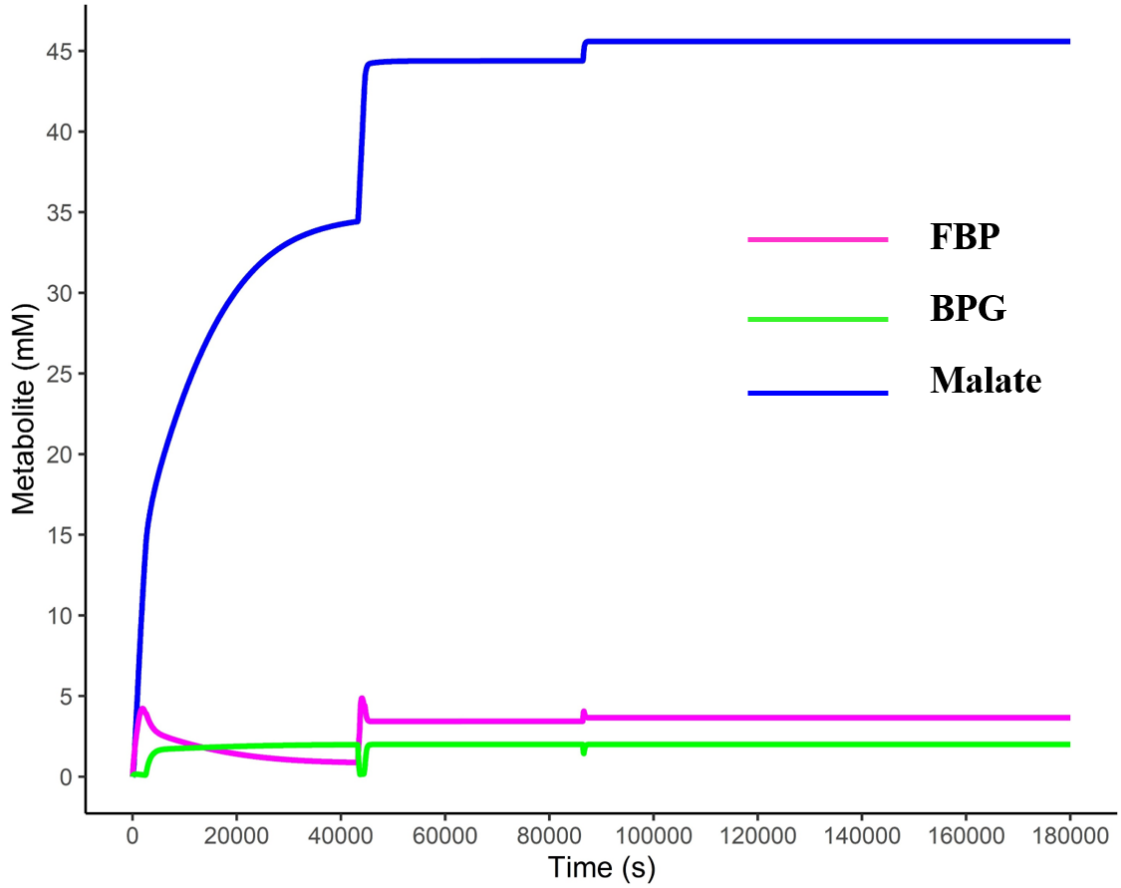


Figure 7.9: The metabolite produced by the new model with 4GT and PPGK. The model contains newly estimated parameter for GAPN, GAPDH and PGK.

In the third step, to enhance the malate production by utilising FBP and BGP, the kinetic model from the second step (model-2) was used. The model-2 parameters are then estimated using the data from Shi *et al.* for 4GT, PPGK added system (Annexe 21). The model-2 parameter was fit to experimental data for the enzymes PFK, GAPN, GAPDH, PGK and MDH. These five enzymes are selected because these reactions involve the cofactors ATP/ADP and NADH/NAD+. The hypotheses is that adjusting these enzyme parameters could help in the regeneration of the cofactors and therefore, would increase the final malate production. The parameter estimation from COPASI is used to fit the kinetic model to experimental data for time-course production of malate from Shi *et al.* (T. Shi *et al.*, 2019) using the genetic algorithm. The kinetic parameter PFKn was omitted from the estimation as it represents the number of allosteric site in the enzyme which is constant (n=4).

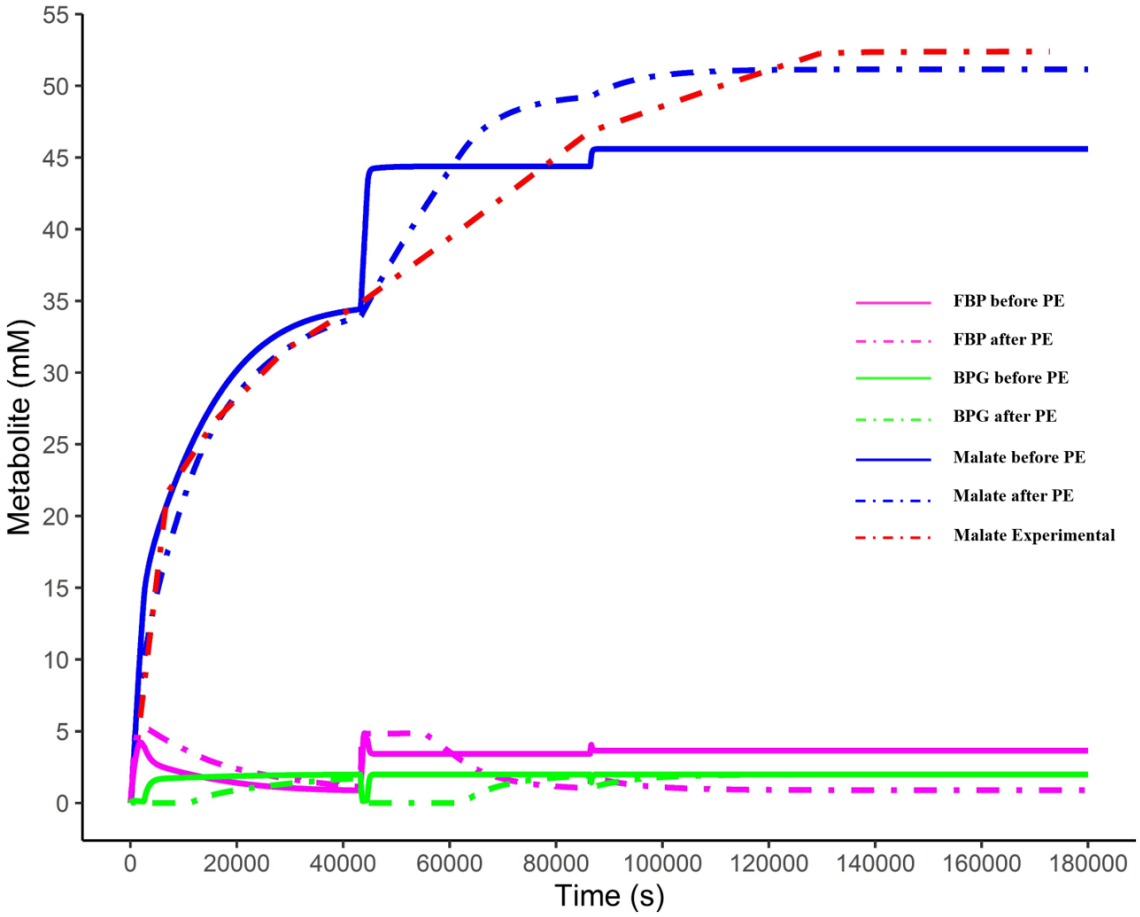


Figure 7.10: The concentration of metabolite produced before and after optimisation of the kinetic parameter. The kinetic parameters of enzyme PFK, GAPDH, GAPN, PGK and MDH are used fit with the experimental data.

From Figure 7.10 it is clear that, by regenerating the cofactors, higher malate concentration can be obtained. The enzyme PFK consumes ATP and converts F6P to FBP. In the system, the ATP concentration is low i.e. 2mM. This implies that ATP should be regenerated to obtain a higher malate concentration, according to the observation from Figure 7.6. PGK catalyses the ATP/ADP regeneration. The NAD+ in the pathway will be reduced to NADH by GAPN whereas the NADH is oxidised to NAD by MDH and hence leading to the regeneration of NADH/NAD+. FBP concentration decreased (Figure 7.11), which means that most of the intermediate FBP is completely utilised through effective regeneration of ATP and ADP by enzymes PGK and PFK respectively. And the NAD+/ NADH regeneration is effectively taking place by GAPN, GAPDH and MDH as observed in Figure 7.12. This proves that the enzymes selected i.e., PFK, GAPDH, GAPN, PGK and MDH, are an appropriate choice for optimising the system.

It should be noted that in microbial fermentation for malate synthesis, malate dehydrogenase (one of the five enzymes selected in model-2 for the optimisation) was found to be the rate-limiting step. And, as of now, other enzymes (PFK, GAPDH, GAPN and PGK) were not found to be rate-limiting in the context of malate synthesis. In the kinetic model, five enzymes (PFK, GAPDH, GAPN, PGK and MDH) were found to be important. This could be explained by the fact the model does not contain all the exact experimental values for some kinetic parameters of the enzymes, thus the experimental behaviour is not exactly observed during the simulation using the kinetic model.

New parameters obtained from optimisation of model-2 are given in Table 7.6. Except for the PFKKaATP, GAPDHkcat and PGKKmADP, all the parameters estimated were found to be in the range experimentally measured kinetic parameter from different sources recorded in the BRENDA database (Annexe 22).

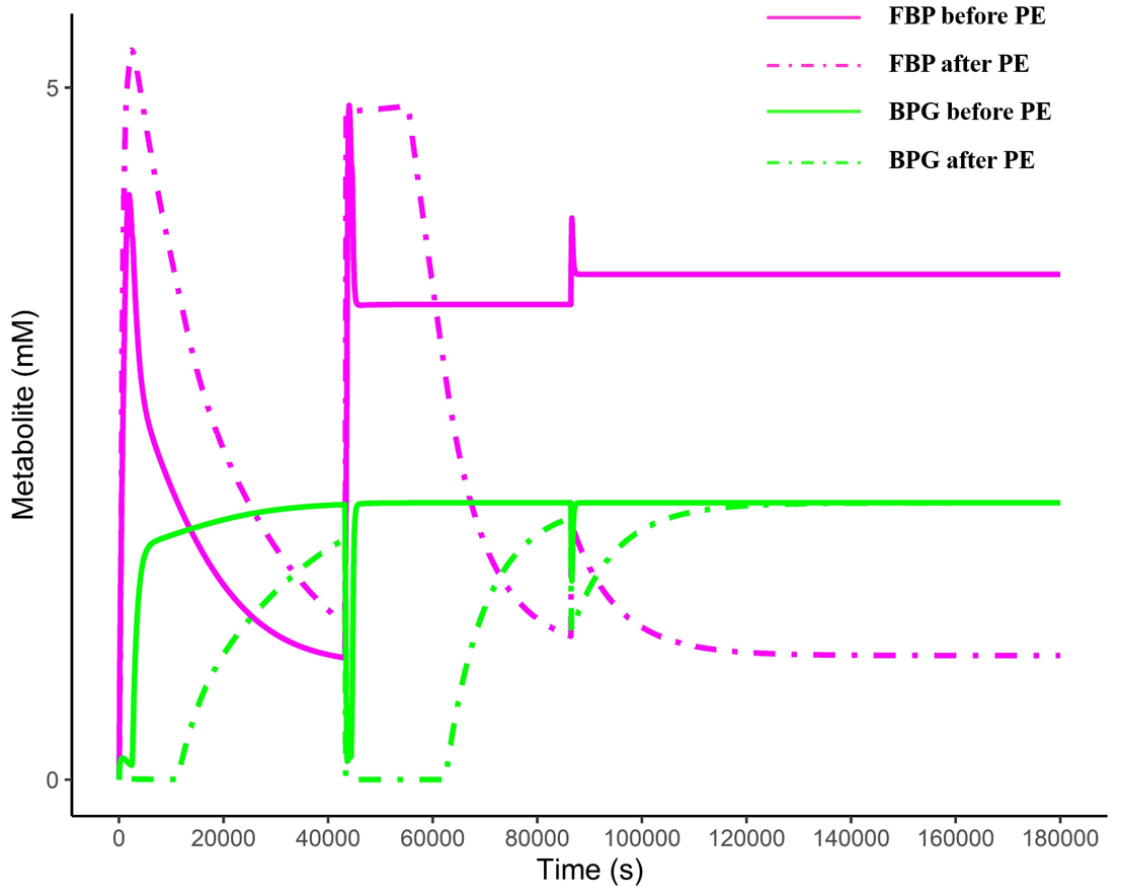


Figure 7.11: The concentration of fructose 1,6-bisphosphate and 1,3 biphosphoglycerate produced before and after optimisation of the kinetic parameter. The kinetic parameters of enzyme PFK, GAPDH, GAPN, PGK and MDH are used fit with the experimental data.

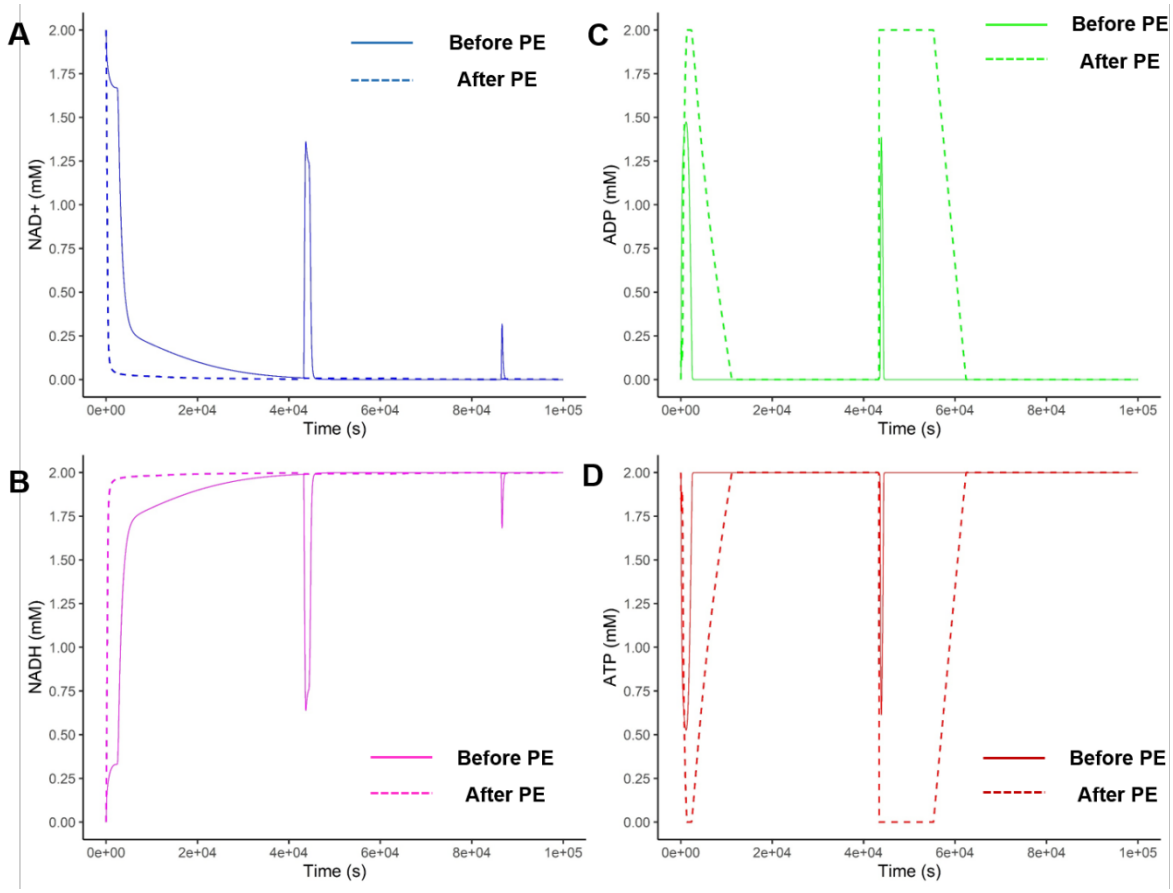


Figure 7.12: The concentrations of (A): NAD+ (B): NADH, (C): ADP (D): ATP before and after the estimation of PFK, GAPDH, GAPN, PGK, MDH parameter.

The above-optimised model (model-2) has an RMSE of 1.96 between the experimentally measured and model-simulated malate concentration. The final malate produced by the optimised model is 51.16 mM which is 97% of the experimentally measured malate concentration (52.4mM). And this is 93% of the theoretical maximum. This is a good validation for the kinetic model. Thus, model-2 can be used for further study. The 100% conversion of malate is not yet possible in the experimental system and some possible reasons could be, for example, i. the degradation of intermediates at high temperature such as 50-80 °C in which thermophilic enzymes are highly active, or ii. the NADH can be easily get decomposed at a higher temperature (Hofmann, Wirtz, Santiago-Schübel, Disko, & Pohl, 2010).

Table 7.6: The kinetic parameters used in the model before and after estimating the parameters (PFk, GAPDH, GAPN, PGK, MDH) using the IA-treated model with 4GT-PPGK. K_m: Michaelis-Menten constant in mM and k_{cat}: turnover number or catalytic constant in s^{-1}. The parameters were estimated using data from Shi *et al.* (T. Shi *et al.*, 2019) in COPASI with genetic algorithm.

