Metabolomics and biosensor approaches to the detection of fever associated diseases

Ana-Maria Năstase

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

University of Glasgow

November 2021

Abstract

Febrile illnesses are still a major cause of mortality and morbidity globally and the failure to detect and correctly diagnose a specific disease associated with fever is partly responsible for this. This thesis aimed to investigate a biosensor-based method for the detection of fever-associated diseases and to further explore the molecular mechanisms and possible biomarkers of febrile illnesses by employing a metabolomics-based approach. The biosensor platform is based on a complementary metal oxide semiconductor technology, which has both technological and economic advantages. Due to the small size of the microchip, accurate signal processing becomes challenging and, thus, computational methods were developed and tested for the quantitative detection of antibodies in a solution tested on the biosensor platform. Three methods, one based on a deterministic approach and two others based on machine learning (ML) algorithms, were tested and compared for the detection of a reaction spot intensity using synthetically generated images. Next, in order to develop an immunoassay protocol for the detection of one specific fever associated infectious disease, human African trypanosomiasis (HAT), several steps were taken. First of all, a suitable and sensitive method of detection was selected, i.e. enzyme linked immunosorbent assay (ELISA). Next, four recombinant antigens currently used for the detection of HAT were selected based on previous evidence and developed using molecular cloning techniques in *E.coli*. These were tested on infected and control humans serum samples obtained from endemic regions of the Democratic Republic of Congo (DRC). Disposable poly-methyl methacrylate (PMMA) slides which were chemically functionalised were used on top of the chip as the immunoassay surface. Titrations for the selected antigens/antibody were tested using an indirect ELISA-like protocol and the best results after fitting a calibration curve were obtained for an antigen concentration of 2.5 µg/ml. The detection of the antibody to the trypanosome antigen invariant surface glycoprotein 65 (ISG65) proved to be successful and the protocol could be replicated for all the other antigens. However, technical challenges and the closure of the laboratory during the Covid-19 pandemic precluded my taking this part of the project to its conclusion. Following this, metabolomics datasets studying disparate febrile infectious illnesses obtained using liquid chromatography coupled to mass spectrometry (LC-MS) were used in order to investigate and detect possible metabolite-based biomarkers common to fever-associated diseases. A warping based method was developed in order to enable integration by alignment of disparate LC-MS metabolomics datasets. Integration was performed by correcting the RT drift between the datasets using fitted Gaussian Process regression models, a supervised ML method, which was followed by direct matching alignment using MZmine2. The correction was performed by using the standard reference mixture (SRM) information. Statistical analysis on the meta-dataset was performed using linear modelling implemented in the limma R-package. Comparison was made between infected and control samples and commonality was established using the fold change values obtained for the individual datasets. Annotation was carried out by matching the compounds against metabomlomics datasets and through mummichog software, which was also used for pathway analysis. The features obtained from this analysis which were putatively annotated were classified into categories (amino acids, sugars, lipids, nucleotides, etc.). Features in common to all datasets were used to make a connection to the previously established molecular basis of fever. Significant changes were identified to several metabolic pathways, with the most notable perturbations being within the kynurenine pathway, a branch of tryptophan metabolism. Also, features specific to each dataset were used to evaluate the accuracy of the fever biomarkers and investigate possible biomarkers for each different fever-associated disease.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank to my supervisors, Prof Michael P. Barrett and Dr Simon Rogers, for their encouragement and patience, and for the guidance and feedback they have provided during this research project. The meetings and conversations were very important and they inspired me to think outside the box and form an objective critique. I would also like to thank to my annual review assessors, Dr Julien Reboud, Dr Richard Burchmore and Prof Roderick Murray-Smith, for their feedback and useful tips they have provided me during the annual reviews of this PhD. I would also like to thank to the members of the Microsystem Technology group (led by Prof David Cumming), Valerio Annese, Chunxiao Hu and Boon Chong Cheah, who provided help in the biosensor related work, to the fellow members of Prof Michael P. Barrett's group, Dharmendra Dheeman and Andreea Stroia, with whom I shared the laboratory space and who provided useful insight in developing the recombinant proteins and to fellow member of the Computational Biology group (led by Dr Simon Rogers), Alex Pancheva, who shared with me useful information about Gaussian Processes.

I would also like to express my gratitude towards the Engineering and Physical Sciences Research Council (EPSRC) and the College of Medical, Veterinary and Life Sciences (MVLS) for the financial support they offered for this work.

Declaration

All of the work in this thesis was carried out by myself unless otherwise explicitly stated. None of this material has been submitted for any other degree at the University of Glasgow or any other institution.

Ana-Maria Năstase

Chapter 1

Introduction

Many infectious diseases are characterised by fever: a generic host response to numerous microbial pathogens. Fever is associated with the hypothalamus which, in response to exogenous pyrogens, activates cyclooxygenase-2 (COX-2) and releases prostaglandin E2 (PGE2), triggering a systemic increase in body temperature which can have microbicidal effects [1]. Although fever generally has a protective effect, acute febrile illnesses are still a major cause of mortality and morbidity globally, particularly in low to middle income countries [2]. The failure to detect and correctly diagnose a specific disease associated with fever is partly responsible for this. Inappropriate treatment of misdiagnosed diseases can contribute to the selection of drug resistant microbes. For example, in many parts of Africa, fever is assumed to be due to malaria and treated with anti-malarial drugs. In cases where the patient may have not been infected with malaria parasites, but subsequently became infected while the drug concentration was waning, a selective pressure on resistant mutants was imposed [3]. Therefore, improved diagnostics of febrile patients and specific biomarker discovery is desirable.

In this PhD, an inter-disciplinary approach was undertaken to investigate the detection of diseases associated with fever, drawing on areas such as molecular biology, computational biology based on machine learning approaches, and engineering. The first part of this thesis focused on developing an immunoassay for the detection of a fever-related disease on a point-of-care platform previously developed as part of the Multicorder project [4, 5]. The initial aim of the Multicorder research project was to develop a complementary metal-oxide semiconductor (CMOS) based biosensor to enable the detection of metabolites such as choline, xanthine, sarcosine and cholesterol through enzyme assays. In [5], the scope of the CMOS-based biosensor was extended to the detection of human immuno-deficiency virus (HIV) specific antibodies through an immunoassay. This part of the PhD used [5] as a reference for the immunoassay previously developed on the biosensor platform and aimed to contribute to the Multicorder programmatic research effort. Based on the results obtained by [5] a machine learning algorithm was first developed for the analysis of the biosensor data. Next, recombinant antigens and a suitable immunoassay were both developed for the biosensor platform. In the second part of the thesis, the mechanism of fever-associated diseases was further investigated using mass spectrometry data. Multiple metabolomics datasets on disparate fever-associated diseases were integrated using a novel algorithm and metabolic biomarkers associated with fever-causing diseases and infectious disease severity were identified.

1.1 Overview of the thesis

The thesis is structured into seven chapters. A brief description of each chapter is outlined below.

Chapter 2 presents the background information, based on relevant literature research, on the main topics addressed in this PhD. The topics included refer to immunosensors, metabolomics analysis overview and concepts of machine learning, in particular of supervised learning.

Chapter 3 compares three computational methods developed for the processing of the immunosensor signal outputs. It also presents an algorithm based on Bayesian inference techniques which was specifically developed for the detection of reaction spots developed on the immunosensor.

Chapter 4 presents the process of developing an immunosensor approach for the detection of one fever-associated disease, Human African Trypanosomiasis (HAT). It also presents the processes of developing recombinant antigens for the detection of HAT and customising the detection platform for running the immunoassays.

Chapter 5 introduces a novel method based on supervised machine learning of integrating multiple disparate metabolomics datasets obtained following mass spectrometry analysis. It also presents the process of identifying common metabolic biomarkers of fever-associated diseases which could help in their diagnosis process.

Chapter 6 presents and discusses the biomarkers previously identified in Chapter 5 and the biochemical pathways relevant to the meta-dataset, as well as to each independent disease, in the context of the pathophysiological mechanisms of fever.

The work described in Chapters 5 and 6 has led to a paper which has been prepared for publication:

1. "Alignment of multiple metabolomics LC-MS datasets from disparate diseases to reveal fever-associated metabolites" [6].

Chapter 7 presents a summary of the work performed in this thesis along with its contributions, and highlights the possible future work directions based on the conducted research.

1.2 Code

All of the code for the analysis presented in this thesis was written using python programming language, unless specified otherwise. Version Python 3.7 was used. The code is available in my github repositories: https://github.com/anamaria-uofg/biosens and https://github.com/anamaria-uofg/mma.

1.3 Figures

All figures and plots were produced by myself either in Python 3.7 or Microsoft Office PowerPoint, unless specified otherwise.

1.4 Abbreviations

Table 1.1 below contains a comprehensive list of the abbreviations or acronyms and their meanings which were used throughout this thesis.

Abbreviation	Meaning
APTES	3-Amino Propyl Tri-Ethoxy-Silane
AuNP	Gold Nanoparicles
CMOS	Complementary Metal Oxide Semiconductor
E.coli	*Escherichia coli*
COX-2	Cyclooxygenase-2
EIC	Extracted Ion Chromatogram
ELISA	Enzyme Linked Immunosorbent Assay
ESI	Electrospray Ionisation
GA	Glutaraldehyde
GP	Gaussian Process
GPR	Gaussian Process Regression
HAT	Human African Trypanosmiasis
HMDB	Human Metabolome Database
HPLC	High Performance Liquid Chromatography
IDO-1	Indoleamine-2,1-dioxygenase
IFN	Interferon
IPTG	Isopropyl β-D-1-Thiogalactopyranoside
ISG65	Invariant Surface Glycoprotein 65
ISG75	Invariant Surface Glycoprotein 75

KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	k-Nearest Neighbours
LB	Lysogen Broth
LED	Light-Emitting Diode
LC-MS	Liquid Chromatography-Mass Spectrometry
logFC	logarithmic Fold Change
LOD	Limit of Detection
LOQ	Limit of Quantitation
m/z	mass-to-charge ratio
MAE	Mean Absolute Error
MC	Monte Carlo
MLP	Multi-Layer Perceptron kernel
ML	Machine Learning
MS2	Tandem Mass spectrometry
MSE	Mean Squared Error
nL1.3	Native Variant Surface glycoprotein LiTat 1.3
nL1.5	Native Variant Surface glycoprotein LiTat 1.5
PD	Photo-diode
pdf	Probability Distribution Function
PGE-2	Prostaglandin E2
PMMA	Poly-Methyl Methacrylate
ppm	parts per million
RBF	Radial Basis Function kernel
rL1.3	Recombinant Variant Surface Glycoprotein LiTat 1.5
rL1.5	Recombinant Variant Surface Glycoprotein LiTat 1.5
RT	Retention Time
SDS-PAGE	Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis
SMC	Sequential Monte Carlo
SRM	Standard Reference Mixture
TIC	Total Ion Current
T.b.gambiense	*Trypanosoma brucei gambiense*
TNF	Tumor Necrosis Factor
VL	Visceral Leishmaniasis

Table 1.1: Abbreviations used often throughout the thesis, listed in alphabetical order.