Kinetic parameter used before the optimisation			**Kinetic parameters estimated for IA treated + 4GT, PPGK added model**		
Reaction	**Kinetic parameter**	**Original value**	**Parameter**	**Kinetic parameter**	**New value**
re4	PFKKaATP	0.006	**re4**	PFKKaATP	0.000672
re4	PFKKaF6P	0.027	**re4**	PFKKaF6P	0.038633
re4	PFKKiPEP	0.00158	**re4**	PFKKiPEP	0.007513
re4	kcatPFK	2.5	**re4**	kcatPFK	0.431056
re7	GAPDHKmG3P	1.54304	**re7**	GAPDHKmG3P	3.77858
re7	GAPDHKmNAD	1	**re7**	GAPDHKmNAD	0.995012
re7	GAPDHKmPho	4.18037	**re7**	GAPDHKmPho	0.418504
re7	kcatGAPDH	377.954	**re7**	kcatGAPDH	2149.92
re8	PGKKmADP	0.011195	**re8**	PGKKmADP	0.00551
re8	PGKKmBPG	2.19344	**re8**	PGKKmBPG	0.219
re8	kcatPGK	32.1124	**re8**	kcatPGK	23.2651
re9	GAPNKmG3P	0.2	**re9**	GAPNKmG3P	0.958756
re9	GAPNKmNAD	0.31	**re9**	GAPNKmNAD	0.033685
re9	kcatGAPN	0.173	**re9**	kcatGAPN	0.72704
re13	MDHKmMal	0.095	**re13**	MDHKmMal	0.0095
re13	MDHKmNAD	0.14	**re13**	MDHKmNAD	1.27271
re13	MDHKmNADH	0.024	**re13**	MDHKmNADH	0.248088
re13	MDHKmOAA	0.043	**re13**	MDHKmOAA	0.164212
re13	kcatMDH	35	**re13**	kcatMDH	32.8082

Here, estimating the values of the kinetic parameters for a given enzyme guide us for the selection of the more appropriate enzymes to add in the system. Indeed, one could choose for example phosphoglycerate kinase from *Chlamydomonas reinhardtii* which has a k_{cat} (28.5 s^{-1}) value close to the optimised parameter (23.26 s^{-1}), or malate dehydrogenase from *Methylomicrobium alcaliphilum* which has a k_{cat}/K_m (2558 $1/mMs^{-1}$) close to the estimated parameters (3452.63 $1/mMs^{-1}$). Nevertheless, there is no guarantee that an enzyme that fits exactly the requirements in term of kinetic parameters can be found, and that the enzyme is thermostable. So, obviously, such an approach can be considered as an interesting step in the development of a process to build a metabolic pathway which will produce a maximum concentration of malate, but not as a final goal. The other interest of such an approach is to provide a guide in terms of enzyme improvement through protein engineering: using a thermostable enzyme, to introduce mutation(s) to modify kinetic parameters so as it fits with the estimated values could be tried. So, introducing enzyme engineering to improve the kinetic

parameters, or substrate channelling between two or more successive enzymes in the cascade reactions to fasten the conversion of intermediates, could be an interesting area of research.

Nevertheless, the optimised model-2 can be already used to optimise the malate synthesis from other substrates such as cellulose, cellobiose *etc.* Currently, poly-malic acid synthesis is studied from the lignocellulose. Synthesising malate from the low-cost substrate using thermophilic enzymes could be a promising field of research for malate synthesis.

7.5 Conclusion

The kinetic model of biosynthesis of malic acid was built in this study based on Shi *et al.* designed an ATP-balanced pathway using hyperthermophilic enzymes. Since the kinetic parameters were not available for all the enzymes used in the experimental study, homologous enzymes were identified for which kinetic parameters were available in the BRENDA database. The kinetic parameters from homologous enzymes are used and classical Michaelis-Menten kinetic equations are utilised in the model. To improve the model accuracy, the kinetic parameters were estimated using the experimental data. The parameter estimation was performed in two steps using malate produced by i. isoamylase treated maltodextrin and ii. isoamylase treated malate with 4GT and PPGK being added. This two-step estimation approach helps to identify the key enzyme in the model to optimise. The important enzymes in the higher malate production by this kinetic model was found to be PFK, GAPN, GAPDH, PGK and MDH. With newly estimated parameters the model could produce malate up to 51.16 mM which was 93% of the maximum theoretical yield. This is 97% of the experimentally obtained malate which validates the model to study the regulators of the pathway and to examine other substrates.

Conclusion and Prospectives

In the past decades, Cell-free systems (CFS) have emerged as a powerful technology to engineer biological systems for the synthetic biology application without using living cells. Compared to cell-based systems, CFS provides a simpler and faster solution with a high degree of freedom to control and to manipulate the process. Cell-free systems are used both in the laboratory, to develop and optimise new pathways and in the industry, for biomanufacturing. CFS offers a way to overcome the drawbacks of cell-based systems like unnecessary side reactions, the toxicity of substrates, intermediates or products, *etc*. The CFS allows not only the use of enzymes from different organisms, but also to use different metabolisms across different phyla. This thesis explores the data-based (machine learning) and knowledge-based (kinetic modelling) modelling approaches to study cell-free systems in two aspects.

First, for selecting the optimum enzyme concentrations: One of the main challenges for CFS is the selection of appropriate enzymes. The choice of low performing enzyme can compromise the final product yield. This issue can be solved by identifying homologous enzymes which perform better. After the identification of the suitable enzyme, the selection of optimum enzyme concentration is crucial. The experimental selection of optimal enzymes for the higher yield is time-consuming and expensive.

Therefore, in Chapter 3 an artificial neural network (ANN) model was developed using part of classical glycolysis pathway. The ANN model predicts flux through the pathway. Different algorithms, architectures and activation functions were examined in the study. Eventually, the model built using algorithm neuralnet and logistic activation function was retained.

The ANN has been already used in fermentation for predicting the culture conditions. In Chapter 3, the artificial neural network was used for the prediction of flux. However, the ANN method is a training based method which means the prediction depends on the data used for the modeling. ANN is known to be inefficient in extrapolating the prediction beyond the training data range used in the model. Indeed, when the model receives new information outside the trained data, the new prediction within the range of training data. Mostly, possible higher values are predicted at the glass ceiling (GC), ie. at a limit maximal value.

In chapter 4, a new ANN-based methodology (GC-ANN for Glass ceiling ANN) was developed to select the optimum enzyme balances (the combination of different enzymes concentration) of higher flux. The methodology uses different statistical analysis methods such

as principal component analysis, data classification along with the neural network. The new methodology aids *in silico* selection of optimum enzyme concentrations for maximum flux through the pathway. The approach was validated in simulation using the kinetic model of the pathway, and the experimentally. This methodology also provides an advantage of selecting the cost-efficient optimum balances. It was expected to obtain slight improvements, i.e., improved flux values close to the highest one that fed into the model. Surprisingly, improvements up to 63% were obtained. Moreover, these improvements are coupled with a cost decrease of up to 25% for the assay. The GC-ANN methodology was implemented for the upper part of glycolysis as a first example. It would be interesting to apply this methodology for different pathways in future.

The ANN model for chapter 3 and chapter 4 was built using the experimental measurement of flux through the upper part of glycolysis. The data consists of 121 enzyme balances with the flux value ranging from 0.79 µM/s to 12.9 µM/s. It is difficult to accurately predict the flux beyond this range of flux. And, increasing the dataset of different enzyme balances, with diverse flux, by performing experiments is intensive and high-budget. Hence, the potential of kinetic modelling of the pathway was explored to increase the size of learning dataset.

The building of the kinetic model requires kinetic information of enzymes. The kinetic parameters measured in chapter 4 helped to build the model. The model built was inefficient in replicating experimental flux (RMSE: 5.14). Thus, the optimisation for kinetic parameters of the enzymes was performed in two steps: i. iterative approach: where individual enzymes parameter were fitted with experimental data, ii. Selective approach: the mean of the less deviated parameters from step (i) was updated to the model, and highly deviated parameters were identified. The hypothesis was that the parameters which deviate less are more crucial to the model than the once which deviate a lot. After the two steps of parameter estimation, the final model had an RMSE of 1.91, which was a significant improvement in the model efficiency. And the final kinetic model obtained can be used to check if the balances found in the glass-ceiling of the ANN are potential high flux or not. Yet, there is room to improve the model to replicate the experimental conditions.

Second, for modelling carbon fixation, the constant increase of carbon dioxide (CO_2) concentration in the atmosphere is one of the major threats to life on earth. Finding ways to reduce global CO_2 is not only fascinating but also an urgent requirement. The fixation of CO_2 is observed naturally in plants, bacteria and cyanobacteria. The CO_2 fixation, by biosynthesis,

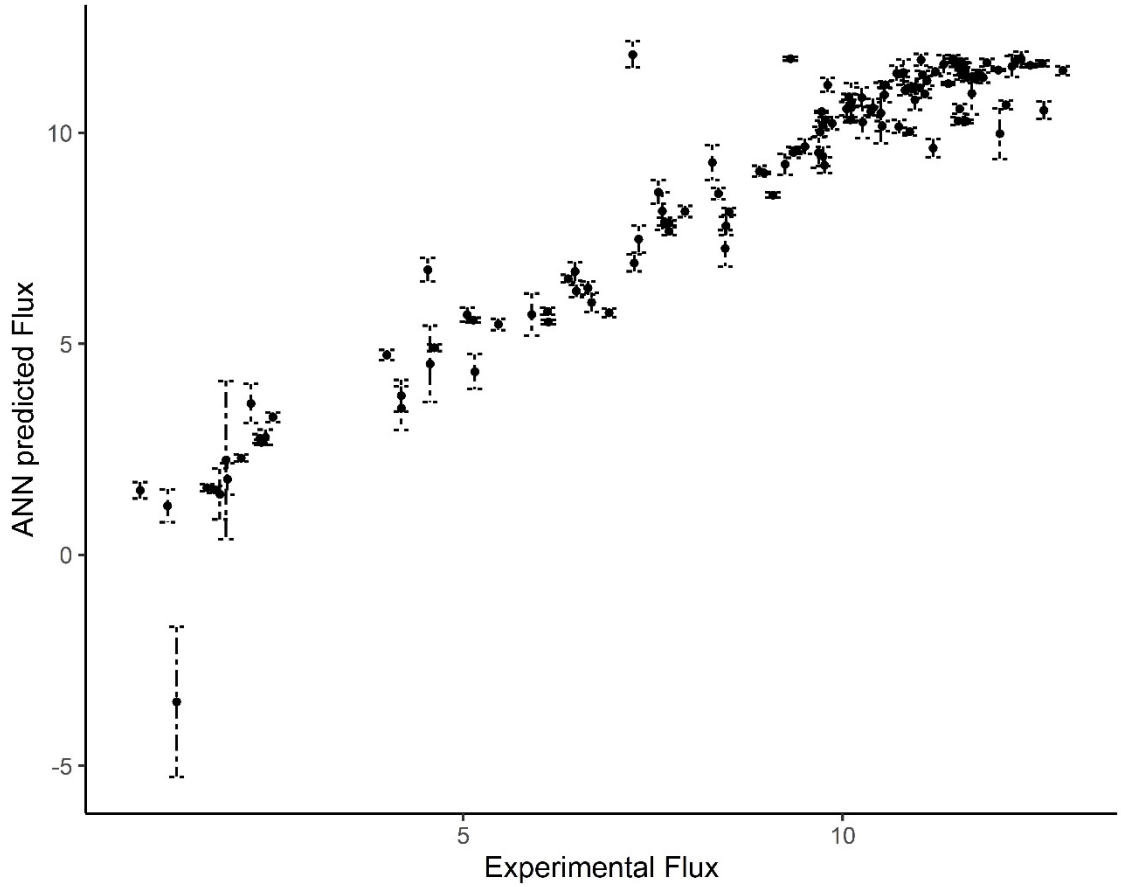
into demanding biochemicals is a good choice. The malic acid is a four-carbon dicarboxylic acid, part of the citric acid cycle. Malic acid is extensively used in the food and beverage industries as an acidulant. Currently, malic acid is synthesised by chemical method from petroleum-based. In chapter 6, the state of art in malic acid synthesis is discussed. Many studies have been performed for malate synthesis using microorganisms. However, microbial synthesis of malate is not yet reached the biomanufacturing stage. This is mostly due to expensive substrates and low yield of malate by microbial fermentation. Recently malate has been synthesised using a cell-free system of thermophilic enzymes in artificial ATP-balanced pathway. The ATP-balanced pathway produced 95% of the theoretical yield from maltodextrin in 48 hours (55 mM).

In Chapter 7, the ATP-balanced pathway was studied using the kinetic modelling approach. The model was built using the kinetic parameters from homologous enzymes, whenever it was necessary. The model utilises the isoamylase treated maltodextrin as a substrate. In *in vitro* experiments of ATP-balanced pathway, no rate-limiting step was found. Using the kinetic model, it was discovered that enzymes involved in cofactor regeneration, i.e. PFK, GAPDH, GAPN, PGK and MDH, were important for higher malate production. The enzymes kinetic parameters of these five enzymes were estimated to fit the model with experimental data. The enzyme MDH was identified as the rate-limiting step in microbial biosynthesis. However, other enzymes are not identified as potential regulators for malate synthesis in microbial fermentation and it will be interesting to study these enzymes in the cell-free system to check if these biocatalysts could be potential regulators for the pathway. The optimised kinetic model obtained, can produce 97% (51.16mM malate) of experimentally produced malate and 93% of the theoretical maximum (55 mM). This is an indication that our kinetic model is good.

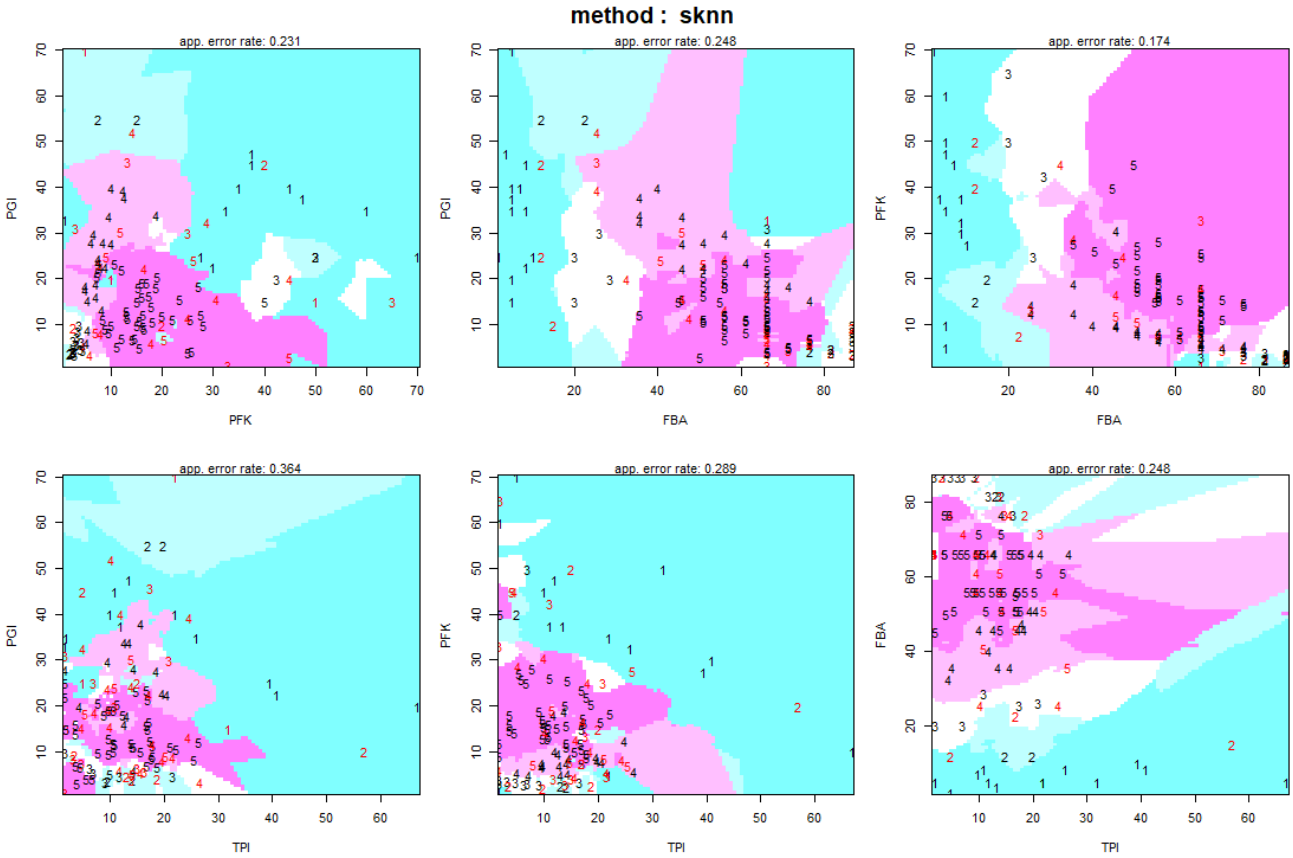
In future, the modelling of cell-free systems can be used to optimise the malate synthesis using low-cost raw materials such as lignocellulose. The already developed kinetic model can be used as a plugin to examine the potential of other substrates. This modelled pathway can be used as a guide for enzyme improvement through protein engineering to improve the efficiency of key enzymes. The substrate channelling between two or more successive enzymes in the cascade reactions could be tested to fasten the conversion of intermediates.

ANNEXE

Annexe 1: The variation of flux predicted by ANN model with different order of input data during the training phase of modelling.



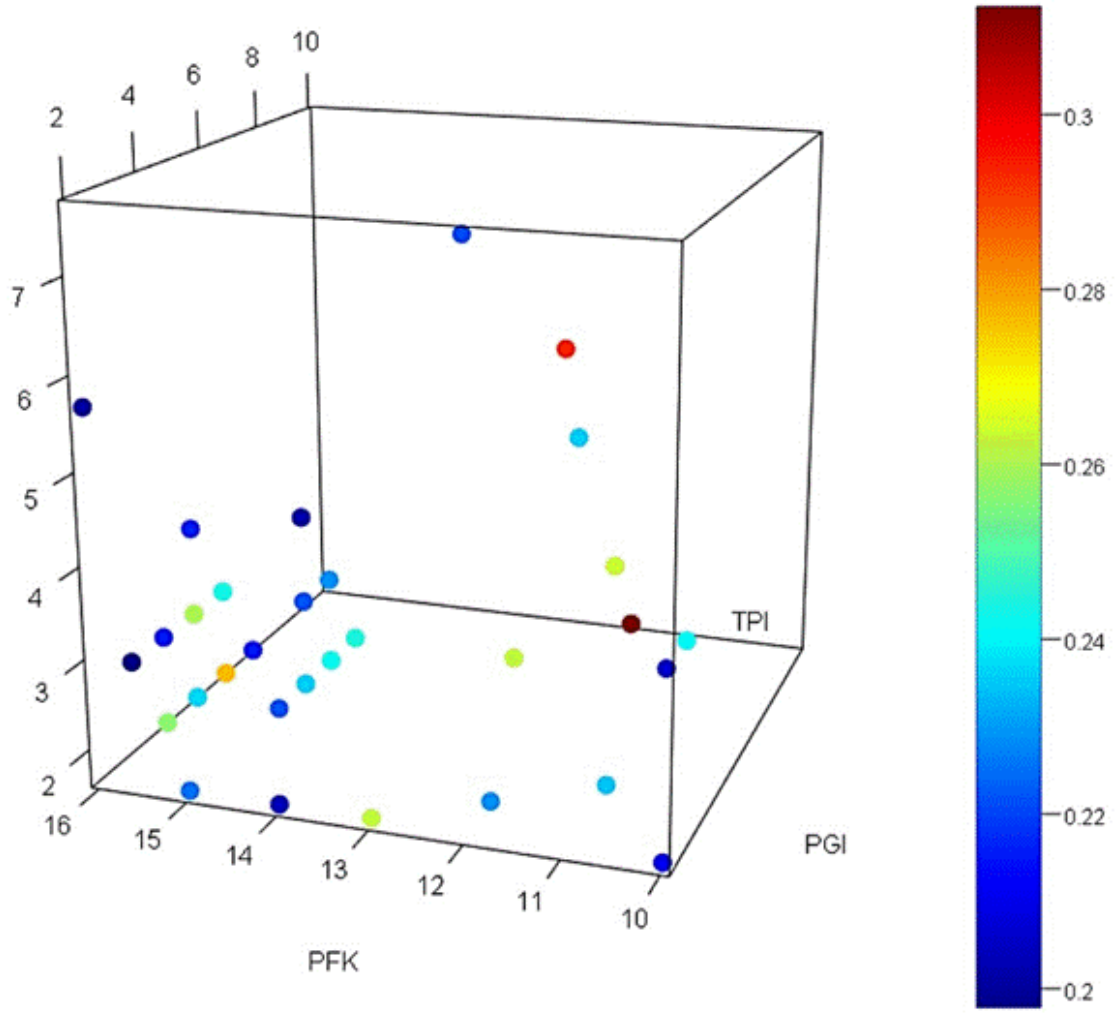
Annexe 2: Discriminant analysis for the classification of data from Fiévet *et al.* (Fiévet *et al.*, 2006) using the rpart (Therneau & Atkinson, 2018) method from R. Colour code according to the feature space of data, where group 1 (flux: 0.728-3.17 µM/s) is shown in light cyan, group 2 (flux: 3.17-5.6 µM/s) darker cyan, group 3 (flux: 5.6-8.04 µM/s), group 4 (8.04-10.5 µM/s), group 5 (10.5-12.9 µM/s).



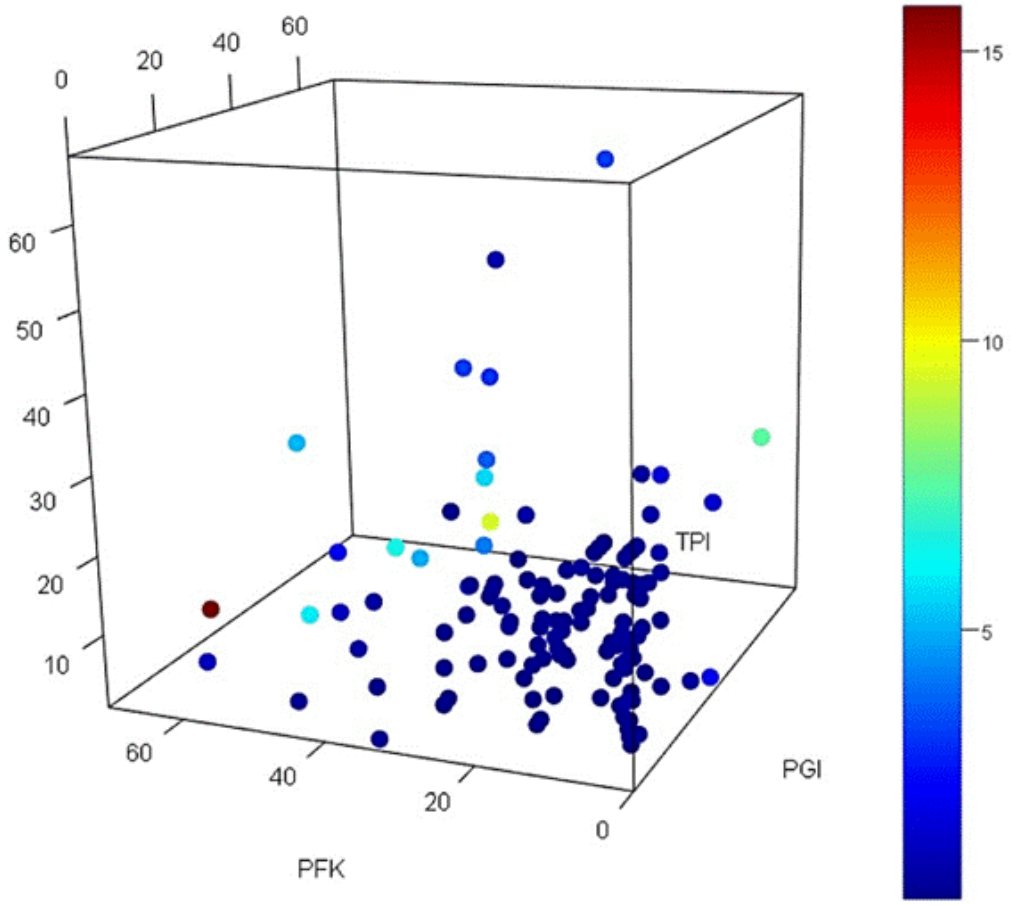
Annexe 3: The calculated price for the µM of NADH consumed per second by the enzyme concentration selected for the experiment.

Index	mg/l				µM/s		in EUR
	PGI	PFK	FBA	TPI	J_{ANN}	J_{EXP}	Price per µM
11	2	10	88.24	1.66	12.24	15.7	0.213
12	2	10	86.24	3.66	12.06	16.3	0.208
13	2	11	82.24	6.66	12	12.1	0.294
14	2	12	80.24	7.66	12.03	16.6	0.222
15	2	13	85.24	1.66	12.7	13.9	0.263
16	2	14	84.24	1.66	12.74	18.3	0.205
17	2	15	83.24	1.66	12.72	17.1	0.226
18	2	16	78.24	5.66	12.16	20.1	0.202
19	3	10	85.24	3.66	12	14.4	0.241
20	3	12	85.24	1.66	12.53	15.8	0.230
21	3	16	80.24	2.66	12.44	20.6	0.198
22	4	11	85.24	1.66	12.32	15.4	0.235
23	4	16	80.24	1.66	12.49	16.1	0.257
24	4	16	79.24	2.66	12.36	19.3	0.216
25	5	15	80.24	1.66	12.48	18.5	0.223
26	5	16	79.24	1.66	12.41	17.8	0.237
27	5	16	78.24	2.66	12.29	16.3	0.261
28	5	16	77.24	3.66	12.18	19.7	0.217
29	6	15	79.24	1.66	12.41	17.8	0.237
30	6	15	78.24	2.66	12.29	19	0.223
31	6	15	77.24	3.66	12.19	21	0.203
32	6	16	78.24	1.66	12.34	15.6	0.277
33	6	16	77.24	2.66	12.23	17.8	0.244
34	7	12	78.24	4.66	12	17.1	0.237
35	7	15	78.24	1.66	12.33	17.7	0.243
36	7	15	77.24	2.66	12.22	18.8	0.230
37	7	16	77.24	1.66	12.27	20.4	0.216
38	8	13	79.24	1.66	12.26	15.9	0.263
39	8	15	77.24	1.66	12.26	17.9	0.245
40	9	12	78.24	2.66	12.04	15.8	0.265
41	10	12	78.24	1.66	12.05	13.6	0.312

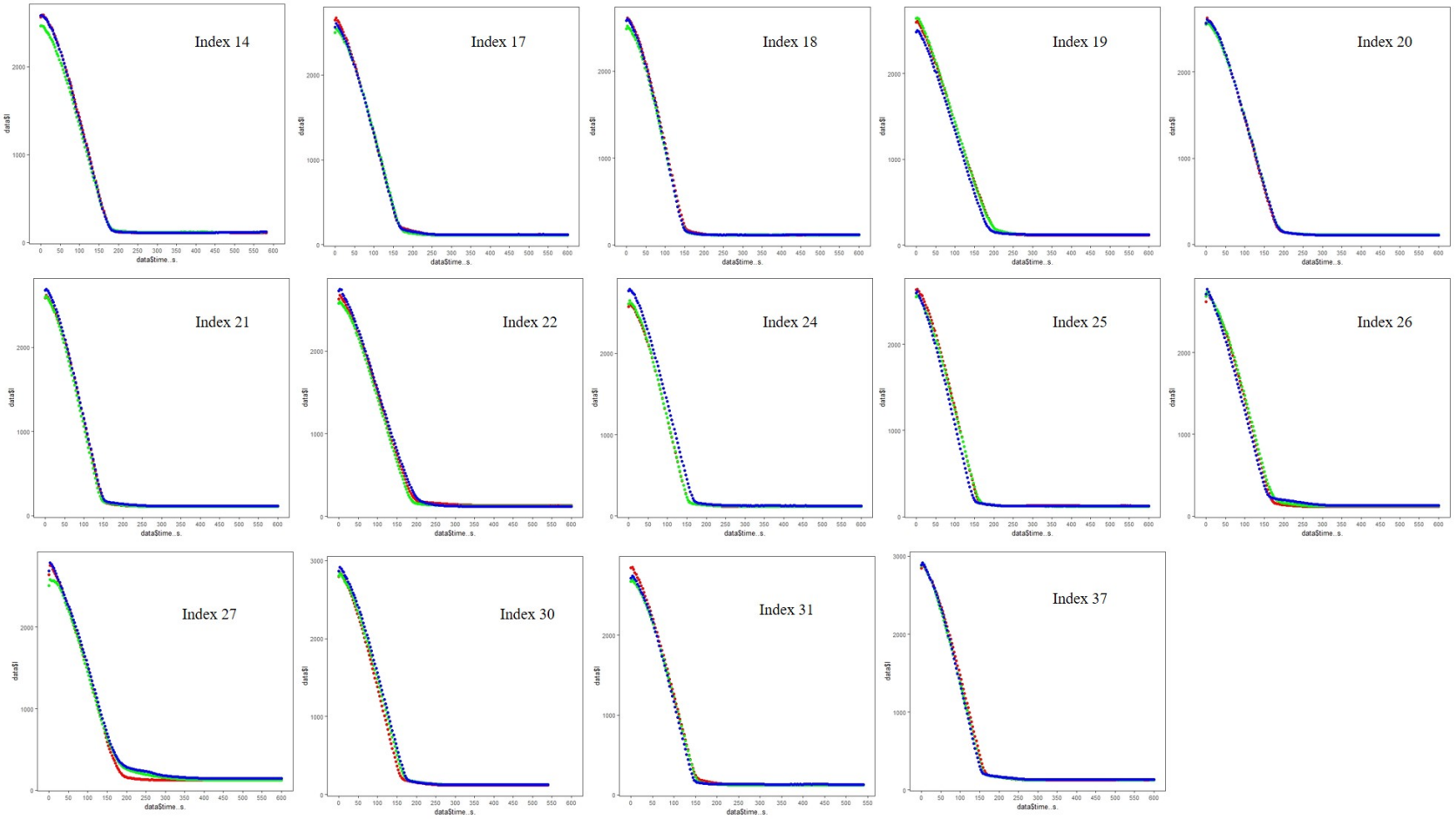
Annexe 4: The cost predicted (in EUR) for the four-enzyme concentration (PGI, PFK, FBA and TPI) selected for experimental validation. The blue is lowest, to highest in red.



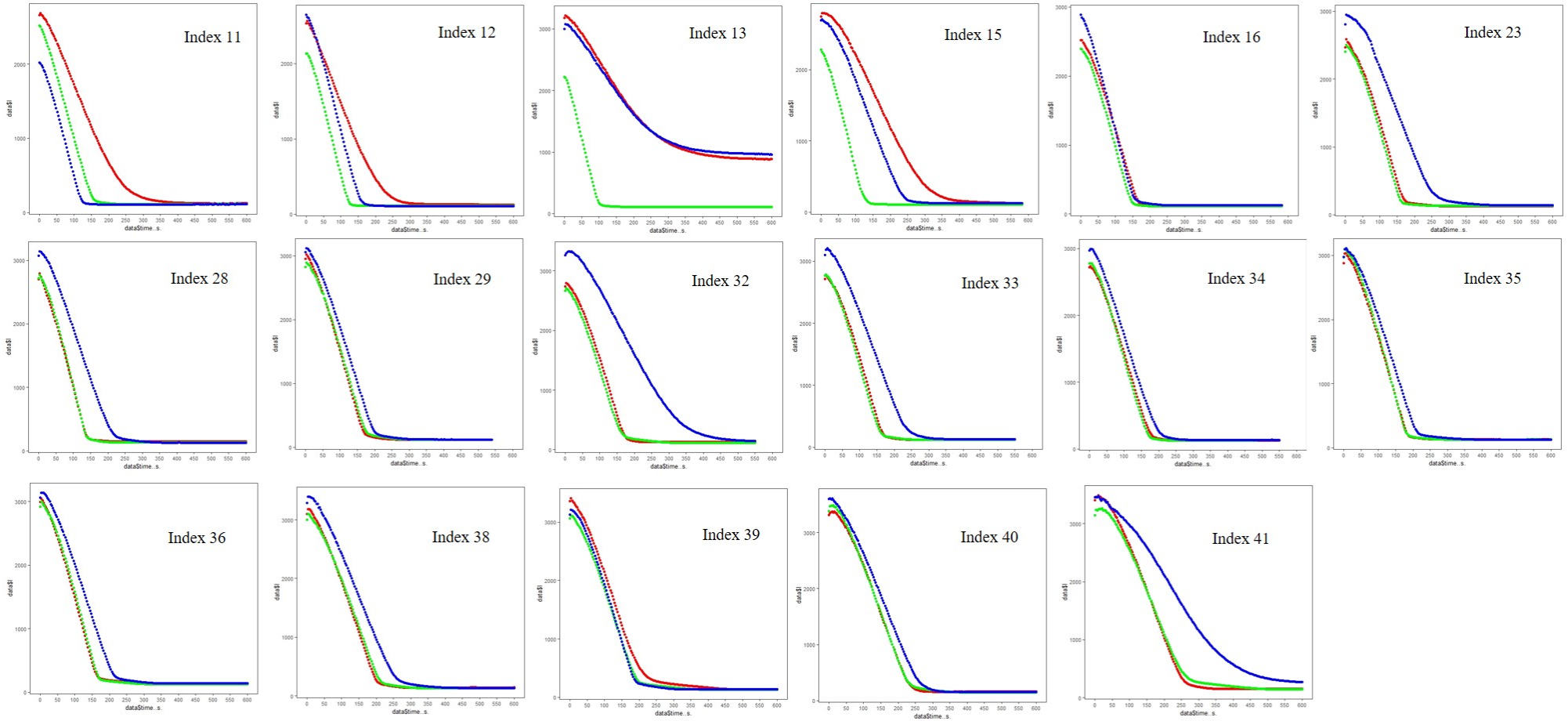
Annexe 5: The cost predicted (in EUR) for the four-enzyme concentration (PGI, PFK, FBA and TPI) selected by Fiévet *et al.* (2006). The blue is lowest, to highest in red.



Annexe 6: The Group 1 data used for the estimation of the kinetic parameters. The dots represent the experimental measurement of NADH consumption by G3PDH through the upper part of glycolysis. The colours of the dot represent different replicates as follows: red: replicate 1; blue: replicate 2, green: replicate 3. The x-axis: time in seconds, y-axis: Concentration of NADH in µM.



Annexe 7: The Group 2 data used for the estimation of the kinetic parameters. The dots represent the experimental measurement of NADH consumption by G3PDH through the upper part of glycolysis. The colours of the dot represent different replicates as follows: red: replicate 1; blue: replicate 2, green: replicate 3. The x-axis: time in seconds, y-axis: Concentration of NADH in µM.



Annexe 8: RMSE variation between experimental and model fitted concentration of NADH across different cycles of parameter estimation.

	Cycle1					Cycle12					Cycle13				
	L1	**L2**	**L3**	**L4**	**L5**	**L1**	**L2**	**L3**	**L4**	**L5**	**L1**	**L2**	**L3**	**L4**	**L5**
index14	13.640	11.575	11.575	11.675	11.690	11.380	11.387	11.243	11.419	11.419	11.251	11.167	11.130	11.241	10.995
index17	14.043	11.844	11.441	11.405	11.404	10.934	10.920	10.883	10.880	10.880	10.880	10.873	10.870	10.870	10.867
index18	13.309	12.157	12.135	11.961	11.989	11.812	11.705	11.705	11.753	11.753	11.441	11.324	11.302	11.330	11.326
index19	20.961	18.514	18.442	18.380	18.264	17.512	17.325	17.287	17.435	17.451	17.215	17.215	17.255	17.373	17.373

Annexe 9: RMSE variation between experimental and model fitted concentration of NADH across the different cycle of parameter estimation after updating PGI, PFK, FBA, TPI, And GDH iteratively. The order of enzyme selected at step n depends on the RMSE between new prediction and experimental concentration at step n-1.

	Cycle 1	**Cycle 2**	**Cycle 3**	**Cycle 4**	**Cycle 5**
index14	11.68973	11.41933	10.99482	10.99482	10.6495
index17	11.40421	10.87999	10.86682	10.85298	10.84276
index18	11.98877	11.75305	11.32596	10.95669	10.18399
index19	18.26374	17.45077	17.37294	17.30587	17.21132

Annexe 10: Kinetic parameters obtained after five cycles of iterative parameter estimation for the indexes which yield an RMSE of less than 2 and positive Kendall Tau between experimental and model-simulated flux. kcat: turnover number in s^{-1}; Km: Michaelis-Menten constant in mM; Ki: inhibition constant in mM; Keq: Equilibrium constant.