Chapter 2

Background literature

This chapter aims to introduce the relevant background knowledge for the main topics addressed in this thesis required for investigating the detection of fever associated diseases. The first topic discussed is that of biosensors, as the first aim of this thesis was to develop an immunosensor on a complementary metal oxide semiconductor (CMOS) based platform for the detection of fever-associated diseases. The next topic introduced is that of machine learning, in particular regression, its importance in the analysis of biological data and its utility in this project. Lastly, the final topic which is discussed is metabolomics and its utility in the discovery of molecular mechanisms.

The topics investigated in this thesis, alongside their application researched in the different chapters aim to provide an insight into the possibilities of better detection and diagnosis of fever-associated infectious diseases. Accurate detection of such infectious diseases is important for avoiding pathogen transmission and long-term complications, thus ensuring therapy effectiveness.

As mentioned previously, acute febrile illnesses, including those caused by neglected tropical pathogens, are still a major cause of death, especially in low to middle income countries. Therefore easy to access, simple operation and low-cost portable technology is required in such settings to facilitate the accurate detection, and, thus, the correct treatment of the disease. Such technology is represented by point-of-care (POC) devices, including biosensors [7] which are presented in the next section.

2.1 Biosensors

A biosensor is an analytical device which converts a biochemical recognition event into a measurable signal enabling, thus, rapid diagnosis [8]. The first biosensor, an electrochemical biosensor, was proposed in 1962 and was used for quantifying the glucose concentration directly from a sample [9]. According to its IUPAC definition, a biosensor is comprised of two main components, i.e. the sensing element and the transducing element, i.e. an element which converts one form of energy into another (Figure 2.1) [10]. The sensing element, or the bioreceptor, is represented by molecules which should exhibit specific and selective interactions with the analyte that needs detection (Figure 2.1). The detection of the analyte, also known as the biorecognition event, results in a change in the property of the bioreceptor that is then detected by the transducer and converted into a measurable signal which is proportional to the amount of analyte-bioreceptor interaction (Figure 2.1) [10, 11]. The transduced signal is then processed by complex electronic circuitry and converted from analog to digital. This signal is then further processed through software to generate results understandable by any user [10].

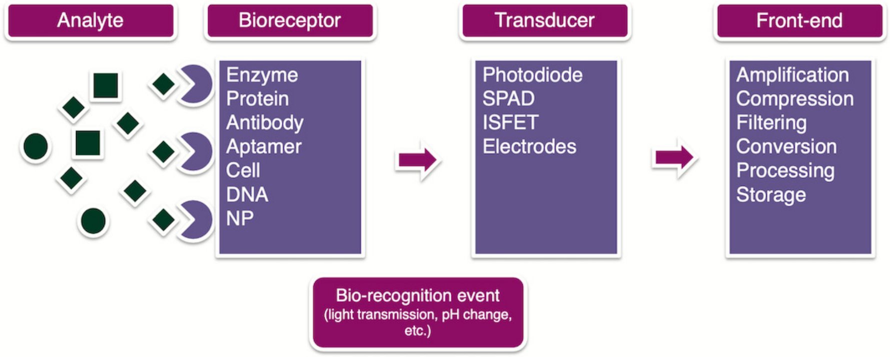


Figure 2.1: General representation of the way biosensors operate: the bioreceptor could be represented by either: DNA molecules, enzymes, antibodies, tissues, microorganisms or cell receptors. Different types of transducers could be used in a biosensor, such as: optical (photodiode, single photon avalanche diode (SPAD)) or electrochemical (ion-sensitive field-effect transistor (ISFET)) sensors. Reproduced and modified from [12].

Depending on their two main components, biosensors can be classified either into tissue-based, enzyme-based, antibody-based (i.e. immunosensors), DNA-based (apta- or geno-sensors) or according to the transducing element into optical or electrochemical sensors [13]. The biosensor used in this project is an optical immunosensor, as it aims to detect antibodies or antigens from a solution using the signal received from photodiodes (PDs) after a change in the solution's colour which causes a change in the light transmittance to the PD.

2.1.1 Biosensors for infectious diseases

Well-established diagnosis techniques for infectious diseases include Enzyme-Linked Immunosorbent Assay (ELISA), microscopy and microorganism culture, and nucleic acid-based assays, with ELISA being one of the most commonly used biochemistry-based assay types which involves the detection of an analyte in a liquid sample. The standard overall process of detection is, however, costly, labour intensive and requires complex sample preparation. Thus, the use of alternative biosensors in this field would offer the possibility of a low-cost portable technology platform that can identify pathogens rapidly and help in establishing the appropriate treatment [14]. Other advantages of using biosensors include the use of small sample volumes, high selectivity and sensitivity and rapid response [15].

The most common types of biosensors used for the detection of infectious diseases are, based on their transducing element, either optical or electrochemical or, based on their bioreceptor, either genosensors or immunosensors [12]. Optical biosensors measure changes in absorbance, fluorescence, chemiluminescence or refractive index resulting from the interaction of the optical field with the receptor. In contrast, the detection in electrochemical biosensors depends on the binding-induced electrical properties of the circuit of which the sensor is an essential component [12]. Since one of the aims of this thesis was to develop an optical immunosensor, these type of biosensors will mostly be discussed next.

An important step in the development of an immunosensor is the immobilisation step, as the antibody-antigen interaction should occur with minimal steric hindrance making the molecular orientation of proteins on the surface very important. Immobilisation methods include adsorption, covalent coupling, antibody-fragment tag, antibody-binding proteins, with the simplest and most commonly used method being adsorption [16]. Most immunosensors are based on the same concept as direct, sandwich or competitive inhibition ELISA assays. In direct immunoassays, the bioreceptor is the antigen from the sample immobilised on the surface and a change in the optical properties of the sensor occurs when conjugated antibodies interact with the immobilised antigens. Sandwich assays, which measure the antigen from a sample by using two layers of antibodies (capture and detection antibodies), are a sensitive method of detection compared to direct assays. Finally, in indirect assays antigens are immobilised on the surface in order to detect specific antibodies in a sample [14].