Parameter	Index29	Index30	Index37	Index39	parameter	Index29	Index30	Index37	Index39
KeqPGI	0.615	1	0.581132	1.000	**kcatPfk**	29.390	27.78415	28.55318	209.998
kcatFbaR	0.000	0	0.007345	0.046	**KmPGIG6P**	84.000	84	84	400.782
KeqTPI	0.209	0.021268	0.049036	0.029	**KmFbaFBP**	3.904	3.977679	20.32414	5000.000
KmG3dhNADH	4.300	4.300936	4.3	21.992	**kcatPgiR**	262.649	543.5177	290.158	247.200
kcatFbaF	1.512	1.541179	3.244096	5.449	**kcatPgiF**	1410.000	1398.289	1397.283	1410.000
KiFbaG3P	9960.214	10000	10000	65.784	**KmFbaDHAP**	2375.261	1238.556	2036.585	1745.441
kcatTpiF	5640.078	16700	4891.016	15471.033	**KmFbaG3P**	2173.774	2500	716.7147	2500.000
kcatG3dh	309.000	256.5478	307.8306	195.335	**KmTpiDHAP**	1230.219	1521.178	1241.007	2240.978
KmPGIF6P	268.106	306.5059	307	307.000	**KmTpiG3P**	2433.737	1370.018	722.3535	2256.048
KmG3dhDHAP	77.176	119.7806	96.13302	75.722	**kcatTpiR**	623.360	942.4913	898.5548	999.936
KmPfkATP	120.722	76.38901	71.03611	72.550	**KmG3dhGly3P**	318.890	768.7406	409.0252	909.000
KmPfkf6p	205.619	143.6157	190.1068	193.792	**KmG3dhNAD**	380.000	380	380	67.510

Annexe 11: The percentage deviation between the kinetic parameters used in the initial model and the parameter estimated after 5 cycles of iterative parameter estimation across 31 experiments. Original value: the kinetic values used in the original model, Average Value: The average of kinetic parameters estimated during iterative PE and Deviation: is the percentage deviation between the original value and average.kcat: turnover number in s^{-1}; Km: Michaelis-Menten constant in mM; Ki: inhibition constant in mM; Keq: Equilibrium constant.

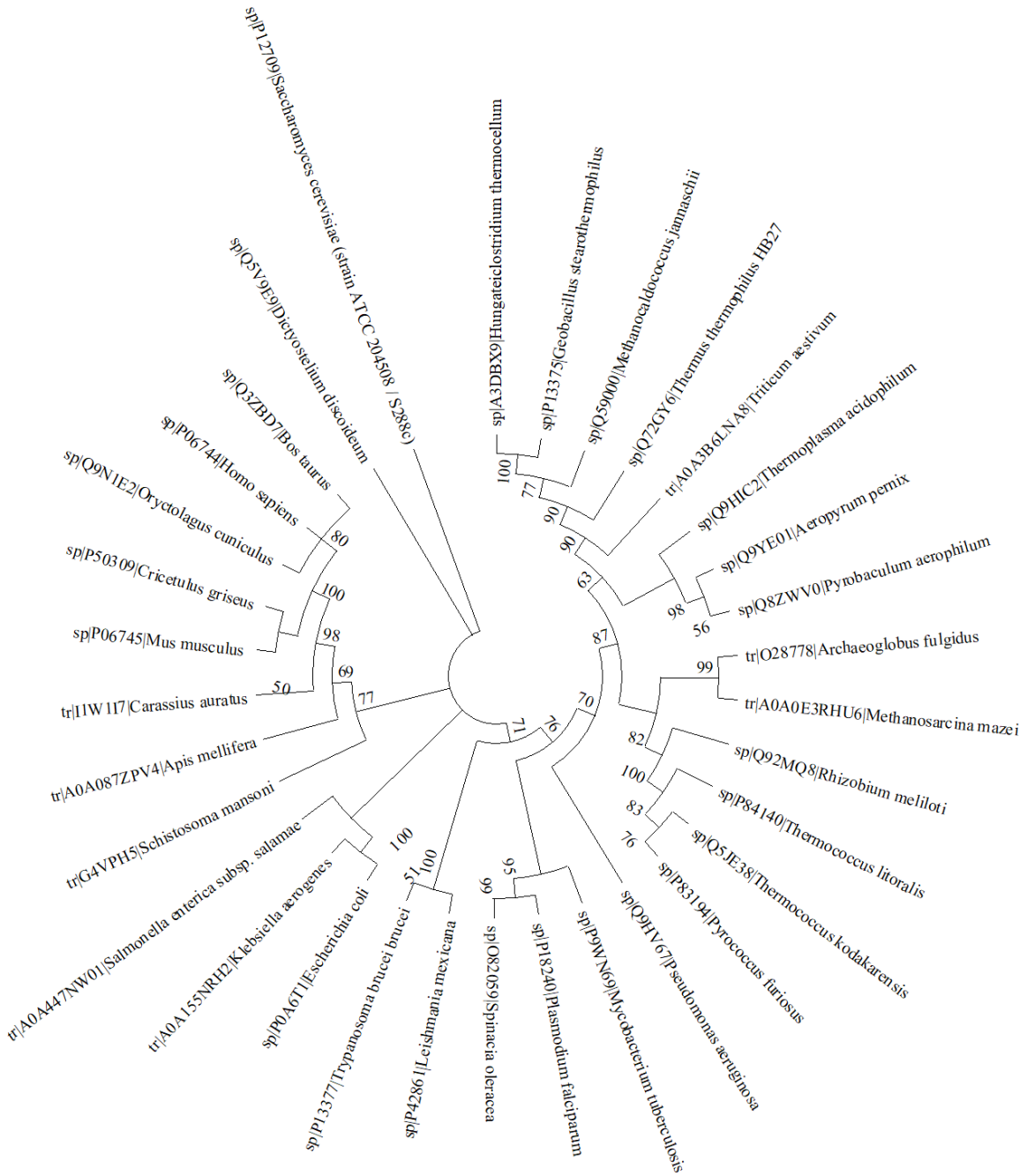
Parameter	Original value	Average value	Deviation %	Parameter	Original value	Average value	Deviation %
kcatG3dh	189.1	287.3792	151.9721	**KiFbaG3P**	10000	6387.225	63.87225
KeqTPI	0.045	0.105636	234.7466	**KmFbaDHAP**	2400	1791.64	74.65168
KmG3dhNAD	83	272.9613	328.869	**kcatPgiF**	1107.37	860.1729	77.6771
KmFbaFBP	140	603.0298	430.7355	**KmTpiG3P**	1270	1044.958	82.28017
KmPfkATP	120	107.7768	89.81401	**KmFbaG3P**	2000	1667.025	83.35125
kcatPfk	166.075	161.9523	97.51757	**kcatTpiF**	8486.67	10394.19	122.4767
kcatTpiR	816.67	839.7529	102.8265	**KmTpiDHAP**	1230	1529.05	124.313
KmPfkf6p	130	143.1877	110.1444				

Annexe 12: The comparison of RMSE, Kendall tau and Spearman coefficients between experimental and model-simulated flux for the model after five cycles of estimation selective parameters.

Index	**RMSE**	**Kendall Tau**	**Spearman Coefficient**	**Index**	**RMSE**	**Kendall Tau**	**Spearman Coefficient**
Index11	3.852	-0.249	-0.354	Index27	2.500	0.331	0.472
Index12	3.816	-0.262	-0.373	Index28	4.458	0.370	0.518
Index13	7.490	0.340	0.483	Index29	2.153	0.249	0.346
Index14	5.517	-0.054	-0.095	Index30	2.131	0.219	0.309
Index15	6.206	0.353	0.495	Index31	3.022	0.219	0.400
Index16	2.533	-0.245	-0.349	Index32	2.042	0.318	0.439
Index17	4.366	-0.058	-0.349	Index33	2.150	0.314	0.435
Index18	5.988	-0.054	-0.095	Index34	2.062	0.374	0.508
Index19	3.351	-0.024	-0.039	Index35	4.407	0.331	0.470
Index20	2.174	0.024	0.018	Index36	3.233	0.348	0.492
Index21	3.146	-0.006	-0.015	Index37	2.544	0.201	0.295
Index22	2.350	0.058	0.087	Index38	3.956	0.318	0.462
Index23	2.172	-0.158	-0.233	Index39	3.196	0.180	0.266
Index24	2.413	0.297	0.415	Index40	8.183	0.227	0.328
Index25	2.084	0.236	0.343	Index41	13.623	0.197	0.286
Index26	1.912	0.292	0.403				

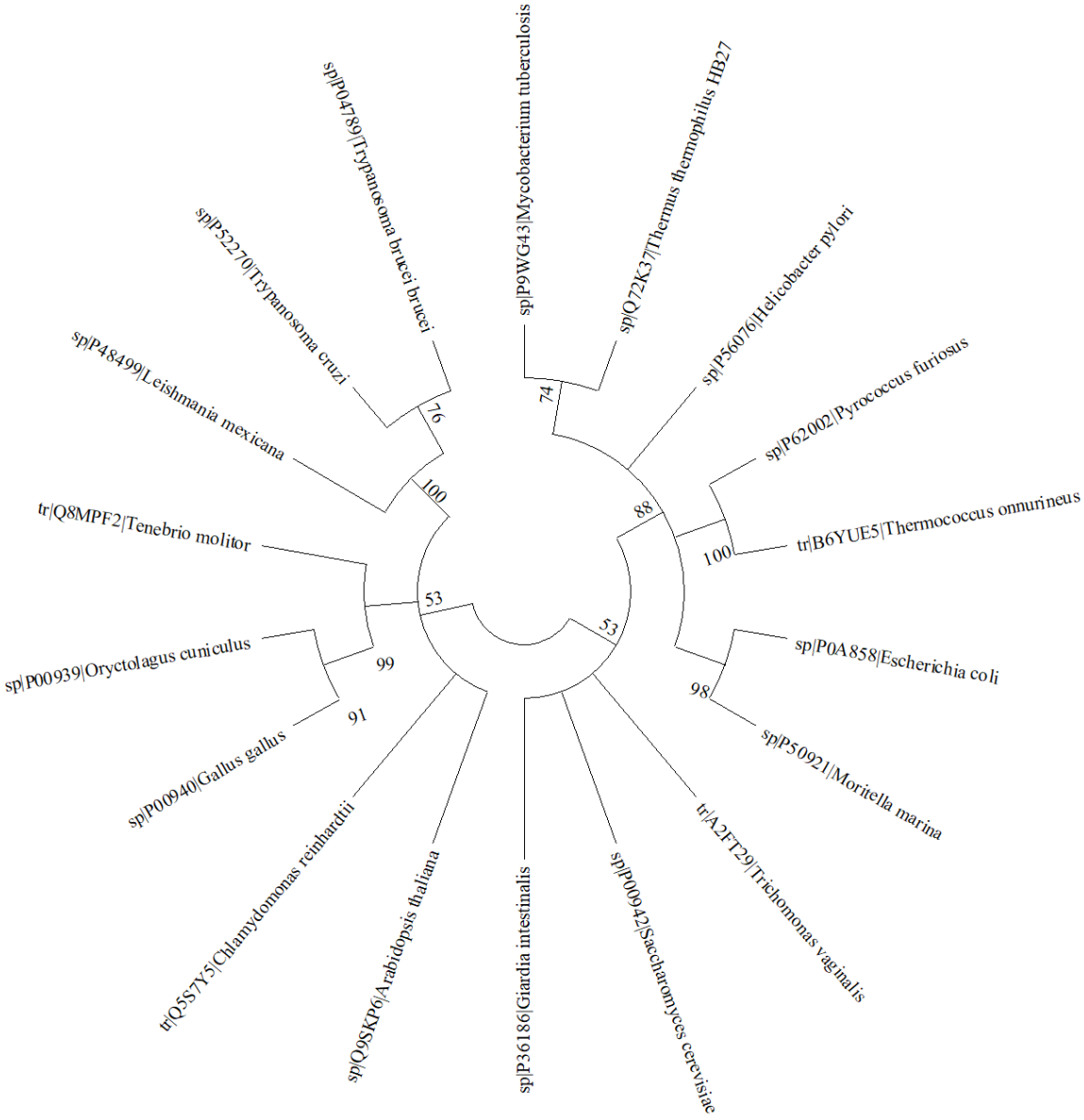
Annexe 13: The phylogenetic tree for the enzyme alpha-glucan transferase (aGP). The phylogeny was constructed using organisms from which enzyme parameters are available in the BRENDA database.

Annexe 14: The phylogenetic tree for the enzyme Phosphoglucoisomerase (PGI). The phylogeny was constructed using organisms from which enzyme parameters are available in the BRENDA database.



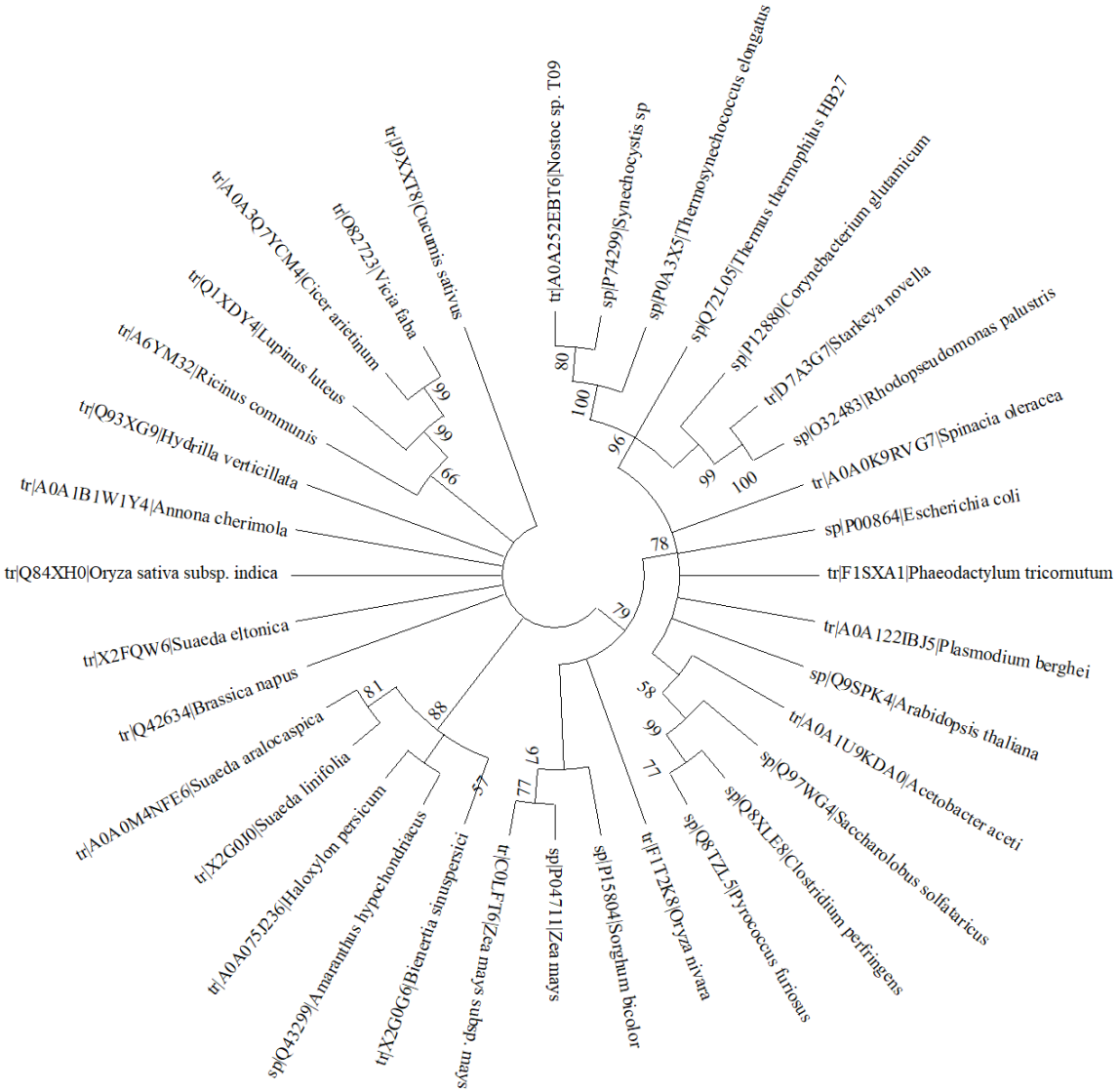
Annexe 15: The phylogenetic tree for the enzyme fructose-bisphosphate aldolase (ALD). The phylogeny was constructed using organisms from which enzyme parameters are available in the BRENDA database.

Annexe 16: The phylogenetic tree for the enzyme triose-phosphate isomerase (TPI). The phylogeny was constructed using organisms from which enzyme parameters are available in the BRENDA database.

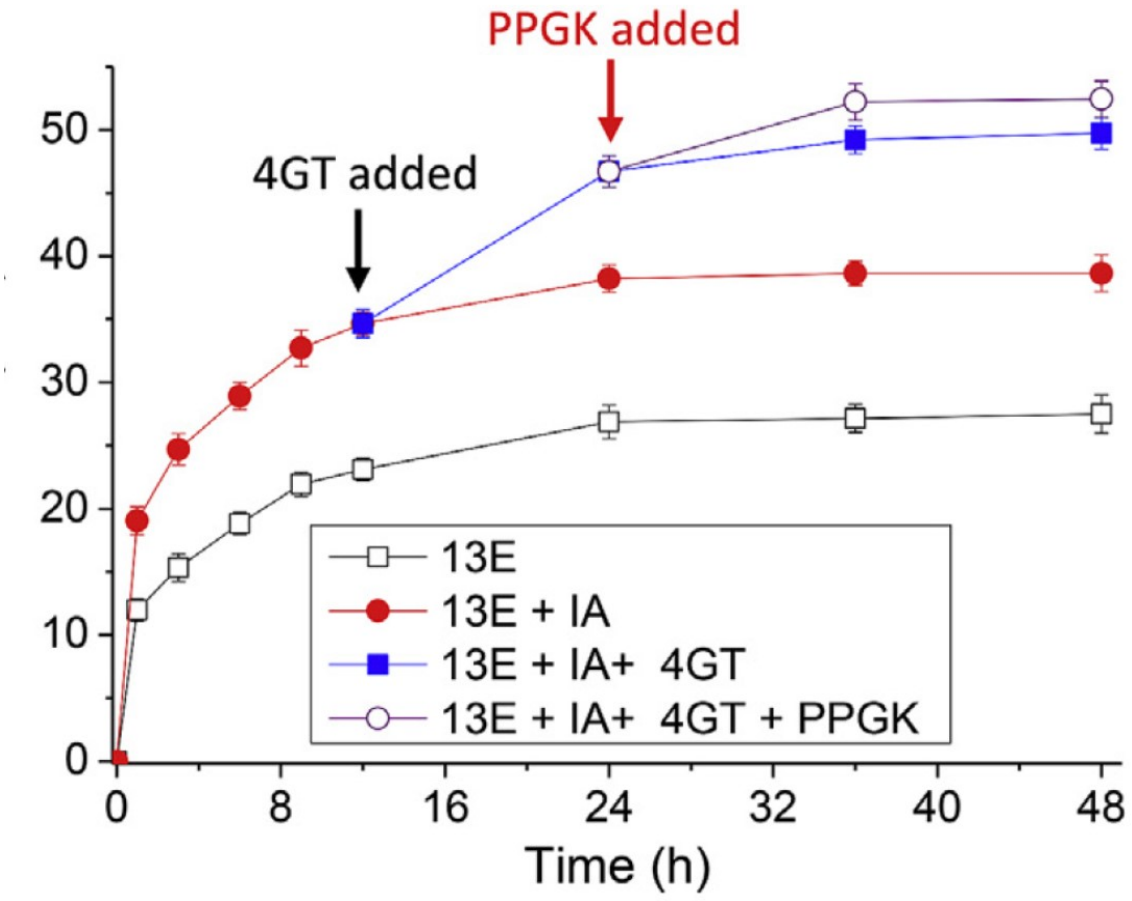


Annexe 17: The phylogenetic tree for the enzyme glyceraldehyde-3-phosphate dehydrogenase (GAPDH). The phylogeny was constructed using organisms from which enzyme parameters are available in the BRENDA database.