The efficiency of biosensors is also greatly dependent on the materials used for their fabrication. Nanomaterials are increasingly used as an essential component in the development of biosensors, as they can enhance their optical, electronic or magnetic properties. Gold nanoparticles (AuNPs), for example, are being used increasingly in both optically and electrochemically based biosensor applications. AuNPs may also be used for the immobilization or labelling of biomolecules [17]. For example, functionalisation of antibodies may use AuNPs conjugated to either of the three main available protein groups: $-NH_2$, $-COOH$ and $-SH$. The use of the thiol group and of the C-terminal are both favourable as they prevent the involvement of the labelling particle with the antigen-binding site and, thus, allow antibody-antigen interaction [17].

2.1.1.1 Examples of biosensors developed for detecting fever associated infectious diseases

A selection of successfully developed biosensors for the detection of infectious diseases are presented in this section. For the detection of malaria, both electrochemical and optical biosensors have been developed using histidine-rich protein-II as the bioreceptor [18–22]. In [22] an electrochemical immunosensor was developed using a sandwich ELISA-like assay using a detection antibody conjugated with horseradish peroxidase (HRP) (Figure 2.2). A limit of detection (LOD) of 2.14 ng/mL was obtained for the buffer samples and an LOD of 36 pg/mL when AuNPs conjugate detection antibody-enzyme were used, which demonstrates the benefit of using colloidal gold nanoparticles for the amplification of the sensor's signal and for lowering the detection limit of the target protein.

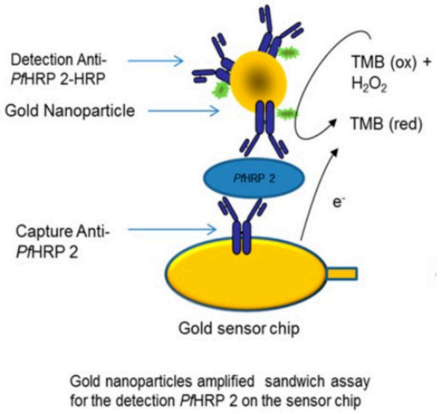


Figure 2.2: Schematic representation of a sandwich-ELISA-like immunosensor which uses AuNPs for signal amplification and HRP as a reporter enzyme. The capture antibody is immobilised by physical adsorption on the surface of the gold working electrodes. The sensor is then washed and the antigen is added. After a second wash, the detection antibody is added. (From [22])

For Leishmaniasis, an optical immunosensor using the surface plasmon resonance (SPR) technique was developed for detecting anti-*L.infantum* antibodies [23]. SPR technology is a label-free method of detection that can monitor the binding of small molecules with high sensitivity. Another reported method of detecting *L.infantum* antigens was by using a piezoelectric immunosensor with antibodies immobilized on a gold surface covered with a thin film of cysteamine and glutaraldehyde [24]. For Zika virus, an electrochemical immunosensor was developed as part of a proof-of-concept work [25]. The bioreceptor used in this case was a monoclonal anti-Zika virus antibody. The immunosensor was based on an interdigitated gold microelectrode array which measured the changes that occur at the interface between the electrodes and electrolytes in the solution. The surface of the immunosensor was modified with a solution which contained esters that react with amines to form amide bonds. For Human African Trypanosomiasis (HAT), an aptasensor using single-walled carbon nanotubes as the transducing element was developed. The recognition element used was protein-specific RNA aptamers used for determining variable surface glycoproteins (VSG) [26]. Since antigen-antibody reactions are more sensitive due to the very specific recognition of the antigen's epitope by the antibody [27], the first aim of the biosensor-related work in the present thesis was to develop recombinant antigens for HAT detection which could be used for an immunoassay on the CMOS-based biosensor platform from the Multicorder project.

2.1.2 Overview on CMOS-biosensors

As previously stated, the method of detection that was used in this PhD for the development of the immunosensor is based on CMOS technology. From an economic point of view, CMOS technology is the most important technology for the fabrication of microelectronic circuits [28]. The CMOS developed for the Multicorder project consists of a 16x16 microelectrode array, with each cluster array having three types of nanophotonic sensors integrated, i.e. single photon avalanche diode (SPAD), ion sensitive field effect transistor (ISFET) and photodiode (PD) (Figure 2.3) [4, 29]. The photodiode is a photo sensitive detector and a single pixel of the CMOS imager is an active pixel sensor. The change in light intensity due to a change in refractive intensity that is measured by the PD is translated to a change in the output voltage. The measurements which were carried out in this project used the PD sensor. As part of the Multicorder project, [5] developed an immunoassay on this CMOS-based platform for the detection of HIV antibodies. This was used as a starting point for the development of the immunosensor described in Chapter 4.

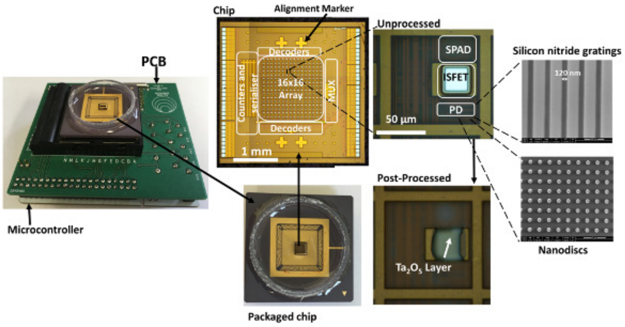


Figure 2.3: The microchip developed in the Multicorder project integrated with the three types of nanosensors including a photodiode coupled to the printer circuit board [29]

.

Other previous studies have also confirmed successful use of CMOS integrated circuits for immunoassays in point-of-care devices [30]. An advantage of using CMOS, rather than other technology such as lateral flow assay, is that CMOS can attain quantitative results. In [31], a CMOS-based biosensor was used to monitor the photon count during the interaction of the HIV antigens and antibodies. The HIV antigen was immobilised on a substrate with Indium nanoparticles of different thicknesses, which proved to influence the binding efficiency between the antigen and antibody. An LOD of 10 fg/mL of HIV antigen was obtained with the CMOS sensor, proving its increased level of sensitivity compared to immunosensors based on different technologies. In another study, an ELISA-like sandwich assay was performed directly on the CMOS chip [32]. The same group successfully developed a multi-analyte CMOS sensor to measure multiple sandwich-ELISA reactions performed on the chip using chemiluminescence [33]. The chemiluminescence was recorded by the integrated photodetector which allowed the detection of different biological targets, such as immunoglobulin E and myoglobin, and showed similar results to the clinical protocols. In terms of its structure, the microchip consisted of 32 photodiodes within 3 μm depth cavities with each diode being coated with chemically functionalised layers for improved antibody protein binding.

2.1.3 Approach to developing the immunoassay for the detection of fever-associated disease

As part of the Multicorder programmatic research effort, the first aim of the thesis consisted of developing an optical ELISA-like immunosensor for the detection of multiple fever associated infectious diseases. The antigens which were immobilised on the sensor surface were developed using molecular cloning techniques. In contrast to [5], surface functionalisation was performed for covalent molecular attachment. Prior to the laboratory work, PD-CMOS biosensor signal processing methods were developed for enhancing the reading of the reaction location and intensity on the chip by using the data already collected from the platform in [5]. Two of the methods were based on supervised machine learning techniques, which are introduced in the next section of this chapter.

2.2 Machine learning overview

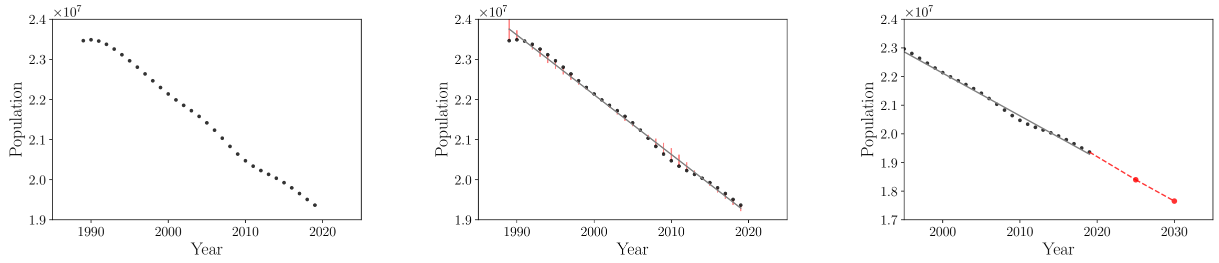
Machine learning (ML) generally aims to identify or *learn* patterns in data by creating *models* using a range of computational algorithms. Based on the type of the process by which the model learns, the ML algorithms can be broadly classified into supervised learning and unsupervised learning. When using supervised learning there is prior knowledge of what the output values for a set of data should be. Thus, for a given samples of data and output values, supervised ML aims to learn a function which best approximates their relationship. Depending on the target data type, i.e. either discrete or continuous variables, the supervised ML algorithms can be classified into classification or regression, respectively [34]. In contrast to supervised ML, unsupervised ML aims to infer the structure within the data without using labelled output values. As such, it can be used for preliminary data exploration. Some algorithms employed for unsupervised ML are clustering, density estimation and visualisation (principal component analysis). The ML related analysis performed in this thesis is based on supervised learning algorithms, mainly on regression algorithms. These will be described in the next section.

2.2.1 Regression models

Mathematical notation For an easier understanding of the concepts explained in this section several mathematical notations were utilised. Scalar variables are notated as x, and vectors and matrices are notated in bold as $\mathbf{x}$ and $\mathbf{X}$, respectively. The superscript T denotes the transpose of a matrix or vector, hence $\mathbf{x}^{\mathrm{T}}$ is a row vector. Conditional probabilities are written as $p(a \mid b)$, which signifies the probability of obtaining a after b has been observed. Lastly, tilde sign $\sim$ signifies that a variable is distributed according to a certain distribution.

As outlined previously, the main aim of regression is to predict a set of continuous target variables. For example, when given a dataset of N attribute scalar variables $\{x_1, x_2, ..., x_N\}$, each described by corresponding target scalar variables $\{t_1, t_2, ..., t_N\}$, the goal of regression is to predict the corresponding target variable (t_{new}) of a new attribute variable (x_{new}). To better explain this, a more practical example is considered. Figure 2.4a shows the population of an European country over the last 30 years. For example, x_1 is approximately 23.5 million for year $t_1 = 1989$. This example uses simple, uni-dimensional attribute and target scalar variables. However, target variables with higher dimensions can also be used to model a relationship; these are normally structured into matrices with N rows and M columns.