Annexe 18: The phylogenetic tree for the enzyme phosphoenolpyruvate carboxylase (PEPC). The phylogeny was constructed using organisms from which enzyme parameters are available in the BRENDA database.



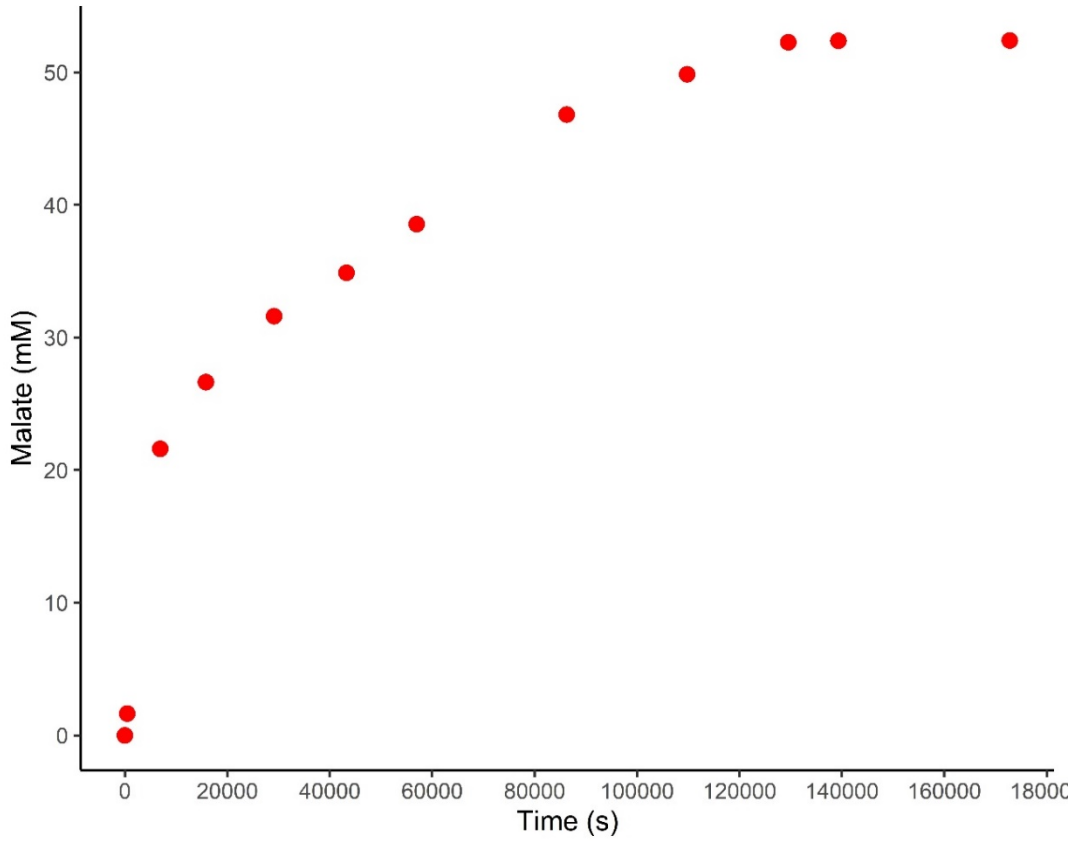
Annexe 19: The experimentally measured malate concentration by Shi *et al*. *via* ATP-balanced malate synthesis pathway (T. Shi *et al.*, 2019). The 13-enzyme cocktail (open square), from IA-treated starch (solid circle), from IA-treated starch supplemented with 4GT (solid square), and from IA-treated starch supplemented with 4GT and PPGK (open circle)



Annexe 20: Comparision of statistics between the three algorithms used for parameter estimation of enzymes GAPDH, GAPN and PGK in the kinetic model, without 4GT and PPGK. RMS is the root mean square between the experimental and model-simulated malate concentration.

	Genetic algorithm			Particle swarm			Hooke Jeeve		
Iteration	**Objective function value**	**RMSE**	**SD**	**Objective function value**	**RMSE**	**SD**	**Objective function value**	**RMSE**	**SD**
1	56.078	2.077	4.324	66.655	2.264	4.714	61.088	2.168	4.512
2	53.573	2.030	4.226	58.549	2.122	4.418	59.168	2.133	4.441
3	53.539	2.029	4.224	55.252	2.062	4.292	58.324	2.118	4.409

Annexe 21: The experimentally measured malate concentration used to fit the kinetic parameters of the PPGK added model. The data points are extracted using WebPlotDegitizer from Annexe 19.



Annexe 22: The range of kinetic parameter values observed in the experimental assays across different sources from the BRENDA database. These kinetic parameters were selected for the estimation to fit the model to experimental data. kcat: turnover number in s^{-1}; Km: Michaelis-Menten constant in mM; Ka: association constant in mM; Ki: inhibition constant in mM.

The experimentally observed parameter values across different sources			
Kinetic parameter	**Biological range**	**Kinetic parameter**	**Biological range**
PFKKiPEP	0.0016 - 0.83	kcatPGK	0.78 - 2633
PFKKaATP	0.005-0.7	GAPNKmG3P	0.02
PFKKaF6P	0.007-254	GAPNKmNAD	1-3.3
kcatPFK	0.015 - 185	kcatGAPN	NA
GAPDHKmG3P	0.00025 - 15	MDHKmMal	0.00012-20
GAPDHKmNAD	0.000032 - 322	MDHKmNAD	0.00087 - 3.32
GAPDHKmPho	0.2 - 37	MDHKmNADH	0.0014 - 1.4
kcatGAPDH	0.002-234	MDHKmOAA	0.0001 - 29.4
PGKKmADP	0.039 - 7.4	kcatMDH	4.71 -4729
PGKKmBPG	0.0005 - 5.6		

List of Equations

List of Tables

List of Figures

References

A Cornish-Bowden and C W Wharton. (1988). *Enzyme kinetics (In Focus)*. oxford: IRL press ltd.

Abe, K., Gomi, K., Hasegawa, F., & Machida, M. (2006). Impact of *Aspergillus oryzae* genomics on industrial production of metabolites. *Mycopathologia*, *162*(3), 143–153. https://doi.org/10.1007/s11046-006-0049-2

Acharjee, A., Kloosterman, B., Visser, R. G. F., & Maliepaard, C. (2016). Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinformatics*, *17*(5), 363–373. https://doi.org/10.1186/s12859-016-1043-4

Ahmed Gamal El-Din, Daniel W. Smith, and M. G. E.-D. (2004). The application of artificial neural network in wastewater treatment. *J. Environ. Eng. Sci*, *3*, 81–95. https://doi.org/doi: 10.1139/S03-067

Ajikumar, P. K., Xiao, W. H., Tyo, K. E. J., Wang, Y., Simeon, F., Leonard, E., … Stephanopoulos, G. (2010). Isoprenoid pathway optimization for Taxol precursor overproduction in Escherichia coli. *Science*, *330*(6000), 70–74. https://doi.org/10.1126/science.1191652

Albe, K. R., & Wright, B. E. (1992). Systems Analysis of the Tricarboxylic Acid Cycle in *Dictyostelium discoideum*. II. control Analysis. *The Journal of Biological Chemistry*, *267*(5), 3106–3114.

Albert A. de Graaf, Mattias Mahle, Michael Mollney, Wolfgang Wiechert, Peter Stahmann, H. S. (2000). Determination of full ^{13}C isotopomer distributions for metabolic flux analysis using heteronuclear spin echo difference NMR spectroscopy. *Journal of Biotechnology*, *77*(1), 25–35. https://doi.org/10.1016/S0168-1656(99)00205-9

Allen, D. K., Libourel, I. G. L., & Shachar-Hill, Y. (2009). Metabolic flux analysis in plants: Coping with complexity. *Plant, Cell and Environment*, *32*(9), 1241–1257. https://doi.org/10.1111/j.1365-3040.2009.01992.x

Almquist, J., Cvijovic, M., Hatzimanikatis, V., Nielsen, J., & Jirstrand, M. (2014). Kinetic models in industrial biotechnology – Improving cell factory performance. *Metabolic Engineering*, *24*, 38–60. https://doi.org/10.1016/j.ymben.2014.03.007

Alsaheb, R. A. A., Aladdin, A., Othman, N. Z., Malek, R. A., Leng, O. M., Aziz, R., & Enshasy, H. A. El. (2015). Lactic acid applications in pharmaceutical and cosmeceutical industries. *Journal of Chemical and Pharmaceutical Research*, *7*(10), 729–735.

Amao, Y., & Ishikawa, M. (2007). Photochemical and Enzymatic Synthesis of Malic Acid from Pyruvic Acid and HCO_3^- with Combination System of Zinc Chlorin-e_6 and Malic Enzyme in Aqueous Medium. *Journal of the Japan Petroleum Institute*, *50*(5), 272–277. https://doi.org/10.1627/jpi.47.222

Anand, P., & Saxena, R. K. (2012). A comparative study of solvent-assisted pretreatment of biodiesel derived crude glycerol on growth and 1,3-propanediol production from *Citrobacter freundii*. *New Biotechnology*, *29*(2), 199–205. https://doi.org/10.1016/j.nbt.2011.05.010

Andersen, M. R., Nielsen, M. L., & Nielsen, J. (2008). Metabolic model integration of the

bibliome, genome, metabolome and reactome of *Aspergillus niger*. *Molecular Systems Biology*, *4*(178). https://doi.org/10.1038/msb.2008.12

Anderson, L. A., Islam, M. A., & Prather, K. L. J. (2018). Synthetic biology strategies for improving microbial synthesis of "green" biopolymers. *Journal of Biological Chemistry*, *293*(14), 5053–5061. https://doi.org/10.1074/jbc.TM117.000368

Angermayr, S. A., Paszota, M., & Hellingwerf, K. J. (2012). Engineering a cyanobacterial cell factory for production of lactic acid. *Applied and Environmental Microbiology*, *78*(19), 7098–7106. https://doi.org/10.1128/AEM.01587-12

Antoniewicz, M. R., Stephanopoulos, G., & Kelleher, J. K. (2006). Evaluation of regression models in metabolic physiology: Predicting fluxes from isotopic data without knowledge of the pathway. *Metabolomics*, *2*(1), 41–52. https://doi.org/10.1007/s11306-006-0018-2

Arturo, J., & Mora, M. (2016). System biology : Mathematical modeling of biological systems. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *5*(8), 2243–2246.

Atsumi, S., Higashide, W., & Liao, J. C. (2009). Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde. *Nature Biotechnology*, *27*(12), 1177–1180. https://doi.org/10.1038/nbt.1586

Ausländer, S., & Fussenegger, M. (2013). From gene switches to mammalian designer cells: Present and future prospects. *Trends in Biotechnology*, *31*(3), 155–168. https://doi.org/10.1016/j.tibtech.2012.11.006

Awan, A. R., Blount, B. A., Bell, D. J., Shaw, W. M., Ho, J. C. H., McKiernan, R. M., & Ellis, T. (2017). Biosynthesis of the antibiotic nonribosomal peptide penicillin in baker's yeast. *Nature Communications*, *8*(May), 1–8. https://doi.org/10.1038/ncomms15202

Balabin, R. M., & Smirnov, S. V. (2012). Interpolation and extrapolation problems of multivariate regression in analytical chemistry : benchmarking the robustness on near-infrared (NIR) spectroscopy data. *Analyst*, *137*(7), 1604–1610. https://doi.org/10.1039/c2an15972d

Battat, E., Peleg, Y., Bercovitz, A., Rokem, J. S., & Goldberg, I. (1991). Optimization of L-malic acid production by *Aspergillus flavus* in a stirred fermentor. *Biotechnology and Bioengineering*, *37*(11), 1108–1116. https://doi.org/10.1002/bit.260371117

Begum, M. F., & Alimon, A. R. (2011). Bioconversion and saccharification of some lignocellulosic wastes by *Aspergillus oryzae* ITCC-4857.01 for fermentable sugar production. *Electronic Journal of Biotechnology*, *14*(5). https://doi.org/10.2225/vol14-issue5-fulltext-2

Bergmeir, C., & Benitez, J. M. (2012). Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. *Journal of Statistical Software*, *46*(7), 1–26. https://doi.org/10.18637/jss.v046.i07

Bisswanger, H. (2014). Enzyme assays. *Perspectives in Science*, *1*, 41–55. https://doi.org/10.1039/b813732c

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., … Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, *326*(5959), 1509–1512. https://doi.org/10.1126/science.1178811

Bradley, R. W., Buck, M., & Wang, B. (2016). Tools and Principles for Microbial Gene Circuit Engineering. *Journal of Molecular Biology*, *428*(5), 862–888. https://doi.org/10.1016/j.jmb.2015.10.004

BRENDA - Information on EC 2.7.1.1 - hexokinase. (n.d.). Retrieved May 22, 2019, from https://www.brenda-enzymes.org/enzyme.php?ecno=2.7.1.1

BRENDA - Information on EC 2.7.1.11 - 6-phosphofructokinase. (n.d.). Retrieved May 22, 2019, from https://www.brenda-enzymes.org/enzyme.php?ecno=2.7.1.9

BRENDA - Information on EC 4.1.2.13 - fructose-bisphosphate aldolase. (n.d.). Retrieved May 22, 2019, from https://www.brenda-enzymes.org/enzyme.php?ecno=4.1.2.13

BRENDA - Information on EC 5.3.1.9 - glucose-6-phosphate isomerase. (n.d.). Retrieved May 22, 2019, from https://www.brenda-enzymes.org/enzyme.php?ecno=5.3.1.9

Brophy, J. A. N., & Voigt, C. A. (2014). Principles of genetic circuit design. *Nature Methods*, *11*(5), 508–520. https://doi.org/10.1038/nmeth.2926

Brown, S. H., Bashkirova, L., Berka, R., Chandler, T., Doty, T., McCall, K., … Berry, A. (2013). Metabolic engineering of *Aspergillus oryzae* NRRL 3488 for increased production of l-malic acid. *Applied Microbiology and Biotechnology*, *97*(20), 8903–8912. https://doi.org/10.1007/s00253-013-5132-2

Cameron, D. E., Bashor, C. J., & Collins, J. J. (2014). A brief history of synthetic biology. *Nature Reviews Microbiology*, *12*(5), 381–390. https://doi.org/10.1038/nrmicro3239

Carbonell, P., Radivojevic, T., & García Martín, H. (2019). Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation. *ACS Synthetic Biology*, *8*(7), 1474–1477. https://doi.org/10.1021/acssynbio.8b00540

Carbonell, P., Wong, J., Swainston, N., Takano, E., Turner, N. J., Scrutton, N. S., … Faulon, J.-L. (2018). Selenzyme: enzyme selection tool for pathway design. *Bioinformatics*, *34*(12), 2153–2154. https://doi.org/10.1093/bioinformatics/bty065

Cardoso Duarte, J., & Costa-Ferreira, M. (1994). *Aspergilli* and lignocellulosics: Enzymology and biotechnological applications. *FEMS Microbiology Reviews*, *13*(2–3), 377–386. https://doi.org/10.1111/j.1574-6976.1994.tb00057.x

Carlson, E. D., Gan, R., Hodgman, C. E., & Jewett, M. C. (2012). Cell-free protein synthesis: Applications come of age. *Biotechnology Advances*, *30*(5), 1185–1194. https://doi.org/10.1016/j.biotechadv.2011.09.016

Chakrabarti, A., Miskovic, L., Soh, K. C., & Hatzimanikatis, V. (2013). Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints. *Biotechnology Journal*, *8*(9), 1043–1057. https://doi.org/10.1002/biot.201300091

Chen, G. Q. (2016). Omics Meets Metabolic Pathway Engineering. *Cell Systems*, *2*(6), 362–363. https://doi.org/10.1016/j.cels.2016.05.005

Chen, Xiulai, Wang, Y., Dong, X., Hu, G., & Liu, L. (2017). Engineering rTCA pathway and C4-dicarboxylate transporter for l-malic acid production. *Applied Microbiology and Biotechnology*, *101*(10), 4041–4052. https://doi.org/10.1007/s00253-017-8141-8

Chen, Xiulai, Xu, G., Xu, N., Zou, W., Zhu, P., Liu, L., & Chen, J. (2013). Metabolic

engineering of *Torulopsis glabrata* for malate production. *Metabolic Engineering*, *19*, 10–16. https://doi.org/10.1016/j.ymben.2013.05.002

Chen, Xuewen, Alonso, A. P., Allen, D. K., Reed, J. L., & Shachar-Hill, Y. (2011). Synergy between $_{13}$C-metabolic flux analysis and flux balance analysis for understanding metabolic adaption to anaerobiosis in *E. coli*. *Metabolic Engineering*, *13*(1), 38–48. https://doi.org/10.1016/j.ymben.2010.11.004

Chen, Y., Banerjee, D., Mukhopadhyay, A., & Petzold, C. J. (2020). Systems and synthetic biology tools for advanced bioproduction hosts. *Current Opinion in Biotechnology*, *64*, 101–109. https://doi.org/10.1016/j.copbio.2019.12.007

Cheng, C., Zhou, Y., Lin, M., Wei, P., & Yang, S. T. (2017). Polymalic acid fermentation by *Aureobasidium pullulans* for malic acid production from soybean hull and soy molasses: Fermentation kinetics and economic analysis. *Bioresource Technology*, *223*, 166–174. https://doi.org/10.1016/j.biortech.2016.10.042

Cherrington, C. A., Hinton, M., Mead, G. C., & Chopra, I. (1991). Organic Acids : Chemistry . Antibacterial Activity and Practical Applications. *Advances in Microbial Physiology*, *32*, 87–108.