The aim of regression in the presented example is to use the data to learn the relationship between the population size (t_n) and year (x_n) and be able to predict the population (t_{new}) in any given year (x_{new}). After making several assumptions with regards to its shape, this relationship could be defined by various mathematical functions with a set of associated parameters. In this case, the relationship can be modelled using a straight line. This is the simplest form of regression, i.e. linear regression, where the relation between the variables can also be described by an equation of the form: $f(x) = t = w_0 + w_1 x$ or $f(x) = t = \mathbf{w}^{\mathrm{T}}\mathbf{x_n}$ if expressed in vectorial form (where $\mathbf{x_n} = \binom{1}{x_n}$ and is a row of $\mathbf{X}$, $\mathbf{w} = \binom{w_0}{w_1}$), or $\mathbf{t} = \mathbf{wX}$ if expressed in matrix form. The values of the parameters of this function, w_0, w_1 determine how good the model is, i.e how close the fitted line is to all of the data point. In order to calculate these, a squared *loss function* can be used which computes the squared difference between the true population size and the population size predicted by the model and after minimisation determines the optimal values for w_0, w_1. In this case, $f(x) = 3.19 \times 10^8 - 1.48 \times 10^5 x$. This function can then be applied to make predictions for future years as illustrated in Figure 2.4c. However, even with optimal parameters, errors can still arise in the model (Figure 2.4b).



(a) **(b)** **(c)**

Figure 2.4: Modelling the population of an European country over the last 30 years using linear regression. a) Scatter plot showing the population size over 30 year. b) Modelling the decrease in population using fitted linear regression. Errors between the true values and the predicted ones are also illustrated. c) Making predictions for year 2025 and 2030. The data was obtained from United Nations - World Population Prospects [35].

Other functions, such as higher order polynomial functions, could be used to model the relationship and minimise these errors. These could lead, however, to over-fitting of the data. Over-fitting of the data refers to using functions which model only the training data very well, but are unable to generalise to other data points [36]. The form of the function $f(x)$, which is determined during the training phase, aims to provide a good generalisation of the whole data and make accurate predictions for new data. Precise predictions are, however, unlikely, and thus it is more useful to predict a range of values rather than a particular one. This is done by using probabilistic regression models.

2.2.2 Probabilistic regression

Basic notions of probabilities will be described firstly in this section. Probabilities can be defined over discrete sets of events or over sets of continuous variables. In the previous example, the sample space is made up of continuous variables. One of the most important probability distribution functions (pdf) is the Gaussian or normal distribution (notation: $N(x \mid \mu, \sigma^2)$). This is defined over a sample space that includes all real numbers and has a probability distribution function for a random variable x:

$$N(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp - \frac{(x-\mu)^2}{2\sigma^2} \tag{2.1}$$

where μ represents the mean and σ^2 the variance of the function. Figure 2.5 shows the pdf of x for various μ and σ^2 values. The width of the distribution is defined by σ^2 and the height by μ about which the distribution is symmetric.

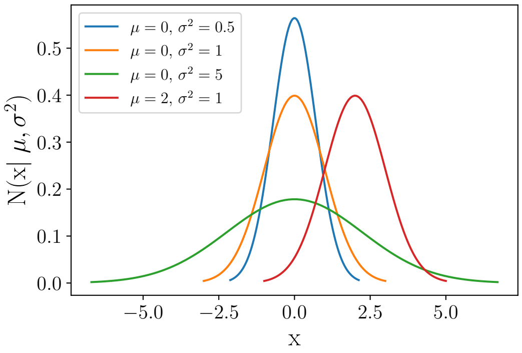


Figure 2.5: The Gaussian distribution represented using different values for mean (μ) and variance (σ^2)

2.2.2.1 Bayes rule

For a better understanding of probabilites, Bayes' rule is presented next. Bayes theorem represents the interaction between probability distribution functions. Bayes' rule (Eq. 2.2) provides a way of updating current knowledge or beliefs about an observation in light of new data. In this case, all of the possible outcomes are described by a probability distribution rather than a single value. In order to explain Bayes' rule, the vectorial form of the linear function will be considered.

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

$$p(\mathbf{w} \mid \mathbf{t}, \mathbf{X}) = \frac{p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}) p(\mathbf{w})}{p(\mathbf{t} \mid \mathbf{x})} \tag{2.2}$$

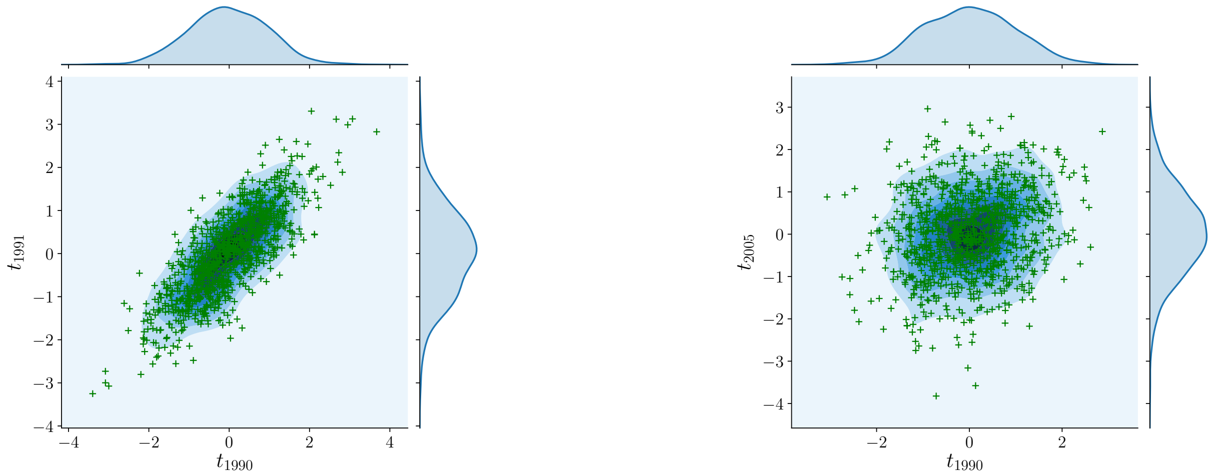
In the Bayesian approach a *prior* belief about the parameters, $p(\mathbf{w})$ needs to be specified, before observing the data $\mathbf{t}$. Bayes' equation tells us how to update what it is known about the distribution over parameters $\mathbf{w}$ from observation $\mathbf{x}_n$ given t_n has been observed, i.e. $p(\mathbf{w} \mid \mathbf{t}, \mathbf{X})$. The main goal of Bayesian inference is, thus, the computation of the posterior probability distribution over the parameters $p(\mathbf{w} \mid \mathbf{t}, \mathbf{X})$. In order to do so, two other distributions need to be computed. The first one is the likelihood distribution $p(\mathbf{t} \mid \mathbf{X}, \mathbf{w})$ which represents how likely it is that for a particular value of $\mathbf{x}_n$ and w_n, t_n would happen. Finding parameters that maximise the likelihood is an important step in ML analysis. The second distribution involved in the equation is the marginal likelihood distribution $p(\mathbf{t} \mid \mathbf{X})$ which is independent of the function parameters and usually acts as a normalising constant. In other words, the posterior probability distribution can be defined as the prior probability distribution which is weighted by the likelihood and, then, re-normalised.

Based on Bayesian inference, generative regression models can be developed. For example, the implementation of a Bayesian framework for regression led to the development of non-parametric supervised machine learning models, such as the Gaussian Process regression models [37].

2.2.2.2 Gaussian Process Regression

Traditionally, parametric models have been used for supervised learning, however they lack flexibility when analysing complex datasets. Such flexible regression models are represented by the non-parametric Gaussian Process (GP) models. In contrast to the parametric Bayesian approach to linear regression where the prior is placed on the parameters ($\mathbf{w}$), in the case of GP regression the prior is placed directly on the values of the function. The dependence in this case is represented between the output values ($\mathbf{t}$) via the covariance function. The covariance function can be regarded as the correlation between two points or a measure of similarity between two vectors. GP regression models also assume that the joint distribution of the function is of Gaussian form.

In order to better explain this, using the population example, only two years were considered. The probabilities $p(t_{1990} \mid \mathbf{x}_{1990})$ and $p(t_{1991} \mid \mathbf{x}_{1991})$ were plotted against each other in a densities graph along with their joint distribution $p([t_{1990}, t_{1991}] \mid \mathbf{x}_{1990}, \mathbf{x}_{1991})$ (Figure 2.6). The mean (μ) of the functions was set to 0, which can be observed in the centre of the ellipse. Because these two functions are assumed to be Gaussian distributed, then any point on this graph (e.g. t_q) can be sampled from a Gaussian distribution with mean vector $\binom{0}{0}$ and a covariance matrix with 2 rows and 1 column (Σ): $t_q \sim N(\mu, \Sigma)$. Based on the shape of the covariance, i.e. the ellipse, it can be said that the correlation between t_{1990} and t_{1991} is quite high and the covariance matrix would look like this: $\Sigma = \left(\begin{smallmatrix} 1 & 0.7 \\ 0.7 & 1 \end{smallmatrix} \right)$ (Figure 2.6 a). If one increases, then it is expected that the other increases as well. In contrast to this, the correlation between t_{2005} and t_{1990} which are more distant points, is lower (e.g. $\Sigma = \left(\begin{smallmatrix} 1 & 0.3 \\ 0.3 & 1 \end{smallmatrix} \right)$) (Figure 2.6 b).



(a)

(b)

Figure 2.6: a) Bi-variate Gaussian distribution of t_{1990} and t_{1991}. The joint distribution $p([t_{1990}, t_{1991}] \mid \mathbf{x}_{1990}, \mathbf{x}_{1991})$ is represented on top of the graph. b) Bi-variate Gaussian distribution of t_{1990} and t_{2005}. The joint distribution $p([t_{1990}, t_{2005}] \mid \mathbf{x}_{1990}, \mathbf{x}_{2005})$ is represented on top of the graph.

A particularity of the Gaussian distribution is that if we slice at any value, we would still see a Gaussian distribution. From the joint distributions presented in Figure 2.6 we would like to obtain the conditional probability of t_{1991} given t_{1990}, $p(t_{1991} \mid t_{1990}, \mathbf{x}_{1990}, \mathbf{x}_{1991})$. This can be obtained through multivariate Gaussian theorem, which, for simplicity purposes, will not be detailed here.

So, in GP regressions, we aim to model the function outputs using a multivariate Gaussian distribution. The prior distribution for this is defined by a mean function and a covariance function (Eq. 2.3). These are then used to compute the elements of the mean vector ($m(\mathbf{x}_n)$) and covariance matrix with N rows and M columns ($k(\mathbf{x}_n, \mathbf{x}_m)$) that define the Gaussian distribution. Usually the mean function is set to 0, which means that the Gaussian mean is a vector of zeros. The covariance matrix, also referred to as a similarity *kernel* has a number of hyper-parameters such as variance and lengthscale which determine the shape of the covariance function [37].