Chi, Z., Wang, Z.-P., Wang, G.-Y., Khan, I., & Chi, Z.-M. (2016). Microbial biosynthesis and secretion of l-malic acid and its applications. *Critical Reviews in Biotechnology*, *36*(1), 99–107. https://doi.org/10.3109/07388551.2014.924474

Choi, K. R., Jang, W. D., Yang, D., Cho, J. S., Park, D., & Lee, S. Y. (2019). Systems Metabolic Engineering Strategies: Integrating Systems and Synthetic Biology with Metabolic Engineering. *Trends in Biotechnology*, *37*(8), 817–837. https://doi.org/10.1016/j.tibtech.2019.01.003

Chou, I. C., & Voit, E. O. (2009). Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Mathematical Biosciences*, *219*(2), 57–83. https://doi.org/10.1016/j.mbs.2009.03.002

Christensen, B., & Nielsen, J. (2000). Metabolic Network Analysis of Penicillium chrysogenum Using ^{13}C-Labeled Glucose. *Biotechnology and Bioengineering*, *68*(6), 652–659.

Claudia, S., Quintero, J. C., & Ochoa, S. (2015). Flux Balance Analysis in the Production of Clavulanic Acid by *Streptomyces clavuligerus*. *Biotechnology Progress*, *31*(5), 1226-36. https://doi.org/10.1002/btpr.2132

Clomburg, J. M., Crumbley, A. M., & Gonzalez, R. (2017). Industrial biomanufacturing: The future of chemical production. *Science*, *355*(eaag0804). https://doi.org/10.1126/science.aag0804

Clomburg, J. M., & Gonzalez, R. (2010). Biofuel production in *Escherichia coli*: The role of metabolic engineering and synthetic biology. *Applied Microbiology and Biotechnology*, *86*, 419–434. https://doi.org/10.1007/s00253-010-2446-1

Costello, Z., & Martin, H. G. (2018). A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *Npj Systems Biology and Applications*, (April), 1–14. https://doi.org/10.1038/s41540-018-0054-3

Covert, M. W., Famili, I., & Palsson, B. O. (2003). Identifying Constraints that Govern Cell Behavior: A Key to Converting Conceptual to Computational Models in Biology?

Biotechnology and Bioengineering, *84*(7), 763–772. https://doi.org/10.1002/bit.10849

Cuperlovic-Culf, M. (2018). Machine learning methods for analysis of metabolic data and metabolic pathway modeling. *Metabolites*, *8*(1). https://doi.org/10.3390/metabo8010004

Dai, Z., Zhou, H., Zhang, S., Gu, H., Yang, Q., Zhang, W., … Xin, F. (2018). Current advance in biological production of malic acid using wild type and metabolic engineered strains. *Bioresource Technology*, *258*, 345–353. https://doi.org/10.1016/j.biortech.2018.03.001

Dakin, B. Y. H. D. (1924). The formation of l-Malic Acid as a product of Alcoholic Fermentation by yeast. *J. Biol. Chem.*, *61*, 139–145.

Day, D. A., & Hanson, J. B. (1977). Pyruvate and malate transport and oxidation in corn mitochondria. *Plant Physiology*, *59*, 630–635. https://doi.org/10.1104/pp.59.4.630

Demple, B., Halbrook, J., & Linn, S. (1983). *Escherichia coli* xth mutants are hypersensitive to hydrogen peroxide. *Journal of Bacteriology*, *153*(2), 1079–1082. https://doi.org/10.1128/jb.153.2.1079-1082.1983

Deng, Y., Mao, Y., & Zhang, X. (2016). Metabolic engineering of a laboratory-evolved *Thermobifida fusca* muC strain for malic acid production on cellulose and minimal treated lignocellulosic biomass. *Biotechnology Progress*, *32*(1), 14–20. https://doi.org/10.1002/btpr.2180

Dexter, J., & Fu, P. (2009). Metabolic engineering of cyanobacteria for ethanol production. *Energy and Environmental Science*, *2*(8), 857–864. https://doi.org/10.1039/b811937f

Dondapati, S. K., Pietruschka, G., Thoring, L., Wüstenhagen, D. A., & Kubick, S. (2019). Cell-free synthesis of human toll-like receptor 9 (TLR9): Optimization of synthesis conditions and functional analysis. *PLoS ONE*, *14*(4), 1–16. https://doi.org/10.1371/journal.pone.0215897

Du, J., Yuan, Y., Si, T., Lian, J., & Zhao, H. (2012). Customized optimization of metabolic pathways by combinatorial transcriptional engineering. *Nucleic Acids Research*, *40*(18). https://doi.org/10.1093/nar/gks549

Duarte, N. C., Herrgård, M. J., & Palsson, B. Ø. (2004). Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Research*, *14*(7), 1298–1309. https://doi.org/10.1101/gr.2250904

Dudley, Q. M., Karim, A. S., & Jewett, M. C. (2015). Cell-free metabolic engineering: Biomanufacturing beyond the cell. *Biotechnology Journal*, *10*(1), 69–82. https://doi.org/10.1002/biot.201400330

Dudley, Q. M., Nash, C. J., & Jewett, M. C. (2019). Cell-free biosynthesis of limonene using enzyme-enriched *Escherichia coli* lysates. *Synthetic Biology*, *4*(1). https://doi.org/10.1093/synbio/ysz003

Edgar, R. C. (2004a). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(113). https://doi.org/10.1186/1471-2105-5-113

Edgar, R. C. (2004b). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Edwards, J. S., & Palsson, B. O. (2000). The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, *97*(10), 5528–5533. https://doi.org/10.1073/pnas.97.10.5528

Eide, Å., Johansson, R., Lindblad, T., & Lindsey, C. S. (1997). Data mining and neural networks for knowledge discovery. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *389*(1–2), 251–254. https://doi.org/10.1016/S0168-9002(97)00145-9

El Karoui, M., Hoyos-Flight, M., & Fletcher, L. (2019). Future trends in synthetic biology—a report. *Frontiers in Bioengineering and Biotechnology*, *7*, 1–8. https://doi.org/10.3389/fbioe.2019.00175

Erb, T. J., & Zarzycki, J. (2016). Biochemical and synthetic biology approaches to improve photosynthetic CO_2-fixation. *Current Opinion in Chemical Biology*, *34*, 72–79. https://doi.org/10.1016/j.cbpa.2016.06.026

Fatma, Z., Hartman, H., Poolman, M. G., Fell, D. A., Srivastava, S., Shakeel, T., & Yazdani, S. S. (2018). Model-assisted metabolic engineering of *Escherichia coli* for long chain alkane and alcohol production. *Metabolic Engineering*, *46*, 1–12. https://doi.org/10.1016/j.ymben.2018.01.002

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., … Palsson, B. Ø. (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, *3*(121), 1–18. https://doi.org/10.1038/msb4100155

Fiévet, J. B., Dillmann, C., Curien, G., & de Vienne, D. (2006). Simplified modelling of metabolic pathways for flux prediction and optimization: lessons from an in vitro reconstruction of the upper part of glycolysis. *The Biochemical Journal*, *396*(2), 317–326. https://doi.org/10.1042/BJ20051520

Fontaine, N., Grondin-Perez, B., Cadet, F., & Offmann, B. (2015). Modeling of a Cell-Free Synthetic System for Biohydrogen Production. *Journal of Computer Science & Systems Biology*, *8*(3), 132–139. https://doi.org/10.4172/jcsb.1000181

Förster, J., Famili, I., Fu, P., Palsson, B. Ø., & Nielsen, J. (2003). Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research*, *13*(2), 244–253. https://doi.org/10.1101/gr.234503

Friedman, A., & Kao, C.-Y. (2012). *Mathematical Modeling of Biological Processes*. *Springer International*. Springer. https://doi.org/10.1007/978-3-319-08314-8

Fukushima, A., Kusano, M., Redestig, H., Arita, M., & Saito, K. (2009). Integrated omics approaches in plant systems biology. *Current Opinion in Chemical Biology*, *13*(5–6), 532–538. https://doi.org/10.1016/j.cbpa.2009.09.022

Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., & Kitano, H. (2008). CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proceedings of the IEEE*, *96*(8), 1254–1265.

Funahashi, A., Morohashi, M., Kitano, H., & Tanimura, N. (2003). CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico*, *1*(5), 159–162. https://doi.org/10.1016/s1478-5382(03)02370-9

Fussenegger, M., Morris, R. P., Fux, C., Rimann, M., Von Stockar, B., Thompson, C. J., &

Bailey, J. E. (2000). Streptogramin-based gene regulation systems for mammalian cells. *Nature Biotechnology*, *18*(11), 1203–1208. https://doi.org/10.1038/81208

Geiser, D. M., Pitt, J. I., & Taylor, J. W. (1998). Cryptic speciation and recombination in the aflatoxin-producing fungus *Aspergillus flavus*. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(1), 388–393. https://doi.org/10.1073/pnas.95.1.388

Giorno, L., Drioli, E., Carvoli, G., Cassano, A., & Donato, L. (2001). Study of an enzyme membrane reactor with immobilized fumarase for production of L-malic acid. *Biotechnology & Bioengineering*, *72*(1), 77–84. https://doi.org/10.1002/1097-0290(20010105)72:1<77::aid-bit11>3.3.co;2-c

Giuseppe Carleo, M. T. (2017). Solving the quantum many-body problem with artificial neural networks. *Science*, *355*(6325), 602–606. https://doi.org/10.1007/978-3-642-75430-2_38

Goldberg, I., Rokem, J. S., & Pines, O. (2006). Organic acids: old metabolites, new themes. *Journal of Chemical Technology & Biotechnology*, *81*, 1601–1611. https://doi.org/10.1002/jctb

Goodey, N. M., & Benkovic, S. J. (2008). Allosteric regulation and catalysis emerge via a common route. *Nature Chemical Biology*, *4*(8), 474–482. https://doi.org/10.1038/nchembio.98

Günther, F., & Fritsch, S. (2010). neuralnet : Training of Neural Networks. *The R Journal*, *2*(1), 30–38. https://doi.org/10.32614/RJ-2010-006

Gupta, R., Rathi, P., Gupta, N., & Bradoo, S. (2003). Lipase assays for conventional and molecular screening: an overview. *Biotechnology and Applied Biochemistry*, *37*(1), 63–71. https://doi.org/10.1042/ba20020059

Hakenberg, J., Schmeier, S., Kowald, A., Klipp, E., & Leser, U. (2004). Finding Kinetic Parameters Using Text Mining. *OMICS: A Journal of Integrative Biology*, *8*(2), 131–152. https://doi.org/10.1089/1536231041388366

Hammerschmidt, A., Boukis, N., Hauer, E., Galla, U., Dinjus, E., Hitzmann, B., … Nygaard, S. D. (2011). Catalytic conversion of waste biomass by hydrothermal treatment. *Fuel*, *90*(2), 555–562. https://doi.org/10.1016/j.fuel.2010.10.007

Haris, S., Fang, C., Bastidas-Oyanedel, J. R., Prather, K. J., Schmidt, J. E., & Thomsen, M. H. (2018). Natural antibacterial agents from arid-region pretreated lignocellulosic biomasses and extracts for the control of lactic acid bacteria in yeast fermentation. *AMB Express*, *8*(1), 1–7. https://doi.org/10.1186/s13568-018-0654-8

Hassoun, M. H. (1996). *Fundamentals of Artificial Neural Networks*. *Proceedings of the IEEE* (Vol. 84). Cambridge, London: The MIT press.

Heckmann, D., Lloyd, C. J., Mih, N., Ha, Y., Zielinski, D. C., Haiman, Z. B., … Palsson, B. O. (2018). Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-07652-6

Hirokawa, Y., Maki, Y., Tatsuke, T., & Hanai, T. (2016). Cyanobacterial production of 1,3-propanediol directly from carbon dioxide using a synthetic metabolic pathway. *Metabolic Engineering*, *34*, 97–103. https://doi.org/10.1016/j.ymben.2015.12.008

Hofmann, D., Wirtz, A., Santiago-Schübel, B., Disko, U., & Pohl, M. (2010). Structure elucidation of the thermal degradation products of the nucleotide cofactors NADH and NADPH by nano-ESI-FTICR-MS and HPLC-MS. *Analytical and Bioanalytical Chemistry*, *398*(7–8), 2803–2811. https://doi.org/10.1007/s00216-010-4111-z

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., … Kummer, U. (2006). COPASI--a COmplex PAthway SImulator. *Bioinformatics*, *22*(24), 3067–3074. https://doi.org/10.1093/bioinformatics/btl485

Hornik, K., Maxwell, S., & White, H. (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, *2*, 359–366. https://doi.org/https://doi.org/10.1016/0893-6080(89)90020-8

Huang, L., Sheng, J., Xu, Z., Zhu, X., & Cai, J. (2014). Reconstitution of the peptidoglycan cytoplasmic precursor biosynthetic pathway in cell-free system and rapid screening of antisense oligonucleotides for Mur enzymes. *Applied Microbiology and Biotechnology*, *98*(4), 1785–1794. https://doi.org/10.1007/s00253-013-5467-8

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., … Wang, J. (2003). The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, *19*(4), 524–531. https://doi.org/10.1093/bioinformatics/btg015

Imlay, J. A., Chin, S. M., & Linnt, S. (1986). Toxic DNA Damage by Hydrogen Peroxide Through the Fenton Reaction in Vivo and in Vitro. *Science*, *240*, 640–642.

Inoue, H., Yamachika, M., & Yoneyama, H. (1992). Photocatalytic conversion of lactic acid to malic acid through pyruvic acid in the presence of malic enzyme and semiconductor photocatalysts. *Journal of the Chemical Society, Faraday Transactions*, *88*(15), 2215–2219. https://doi.org/10.1039/FT9928802215

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer*, *29*(3), 31–44. https://doi.org/10.1109/2.485891

Jamshidi, N., & Palsson, B. (2007). Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. *BMC Systems Biology*, *1*(26), 1–20. https://doi.org/10.1186/1752-0509-1-26

Jantama, K., Haupt, M. J., Svoronos, S. A., Zhang, X., Moore, J. C., Shanmugam, K. T., & Ingram, L. O. (2008). Combining Metabolic Engineering and Metabolic Evolution to Develop Nonrecombinant Strains of *Escherichia coli* C That Produce Succinate and Malate. *Biotechnology & Bioengineering*, *99*(5), 1140–1153. https://doi.org/10.1002/bit.21694

Joseph M Dale, Liviu Popescu, P. D. K. (2010). Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics*, *11*, 15. https://doi.org/10.1186/1471-2105-11-15

Kamruzzaman, S. M., & Jehad Sarkar, A. M. (2011). A new data mining scheme using artificial neural networks. *Sensors*, *11*, 4622–4647. https://doi.org/10.3390/s110504622

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1), 27–30. https://doi.org/10.1016/j.meegid.2016.07.022

Keasling, J. D. (2012). Synthetic biology and the development of tools for metabolic engineering. *Metabolic Engineering*, *14*(3), 189–195.

https://doi.org/10.1016/j.ymben.2012.01.004

Khattak, W. A., Ul-Islam, M., Ullah, M. W., Yu, B., Khan, S., & Park, J. K. (2014). Yeast cell-free enzyme system for bio-ethanol production at elevated temperatures. *Process Biochemistry*, *49*(3), 357–364. https://doi.org/10.1016/j.procbio.2013.12.019

Kim, Y. G., & Chandrasegaran, S. (1994). Chimeric restriction endonuclease. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(3), 883–887. https://doi.org/10.1073/pnas.91.3.883

Kimura, T., Kawabata, Y., & Sato, E. (1986). Enzymatic production of l-malate from maleate by *Alcaligenes* sp. *Agricultural and Biological Chemistry*, *50*(1), 89–94. https://doi.org/10.1080/00021369.1986.10867349

Klement, T., & Büchs, J. (2013). Itaconic acid - A biotechnological process in change. *Bioresource Technology*, *135*, 422–431. https://doi.org/10.1016/j.biortech.2012.11.141

Knuf, C., Nookaew, I., Brown, S. H., McCulloch, M., Berry, A., & Nielsen, J. (2013). Investigation of malic acid production in *Aspergillus oryzae* under nitrogen starvation conditions. *Applied and Environmental Microbiology*, *79*(19), 6050–6058. https://doi.org/10.1128/AEM.01445-13

Kojima, R., Uchiya, K., Manshio, H., & Masuda, K. (2020). Cell-free synthesis of functionally active HSPB5. *Cell Stress and Chaperones*, *25*(4), 1–15. https://doi.org/10.1007/s12192-020-01073-5

Kolanoski, H. (1995). Application of artificial neural networks in particle physics. *Nuclear Instruments and Methods in Physics Research, Section A*, *367*, 14–20. https://doi.org/10.1007/3-540-61510-5_1

Kotula, J. W., Kerns, S. J., Shaket, L. A., Siraj, L., Collins, J. J., Way, J. C., & Silver, P. A. (2014). Programmable bacteria detect and record an environmental signal in the mammalian gut. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(13), 4838–4843. https://doi.org/10.1073/pnas.1321321111

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Kumar, V., Bhalla, A., & Rathore, A. S. (2014). Design of experiments applications in bioprocessing: Concepts and approach. *Biotechnology Progress*, *30*(1), 86–99. https://doi.org/10.1002/btpr.1821

Kuttler, C. (2009). Mathematical models in biology. Retrieved from http://www-m6.ma.tum.de/~kuttler/script1.pdf

Lan, Z., Zhao, C., Guo, W., Guan, X., & Zhang, X. (2015). Optimization of culture medium for maximal production of spinosad using an artificial neural network-genetic algorithm modeling. *Journal of Molecular Microbiology and Biotechnology*, *25*(4), 253–261. https://doi.org/10.1159/000381312

Le, S., Josse, J., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, *25*(1). https://doi.org/10.18637/jss.v025.i01

Lee, J. W., Na, D., Park, J. M., Lee, J., Choi, S., & Lee, S. Y. (2012). Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nature Chemical*

Biology, *8*(6), 536–546. https://doi.org/10.1038/nchembio.970

Lettvin, J. Y., Maturana, H. R., Mcculloch, W. S., & Pitts, W. H. (1959). What the Frog's Eye Tells the Frog's Brain. Proceedings of the. *Proceedings of the IRE*, *47*, 1940–1959.

Li, C., Zhang, R., Wang, J., Wilson, L. M., & Yan, Y. (2020). Protein Engineering for Improving and Diversifying Natural Product Biosynthesis. *Trends in Biotechnology*, 1–16. https://doi.org/10.1016/j.tibtech.2019.12.008

Li, Y., & Smolke, C. D. (2016). Engineering biosynthesis of the anticancer alkaloid noscapine in yeast. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms12137

Li, Z., Hong, P., Da, Y., Li, L., & Stephanopoulos, G. (2018). Metabolic engineering of *Escherichia coli* for the production of L-malate from xylose. *Metabolic Engineering*, *48*, 25–32. https://doi.org/10.1016/j.ymben.2018.05.010

Lin, S., Wang, L., Jones, G., Trang, H., Yin, Y., & Liu, J. (2012). Optimized extraction of calcium malate from eggshell treated by PEF and an absorption assessment in vitro. *International Journal of Biological Macromolecules*, *50*(5), 1327–1333. https://doi.org/10.1016/j.ijbiomac.2012.03.013

Liu, C. C., Qi, L., Lucks, J. B., Segall-Shapiro, T. H., Wang, D., Mutalik, V. K., & Arkin, A. P. (2012). An adaptor from translational to transcriptional control enables predictable assembly of complex regulation. *Nature Methods*, *9*(11), 1088–1094. https://doi.org/10.1038/nmeth.2184

Liu, J., Li, J., Shin, H., Liu, L., Du, G., & Chen, J. (2017). Protein and metabolic engineering for the production of organic acids. *Bioresource Technology*, *239*, 412–421. https://doi.org/10.1016/J.BIORTECH.2017.04.052

Liu ZeLin, Peng ChangHui, Xiang WenHua, Tian DaLun, Deng XiangWen, Z. M. (2010). Application of artificial neural networks in global climate change and ecological research: An overview. *Chinese Science Bulletin*, *55*(34), 3853–3863. https://doi.org/10.1007/s11434-010-4183-3

Lu, Y. (2017). Cell-free synthetic biology: Engineering in an open world. *Synthetic and Systems Biotechnology*, *2*(1), 23–27. https://doi.org/10.1016/j.synbio.2017.02.003

Ludwig, M. (2016). The roles of organic acids in C_4 photosynthesis. *Frontiers in Plant Science*, *7*, 1–11. https://doi.org/10.3389/fpls.2016.00647

Lumyong, S., & Tomita, F. (1993). L-malic acid production by an albino strain of *Monascus araneosus*. *World Journal of Microbiology and Biotechnology*, *9*, 383–384. https://doi.org/10.1007/BF00383086

Manuela, C.-R., & Leao, C. (1990). Transport of Malic Acid and Other Dicarboxylic Acids in the Yeast *Hansenula anomala*. *Applied and Environmental Microbiology*, *56*(4), 1109–1113.

Marion M. Bradford. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, *72*(1–2), 248–254. https://doi.org/https://doi.org/10.1016/0003-2697(76)90527-3

Martínez, J. A., Bolívar, F., & Escalante, A. (2015). Shikimic acid production in *Escherichia coli*: from classical metabolic engineering strategies to omics applied to improve its

production. *Frontiers in Bioengineering and Biotechnology*, *3*(September), 1–16. https://doi.org/10.3389/fbioe.2015.00145

Mcculloch, W. S., & Pitts, W. (1990). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, *52*(1/2), 99–115. https://doi.org/10.1007/BF02478259

Minns, A. W., & Hall, M. J. (1996). Artificial neural networks as rainfall- runoff models. *Hydrological Sciences Journal*, *41*(3), 399–417. https://doi.org/10.1080/02626669609491511

Mishra, B., Kumar, N., & Mukhtar, M. S. (2019). Systems biology and machine learning in plant–pathogen interactions. *Molecular Plant-Microbe Interactions*, *32*(1), 45–55. https://doi.org/10.1094/MPMI-08-18-0221-FI

Mistry, A. N., Ganta, U., Chakrabarty, J., & Dutta, S. (2019). A review on biological systems for CO_2 sequestration: Organisms and their pathways. *Environmental Progress and Sustainable Energy*, *38*(1), 127–136. https://doi.org/10.1002/ep.12946

Miyawaki, A., Llopis, J., Heim, R., Michael McCaffery, J., Adams, J. A., Ikura, M., & Tsien, R. Y. (1997). Fluorescent indicators for Ca2+ based on green fluorescent proteins and calmodulin. *Nature*, *388*(6645), 882–887. https://doi.org/10.1038/42264

Mondala, A. H. (2015). Direct fungal fermentation of lignocellulosic biomass into itaconic, fumaric, and malic acids: current and future prospects. *Journal of Industrial Microbiology and Biotechnology*, *42*(4), 487–506. https://doi.org/10.1007/s10295-014-1575-4

Morowvat, M. H., & Ghasemi, Y. (2016). Medium optimization by artificial neural networks for maximizing the triglycerides-rich lipids from biomass of *Chlorella vulgaris*. *International Journal of Pharmaceutical and Clinical Research*, *8*(10), 1414–1417.

Moscou, M. J., & Bogdanove, A. J. (2009). A Simple Cipher Governs DNA Recognition by TAL Effectors. *Science (New York, N.Y.)*, *326*(December), 1501. https://doi.org/10.1126/science.1178817

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., … Fraser, C. M. (1999). Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, *399*, 323–329. https://doi.org/10.1038/20601

Neufeid, R. J., Peleg, Y., Rokem, J. S., Pinest, O., & Goldberg, I. (1998). L-Malic acid formation by immobilized *Saccharomyces cerevisiae* amplified for fumarase. *Enzyme Microb. Technol.*, *13*(11), S.10-S.11. https://doi.org/10.1016/S0168-9525(98)80006-1

Nieves, L. M., Panyon, L. A., & Wang, X. (2015). Engineering sugar utilization and microbial tolerance toward lignocellulose conversion. *Frontiers in Bioengineering and Biotechnology*, *3*, 1–10. https://doi.org/10.3389/fbioe.2015.00017

Nikoloski, Z., Perez-Storey, R., & Sweetlove, L. J. (2015). Inference and prediction of metabolic network fluxes. *Plant Physiology*, *169*, 1443–1455. https://doi.org/10.1104/pp.15.01082

Nookaew, I., Olivares-Hernández, R., Bhumiratana, S., & Nielsen, J. (2011). Genome-Scale Metabolic Models of *Saccharomyces cerevisiae*. In *Methods in molecular biology* (Vol. 759, pp. 445–463). https://doi.org/10.1007/978-1-61779-173-4_25

Orth, J. D., Thiele, I., & Palsson, B. O. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*(3), 245–248. https://doi.org/10.1038/nbt.1614

Osothslip, C., & Subden, R. E. (1986). Malate transport in *Schizosaccharomyces pombe*. *Journal of Bacteriology*, *168*(3), 1439–1443.

Oswald, F., Dörsam, S., Veith, N., Zwick, M., & Neumann, A. (2016). Sequential Mixed Cultures : From Syngas to Malic Acid. *Frontiers in Microbiology*, *7*(June), 1–12. https://doi.org/10.3389/fmicb.2016.00891

Pardee, K., Green, A. A., Ferrante, T., Cameron, D. E., Daleykeyser, A., Yin, P., & Collins, J. J. (2014). Paper-based synthetic gene networks. *Cell*, *159*(4), 940–954. https://doi.org/10.1016/j.cell.2014.10.004

Pardee, K., Green, A. A., Takahashi, M. K., Braff, D., Lambert, G., Lee, J. W., … Collins, J. J. (2016). Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components. *Cell*, *165*(5), 1255–1266. https://doi.org/10.1016/j.cell.2016.04.059

Park, S., Chang, K. S., Jin, E., Pack, S. P., & Lee, J. (2013). Oxaloacetate and malate production in engineered *Escherichia coli* by expression of codon-optimized phosphoenolpyruvate carboxylase2 gene from *Dunaliella salina*. *Bioprocess and Biosystems Engineering*, *36*(1), 127–131. https://doi.org/10.1007/s00449-012-0759-4

Pawul, M., & Śliwka, M. (2016). Application of artificial neural networks for prediction of air pollution levels in environmental monitoring. *Journal of Ecological Engineering*, *17*(4), 190–196. https://doi.org/10.12911/22998993/64828

Peleg, Y., Barak, A., Scrutton, M. C., & Goldberg, I. (1989). Malic acid accumulation by Aspergillus flavus - III. 13C NMR and isoenzyme analyses. *Applied Microbiology and Biotechnology*, *30*(2), 176–183. https://doi.org/10.1007/BF00264008

Peleg, Y., Stieglitz, B., & Goldberg, I. (1988). Malic acid accumulation by *Aspergillus flavus* I. Biochemical aspects of acid biosynthesis. *Applied Microbiology Biotechnology*, *28*, 69–75.

Pereira, B., Miguel, J., Vilaça, P., Soares, S., Rocha, I., & Carneiro, S. (2018). Reconstruction of a genome-scale metabolic model for *Actinobacillus succinogenes* 130Z. *BMC Systems Biology*, *12*. https://doi.org/10.1186/s12918-018-0585-7

Plaimas, K., Mallm, J. P., Oswald, M., Svara, F., Sourjik, V., Eils, R., & König, R. (2008). Machine learning based analyses on metabolic networks supports high-throughput knockout screens. *BMC Systems Biology*, *2*, 1–11. https://doi.org/10.1186/1752-0509-2-67

Prathumpai, W., Gabelgaard, J. B., Wanchanthuek, P., Van De Vondervoort, P. J. I., De Groot, M. J. L., McIntyre, M., & Nielsen, J. (2003). Metabolic control analysis of xylose catabolism in *Aspergillus*. *Biotechnology Progress*, *19*(4), 1136–1141. https://doi.org/10.1021/bp034020r

Price, N. D., Reed, J. L., & Palsson, B. (2004). Genome-scale models of microbial cells: Evaluating the consequences of constraints. *Nature Reviews Microbiology*, *2*(11), 886–897. https://doi.org/10.1038/nrmicro1023

Raman, K., & Chandra, N. (2009). Flux balance analysis of biological systems: Applications and challenges. *Briefings in Bioinformatics*, *10*(4), 435–449. https://doi.org/10.1093/bib/bbp011

Raman, S., Taylor, N., Genuth, N., Fields, S., & Church, G. M. (2014). Engineering allostery. *Trends in Genetics*, *30*(12), 521–528. https://doi.org/10.1016/j.tig.2014.09.004

Reed, J. L., & Palsson, B. Ø. (2003). Thirteen Years of Building Constraint-Based In Silico Models of *Escherichia coli*. *Journal of Biotechnology*, *185*(9), 2692–2699. https://doi.org/10.1128/JB.185.9.2692

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3), 303–304. https://doi.org/10.1038/nbt0308-303

Rios-Estepa, R., & Lange, B. M. (2007). Experimental and mathematical approaches to modeling plant metabolic networks. *Phytochemistry*, *68*(16–18), 2351–2374. https://doi.org/10.1016/j.phytochem.2007.04.021

Rohwer, J. M. (2012). Kinetic modelling of plant metabolic pathways. *Journal of Experimental Botany*, *63*(6), 2275–2292. https://doi.org/10.1093/jxb/ers080

Rokni, M. (2015). Thermodynamic analyses of municipal solid waste gasification plant integrated with solid oxide fuel cell and Stirling hybrid system. *International Journal of Hydrogen Energy*, *40*(24), 7855–7869. https://doi.org/10.1016/j.ijhydene.2014.11.046

Rollin, J. A., Tam, T. K., & Zhang, Y. H. P. (2013). New biotechnology paradigm: Cell-free biosystems for biomanufacturing. *Green Chemistry*, *15*(7), 1708–1719. https://doi.org/10.1039/c3gc40625c

Rosenberg, M., Miková, H., & Krištofíková, L. (1999). Formation of L-malic acid by yeasts of the genus *Dipodascus*. *Letters in Applied Microbiology*, *29*(4), 221–223. https://doi.org/10.1046/j.1365-2672.1999.00601.x

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rudolph, B., Hansen, T., & Schönheit, P. (2004). Glucose-6-phosphate isomerase from the hyperthermophilic archaeon *Methanococcus jannaschii*: Characterization of the first archaeal member of the phosphoglucose isomerase superfamily. *Archives of Microbiology*, *181*(1), 82–87. https://doi.org/10.1007/s00203-003-0626-4

Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S., & Palsson, B. O. (2002). Genome-Scale Metabolic Model of *Helicobacter pylori* 26695. *Journal of Biotechnology*, *184*(16), 4582–4593. https://doi.org/10.1128/JB.184.16.4582

Schilling, C. H., Edwards, J. S., Letscher, D., & Palsson, B. Ø. (2000). Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnology and Bioengineering*, *71*(4), 286–306. https://doi.org/10.1002/1097-0290(2000)71:4<286::aid-bit1018>3.3.co;2-i

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schoborg, J. A., Hodgman, C. E., Anderson, M. J., & Jewett, M. C. (2014). Substrate replenishment and byproduct removal improve yeast cell-free protein synthesis. *Biotechnology Journal*, *9*(5), 630–640. https://doi.org/10.1002/biot.201300383

Schomburg, I., Chang, A., & Schomburg, D. (2002). BRENDA, enzyme data and metabolic information. *Nucleic Acids Research*, *30*(1), 47–49. Retrieved from

http://www.ncbi.nlm.nih.gov/pubmed/11752250%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC99121

Schuster, S., Fell, D. A., & Dandekar, T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, *18*(3), 326–332. https://doi.org/10.1038/73786

Shi, J., Jiang, Y., Jiang, Z., Wang, X., Wang, X., Zhang, S., … Yang, C. (2015). Enzymatic conversion of carbon dioxide. *Chemical Society Reviews*, *44*(17), 5981–6000. https://doi.org/10.1039/c5cs00182j

Shi, T., Liu, S., & Zhang, Y. P. J. (2019). CO_2 fixation for malate synthesis energized by starch via in vitro metabolic engineering. *Metabolic Engineering*, *55*, 152–160. https://doi.org/10.1016/j.ymben.2019.07.005

Shrestha, P., Holland, T. M., & Bundy, B. C. (2012). Streamlined extract preparation for *Escherichia coli*-based cell-free protein synthesis by sonication or bead vortex mixing. *BioTechniques*, *53*(3), 163–174. https://doi.org/10.2144/0000113924

Siciliano, V., Diandreth, B., Monel, B., Beal, J., Huh, J., Clayton, K. L., … Weiss, R. (2018). Engineering modular intracellular protein sensor-actuator devices. *Nature Communications*, *9*(1881), 1–7. https://doi.org/10.1038/s41467-018-03984-5

Singh, R. K., Singh, R., Sivakumar, D., Kondaveeti, S., Kim, T., Li, J., … Lee, J. K. (2018). Insights into Cell-Free Conversion of CO_2 to Chemicals by a Multienzyme Cascade Reaction. *ACS Catalysis*, *8*(12), 11085–11093. research-article. https://doi.org/10.1021/acscatal.8b02646

Soetaert, K. (2016). plot3Drgl: Plotting Multi-Dimensional Data - Using "rgl." Retrieved from https://cran.r-project.org/package=plot3Drgl

Soetaert, K. (2017). plot3D: Plotting Multi-Dimensional Data. Retrieved from https://cran.r-project.org/package=plot3D

Somasundaram, S., Eom, G. T., & Hong, S. H. (2018). Efficient Malic Acid Production in *Escherichia coli* Using a Synthetic Scaffold Protein Complex. *Applied Biochemistry and Biotechnology*, *184*(4), 1308–1318. https://doi.org/10.1007/s12010-017-2629-7

Song, C. W., Kim, D. I., Choi, S., Jang, J. W., & Lee, S. Y. (2013). Metabolic engineering of *Escherichia coli* for the production of fumaric acid. *Biotechnology and Bioengineering*, *110*(7), 2025–2034. https://doi.org/10.1002/bit.24868

Song, X., Yu, H., & Zhu, K. (2016). Improving alkane synthesis in *Escherichia coli* via metabolic engineering. *Applied Microbiology and Biotechnology*, *100*(2), 757–767. https://doi.org/10.1007/s00253-015-7026-y

Soo Yun Moon; Soon Ho Hong; Tae Yong Kim; Sang Yup Lee. (2008). Metabolic engineering of *Escherichia coli* for the production of malic acid. *Biochemical Engineering Journal*, *40*, 312–320. https://doi.org/10.1016/j.bej.2008.01.001

Srinivasan, S., Cluett, W. R., & Mahadevan, R. (2015). Constructing kinetic models of metabolism at genome-scales: A review. *Biotechnology Journal*, *10*(9), 1345–1359. https://doi.org/10.1002/biot.201400522

Stelling, J. (2004). Mathematical models in microbial systems biology. *Current Opinion in Microbiology*, *7*(5), 513–518. https://doi.org/10.1016/j.mib.2004.08.004

Stephen A. Osmani and Michael C. Scrutton. (1983). The Sub-Cellular Localisation of Pyruvate Carboxylase and of Some Other Enzymes in *Aspergillus nidulans*. *European Journal of Biochemistry*, *133*(3), 551–560. https://doi.org/10.1111/j.1432-1033.1983.tb07499.x

Steuer, R., Gross, T., Selbig, J., & Blasius, B. (2006). Structural kinetic modeling of metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(32), 11868–11873. https://doi.org/10.1073/pnas.0600013103

Taing, O., & Taing, K. (2007). Production of malic and succinic acids by sugar-tolerant yeast *Zygosaccharomyces rouxii*. *European Food Research and Technology*, *224*(3), 343–347. https://doi.org/10.1007/s00217-006-0323-z

Tan, S. Z., & Prather, K. L. (2017). Dynamic pathway regulation: recent advances and methods of construction. *Current Opinion in Chemical Biology*, *41*, 28–35. https://doi.org/10.1016/j.cbpa.2017.10.004

Taniguchi, H., Okano, K., & Honda, K. (2017). Modules for in vitro metabolic engineering: Pathway assembly for bio-based production of value-added chemicals. *Synthetic and Systems Biotechnology*, *2*(2), 65–74. https://doi.org/10.1016/j.synbio.2017.06.002

Teusink, B., Passarge, J., Reijenga, C. A., Esgalhado, E., van der Weijden, C. C., Schepper, M., … Snoep, J. L. (2000). Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *European Journal of Biochemistry*, *267*(17), 5313–5329. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10951190

Therneau, T., & Atkinson, B. (2018). rpart: Recursive Partitioning and Regression Trees.

Trichez, D., Auriol, C., Baylac, A., Irague, R., Dressaire, C., Carnicer-Heras, M., … Walther, T. (2018). Engineering of Escherichia coli for Krebs cycle-dependent production of malic acid. *Microbial Cell Factories*, *17*(1), 1–12. https://doi.org/10.1186/s12934-018-0959-y

Vasilakou, E., Machado, D., Theorell, A., Rocha, I., Nöh, K., Oldiges, M., & Wahl, S. A. (2016). Current state and challenges for dynamic metabolic modeling. *Current Opinion in Microbiology*, *33*(1), 97–104. https://doi.org/10.1016/j.mib.2016.07.008

Vater, A., & Klussmann, S. (2015). Turning mirror-image oligonucleotides into drugs: The evolution of Spiegelmer® therapeutics. *Drug Discovery Today*, *20*(1), 147–155. https://doi.org/10.1016/j.drudis.2014.09.004

Vemuri, G. N., & Aristidou, A. A. (2005). Metabolic Engineering in the -omics Era: Elucidating and Modulating Regulatory Networks. *Microbiology and Molecular Biology Reviews*, *69*(2), 197–216. https://doi.org/10.1128/mmbr.69.2.197-216.2005

Venables, W. N., & B. D. Ripley. (2002). *Modern Applied Statistics with S-Plus*. *Springer* (4th ed.). New York: Springer. https://doi.org/10.2307/2685660

Vijayakumar, S., Conway, M., Lió, P., & Angione, C. (2017). Seeing the wood for the trees: a forest of methods for optimization and omic-network integration in metabolic modelling. *Briefings in Bioinformatics*, *19*(6), 1–18. https://doi.org/10.1093/bib/bbx053

Wals, K., & Ovaa, H. (2014). Unnatural amino acid incorporation in E. coli: Current and future applications in the design of therapeutic proteins. *Frontiers in Chemistry*, *2*(APR), 1–12. https://doi.org/10.3389/fchem.2014.00015

Wang, P., Wang, Y., Guo, X., Huang, S., & Zhu, G. (2020). Biochemical and phylogenetic

characterization of a monomeric isocitrate dehydrogenase from a marine methanogenic archaeon *Methanococcoides methylutens*. *Extremophiles*, *20*, 319–328. https://doi.org/10.1007/s00792-020-01156-2

Wang, Y. K., Chi, Z., Zhou, H. X., Liu, G. L., & Chi, Z. M. (2015). Enhanced production of Ca^{2+}-polymalate (PMA) with high molecular mass by *Aureobasidium pullulans* var. pullulans MCW. *Microbial Cell Factories*, *14*(1), 1–11. https://doi.org/10.1186/s12934-015-0296-3

Ward, V. C. A., Chatzivasileiou, A. O., & Stephanopoulos, G. (2019). Cell free biosynthesis of isoprenoids from isopentenol. *Biotechnology and Bioengineering*, *116*(12), 3269–3281. https://doi.org/10.1002/bit.27146

Weaver, D. S., Keseler, I. M., Mackie, A., Paulsen, I. T., & Karp, P. D. (2014). A genome-scale metabolic flux model of *Escherichia coli* K – 12 derived from the EcoCyc database. *BMCSystems Biology*, *8*(79), 1–24. https://doi.org/10.1186/1752-0509-8-79

Weber, W., Fux, C., Daoud-El Baba, M., Keller, B., Weber, C. C., Kramer, B. P., … Fussenegger, M. (2002). Macrolide-based transgene control in mammalian cells and mice. *Nature Biotechnology*, *20*(9), 901–907. https://doi.org/10.1038/nbt731

Wei, P., Cheng, C., Lin, M., Zhou, Y., & Yang, S. T. (2017). Production of poly(malic acid) from sugarcane juice in fermentation by Aureobasidium pullulans: Kinetics and process economics. *Bioresource Technology*, *224*, 581–589. https://doi.org/10.1016/j.biortech.2016.11.003

Weihs, C., Ligges, U., Luebke, K., & Raabe, N. (2005). klaR Analyzing German Business Cycles. In D. Baier, R. Decker, & L. Schmidt-Thieme (Eds.), *Data Analysis and Decision Support* (pp. 335–343). Berlin: Springer-Verlag. https://doi.org/10.1007/3-540-28397-8_36

Weissman, K. J., & Leadlay, P. F. (2005). Combinatorial biosynthesis of reduced polyketides. *Nature Reviews Microbiology*, *3*(12), 925–936. https://doi.org/10.1038/nrmicro1287

Werpy, T., & Petersen, G. (2004). *Top Value Added Chemicals from Biomass Volume I — Results of Screening for Potential Candidates from Sugars and Synthesis Gas*. https://doi.org/10.2172/15008859

West, T. P. (2011). Malic acid production from thin stillage by *Aspergillus* species. *Biotechnology Letters*, *33*(12), 2463–2467. https://doi.org/10.1007/s10529-011-0720-7

West, T. P. (2015a). Fungal biotransformation of crude glycerol into malic acid. *Zeitschrift Fur Naturforschung - Section C Journal of Biosciences*, *70*(5–6), 165–167. https://doi.org/10.1515/znc-2015-0115

West, T. P. (2015b). Fungal biotransformation of crude glycerol into malic acid. *Zeitschrift Fur Naturforschung - Section C Journal of Biosciences*, *70*(5–6), 165–167. https://doi.org/10.1515/znc-2015-0115

Wheeldon, I., Minteer, S. D., Banta, S., Barton, S. C., Atanassov, P., & Sigman, M. (2016). Substrate channelling as an approach to cascade reactions. *Nature Chemistry*, *8*(4), 299–309. https://doi.org/10.1038/nchem.2459

Wheelock, C. E., Goss, V. M., Balgoma, D., Nicholas, B., Brandsma, J., Skipp, P. J., … U-BIOPRED Study group. (2013). Application of 'omics technologies to biomarker discovery in inflammatory lung diseases. *European Respiratory Journal*, *42*(3), 802–825.

https://doi.org/10.1183/09031936.00078812

Wiechert, W. (2001). 13C Metabolic Flux Analysis. *Metabolic Engineering*, *3*, 195–206. https://doi.org/10.1006/MBEN.2001.0187

Win, M. N., & Smolke, C. D. (2008). Higher-Order Cellular Information Processing with Synthetic RNA Devices, (October).

Wittig, U., Kania, R., Golebiewski, M., Rey, M., Shi, L., Jong, L., … Müller, W. (2012). SABIO-RK - Database for biochemical reaction kinetics. *Nucleic Acids Research*, *40*(Database), 790–796. https://doi.org/10.1093/nar/gkr1046

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Wright, B. E., Butler, M. H., & Albe, K. R. (1992). Systems analysis of the tricarboxylic acid cycle in dictyostelium discoideum. I. The basis for model construction. *Journal of Biological Chemistry*, *267*(5), 3101–3105.

Xia, J., Xu, J., Hu, L., & Liu, X. (2016). Enhanced poly(L-malic acid) production from pretreated cane molasses by *Aureobasidium pullulans* in fed-batch fermentation. *Preparative Biochemistry and Biotechnology*, *46*(8), 798–802. https://doi.org/10.1080/10826068.2015.1135464

Yang, Jiangang, Wang, Z., Zhu, N., Wang, B., Chen, T., & Zhao, X. (2014). Metabolic engineering of *Escherichia coli* and in silico comparing of carboxylation pathways for high succinate productivity under aerobic conditions. *Microbiological Research*, *169*(5–6), 432–440. https://doi.org/10.1016/j.micres.2013.09.002

Yang, Junhao, Voloshin, A., Swartz, J. R., Velkeen, H., Levy, R., Michel-Reydellet, N., & Kanter, G. (2005). Rapid expression of vaccine proteins for B-cell lymphoma in a cell-free system. *Biotechnology and Bioengineering*, *89*(5), 503–511. https://doi.org/10.1002/bit.20283

Yang, X., Xu, M., & Yang, S. T. (2015). Metabolic and process engineering of *Clostridium cellulovorans* for biofuel production from cellulose. *Metabolic Engineering*, *32*, 39–48. https://doi.org/10.1016/j.ymben.2015.09.001

Ye, Xiaoting, Honda, K., Morimoto, Y., Okano, K., & Ohtake, H. (2013). Direct conversion of glucose to malate by synthetic metabolic engineering. *Journal of Biotechnology*, *164*(1), 34–40. https://doi.org/10.1016/j.jbiotec.2012.11.011

Ye, Xinhao, Wang, Y., Hopkins, R. C., Adams, M. W. W., Evans, B. R., Mielenz, J. R., & Zhang, Y. H. P. (2009). Spontaneous high-yield production of hydrogen from cellulosic materials and water catalyzed by enzyme cocktails. *ChemSusChem*, *2*, 149–152. https://doi.org/10.1002/cssc.200900017

Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., … Van Dien, S. (2011). Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nature Chemical Biology*, *7*(7), 445–452. https://doi.org/10.1038/nchembio.580

Zambanini, T., Kleineberg, W., Sarikaya, E., Buescher, J. M., Meurer, G., Wierckx, N., & Blank, L. M. (2016). Enhanced malic acid production from glycerol with high - cell density *Ustilago trichophora* TZ1 cultivations. *Biotechnology for Biofuels*, *9*(135), 1–10. https://doi.org/10.1186/s13068-016-0553-7

Zambanini, T., Sarikaya, E., Kleineberg, W., Buescher, J. M., Meurer, G., Wierckx, N., & Blank, L. M. (2016). Efficient Malic Acid Production from Glycerol with *Ustilago trichophora* TZ1. *Biotechnology for Biofuels*, *9*(67), 1245. https://doi.org/10.1186/s13068-016-0483-4

Zampieri, G., Vijayakumar, S., Yaneske, E., & Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. *PLoS Computational Biology*, *15*(7), 1–24. https://doi.org/10.1371/journal.pcbi.1007084

Zelle, R. M., De Hulster, E., Van Winden, W. A., De Waard, P., Dijkema, C., Winkler, A. A., … Van Maris, A. J. A. (2008). Malic acid production by *Saccharomyces cerevisiae*: Engineering of pyruvate carboxylation, oxaloacetate reduction, and malate export. *Applied and Environmental Microbiology*, *74*(9), 2766–2777. https://doi.org/10.1128/AEM.02591-07

Zelle, R. M., Hulster, E. De, Kloezen, W., Pronk, J. T., & Maris, A. J. A. Van. (2010). Key Process Conditions for Production of C_4 Dicarboxylic Acids in Bioreactor Batch Cultures of an Engineered *Saccharomyces cerevisiae* Strain. *Applied and Environmental Microbiology*, *76*(3), 744–750. https://doi.org/10.1128/AEM.02396-09

Zeng, J., Teo, J., Banerjee, A., Chapman, T. W., Kim, J., & Sarpeshkar, R. (2018). A Synthetic Microbial Operational Amplifier. *ACS Synthetic Biology*, *7*(9), 2007–2013. https://doi.org/10.1021/acssynbio.8b00138

Zhang, K., Zhang, B., & Yang, S.-T. (2013). Production of Citric, Itaconic, Fumaric, and Malic Acids in Filamentous Fungal Fermentations. In S. Yang, H. A. El-Enshasy, & N. Thongchul (Eds.), *Bioprocessing Technologies in Biorefinery for Sustainable Production of Fuels, Chemicals, and Polymers* (pp. 375–398). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118642047.ch20

Zhang, T., Ge, C., Li Deng, Tan, T., & Wang, F. (2015). C4-dicarboxylic acid production by overexpressing the reductive TCA pathway. *FEMS Microbiology Letters*, *362*(9), 1–7. https://doi.org/10.1093/femsle/fnv052

Zhang, X., Wang, X., Shanmugam, K. T., & Ingram, L. O. (2011). L-malate production by metabolically engineered *Escherichia coli*. *Applied and Environmental Microbiology*, *77*(2), 427–434. https://doi.org/10.1128/AEM.01971-10

Zhang, Y. H. Percival. (2010). Renewable carbohydrates are a potential high-density hydrogen carrier. *International Journal of Hydrogen Energy*, *35*(19), 10334–10342. https://doi.org/10.1016/j.ijhydene.2010.07.132

Zhang, Y. H. Percival. (2011). Substrate channeling and enzyme complexes for biotechnological applications. *Biotechnology Advances*, *29*(6), 715–725. https://doi.org/10.1016/j.biotechadv.2011.05.020

Zhang, Y. H. Percival. (2013). Next generation biorefineries will solve the food, biofuels, and environmental trilemma in the energy-food-water nexus. *Energy Science and Engineering*, *1*(1), 27–41. https://doi.org/10.1002/ese3.2

Zhang, Y. H. Percival, Sun, J., & Zhong, J. J. (2010). Biofuel production by in vitro synthetic enzymatic pathway biotransformation. *Current Opinion in Biotechnology*, *21*(5), 663–669. https://doi.org/10.1016/j.copbio.2010.05.005

Zhang, Y. H.Percival. (2010). Production of biocommodities and bioelectricity by cell-free

synthetic enzymatic pathway biotransformations: Challenges and opportunities. *Biotechnology and Bioengineering*, *105*(4), 663–667. https://doi.org/10.1002/bit.22630

Zhang, Y., Nielsen, J., & Liu, Z. (2018). Metabolic engineering of *Saccharomyces cerevisiae* for production of fatty acid–derived hydrocarbons. *Biotechnology and Bioengineering*, *115*(9), 2139–2147. https://doi.org/10.1002/bit.26738

Zhang, Yi Heng Percival. (2015). Production of biofuels and biochemicals by in vitro synthetic biosystems: Opportunities and challenges. *Biotechnology Advances*, *33*(7), 1467–1483. https://doi.org/10.1016/j.biotechadv.2014.10.009

Zhang, Yi Heng Percival, Sun, J., & Ma, Y. (2017). Biomanufacturing: history and perspective. *Journal of Industrial Microbiology and Biotechnology*, *44*(4–5), 773–784. https://doi.org/10.1007/s10295-016-1863-2

Zheng, H., Ohno, Y., Nakamori, T., & Suye, S. ichiro. (2009). Production of l-malic acid with fixation of HCO_3^- by malic enzyme-catalyzed reaction based on regeneration of coenzyme on electrode modified by layer-by-layer self-assembly method. *Journal of Bioscience and Bioengineering*, *107*(1), 16–20. https://doi.org/10.1016/j.jbiosc.2008.09.009

Zheng, Y., & Sriram, G. (2010). Mathematical modeling: Bridging the gap between concept and realization in synthetic biology. *Journal of Biomedicine and Biotechnology*, *2010*(Figure 1). https://doi.org/10.1155/2010/541609

Zoglowek, C., Krömer, S., & Heldt, H. W. (1988). Oxaloacetate and malate transport by plant mitochondria. *Plant Physiology*, *87*(1), 109–115. https://doi.org/10.1104/pp.87.1.109

Zou, X., Wang, Y., Tu, G., Zan, Z., & Wu, X. (2015). Adaptation and transcriptome analysis of *Aureobasidium pullulans* in corncob hydrolysate for increased inhibitor tolerance to malic acid production. *PLoS ONE*, *10*(3), 1–17. https://doi.org/10.1371/journal.pone.0121416

UNIVERSITÉ DE LA RÉUNION

Pôle Recherche
Ecoles doctorales

LETTRE D'ENGAGEMENT DE NON-PLAGIAT

Je, soussigné(e) Anamya AJJOLLI NAGARAJA en ma qualité de doctorant(e) de l'Université de La Réunion, déclare être conscient(e) que le plagiat est un acte délictueux passible de sanctions disciplinaires. Aussi, dans le respect de la propriété intellectuelle et du droit d'auteur, je m'engage à systématiquement citer mes sources, quelle qu'en soit la forme (textes, images, audiovisuel, internet), dans le cadre de la rédaction de ma thèse et de toute autre production scientifique, sachant que l'établissement est susceptible de soumettre le texte de ma thèse à un logiciel anti-plagiat.

Fait à Saint-Denis le : 26/06/2020

Signature :

Extrait du Règlement intérieur de l'Université de La Réunion
(validé par le Conseil d'Administration en date du 11 décembre 2014)

Article 9. Protection de la propriété intellectuelle – Faux et usage de faux, contrefaçon, plagiat

L'utilisation des ressources informatiques de l'Université implique le respect de ses droits de propriété intellectuelle ainsi que ceux de ses partenaires et plus généralement, de tous tiers titulaires de tes droits.
En conséquence, chaque utilisateur doit :
- utiliser les logiciels dans les conditions de licences souscrites ;
- ne pas reproduire, copier, diffuser, modifier ou utiliser des logiciels, bases de données, pages Web, textes, images, photographies ou autres créations protégées par le droit d'auteur ou un droit privatif, sans avoir obtenu préalablement l'autorisation des titulaires de ces droits.

La contrefaçon et le faux
Conformément aux dispositions du code de la propriété intellectuelle, toute représentation ou reproduction intégrale ou partielle d'une œuvre de l'esprit faite ans le consentement de son auteur est illicite et constitue un délit pénal.
L'article 444-1 du code pénal dispose : « Constitue un faux toute altération frauduleuse de la vérité, de nature à cause un préjudice et accomplie par quelque moyen que ce soit, dans un écrit ou tout autre support d'expression de la pensée qui a pour objet ou qui peut avoir pour effet d'établir la preuve d'un droit ou d'un fait ayant des conséquences juridiques ».
L'article L335_3 du code de la propriété intellectuelle précise que : « Est également un délit de contrefaçon toute reproduction, représentation ou diffusion, par quelque moyen que ce soit, d'une œuvre de l'esprit en violation des droits de l'auteur, tels qu'ils sont définis et réglementés par la loi. Est également un délit de contrefaçon la violation de l'un des droits de l'auteur d'un logiciel (…) ».

Le plagiat est constitué par la copie, totale ou partielle d'un travail réalisé par autrui, lorsque la source empruntée n'est pas citée, quel que soit le moyen utilisé. Le plagiat constitue une violation du droit d'auteur (au sens des articles L 335-2 et L 335-3 du code de la propriété intellectuelle). Il peut être assimilé à un délit de contrefaçon. C'est aussi une faute disciplinaire, susceptible d'entraîner une sanction.
Les sources et les références utilisées dans le cadre des travaux (préparations, devoirs, mémoires, thèses, rapports de stage…) doivent être clairement citées. Des citations intégrales peuvent figurer dans les documents rendus, si elles sont assorties de leur référence (nom d'auteur, publication, date, éditeur…) et identifiées comme telles par des guillemets ou des italiques.

Les délits de contrefaçon, de plagiat et d'usage de faux peuvent donner lieu à une sanction disciplinaire indépendante de la mise en œuvre de poursuites pénales.