$$t \sim (m(\mathbf{x}_n), k(\mathbf{x}_n, \mathbf{x}_m)). \tag{2.3}$$

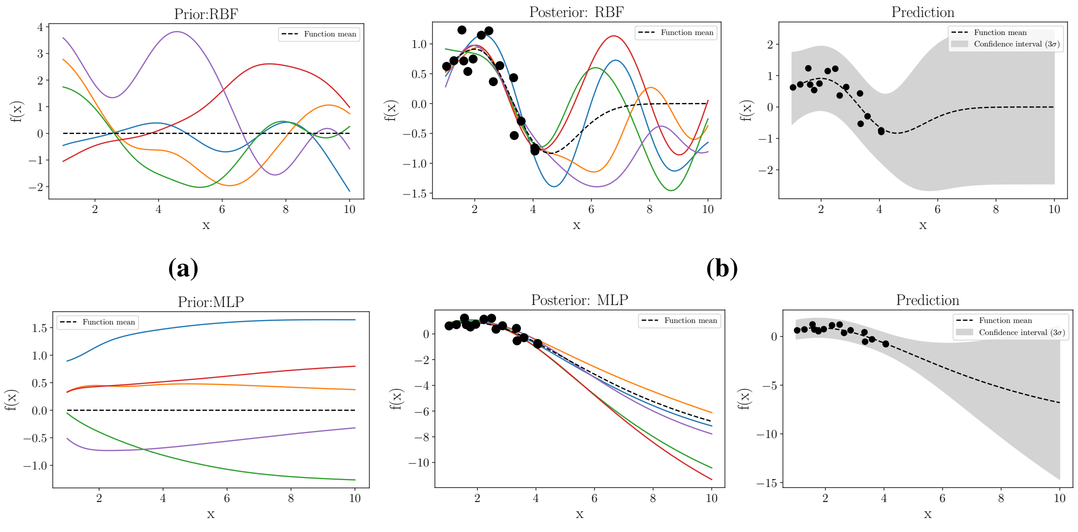
GP regression models can also be constructed to include noise or errors. For instance, in the population example whilst a general downward trend is captured, errors still occur, as illustrated in Figure 2.4 b). In this case, the observed target data also includes the error ε: $t = f(x) + \varepsilon$. For GPs, it is assumed that the error is independently sampled for each target from a Gaussian distribution.

Kernel Types There are several kernels which can be used to model the relationship between the output functions at different input values. The type of kernel used describes the type of correlations between the output functions. Kernels can be stationary and non-stationary. Stationary kernels refer to the fact their value only depends on the difference $(\mathbf{x}_n - \mathbf{x}_m)$ and not on the absolute values of $\mathbf{x}_n$ and $\mathbf{x}_m$ which means that the function prior has the same properties along the x axis, or the same "wobbliness" [36]. On the other hand, for non-stationary kernels, the statistical properties of the function change as $(\mathbf{x}_n - \mathbf{x}_m)$ changes [36]. Therefore, if the function is known to have different characteristics in different regions of the input space, then non-stationary kernels are more suitable.

One of the most used kernels is the squared-exponential or, also known as, radial basis function (RBF) kernel (Figure 2.7). The RBF kernel is stationary, universal and capable of learning any continuous function if given enough training data [38]. The radial basis function has two hyper-parameters: lengthscale and output variance. The lengthscale parameter γ specifies the width of the kernel which determines the smoothness of the functions in the model and the variance parameter α which controls the vertical variation. The difference in different γ values is also illustrated in Figure 2.9 above. The optimal hyper-parameters can be determined through several methods such as: cross-validation, computation of maximal likelihood or Bayesian optimisation [39].

$$k(\mathbf{x}_n, \mathbf{x}_m) = \alpha \exp\left(-\gamma(\mid \mathbf{x}_n - \mathbf{x}_m \mid^2)\right) \tag{2.4}$$

An example of a non-stationary kernel is the neural network kernel which is obtained by marginalising the parameters from a neural network model (Figure 2.7 b) [36].



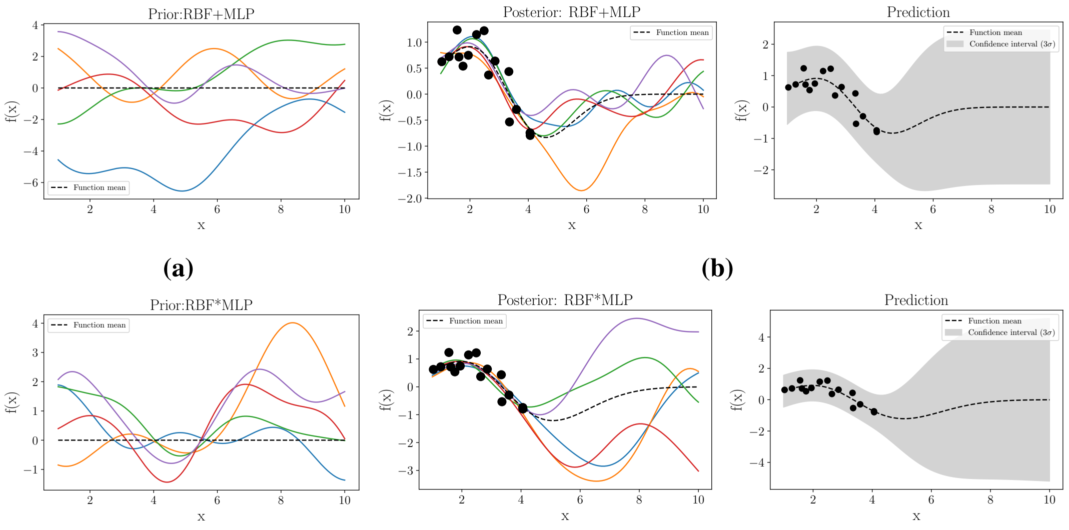
(a) **(b)**

(c) **(d)**

Figure 2.7: Sampling from the prior and posterior distributions of a GP function. In a) and c) it was sampled 5 times from a multivariate Gaussian distribution with mean 0 and kernel ($k(\mathbf{x}_n, \mathbf{x}_m)$). Kernels included were RBF and neural network (MLP). In b) and d) it was sampled 5 times from the posterior distribution after 15 sample points were modelled using GP. The mean of the prediction and the posterior predictive variance ($\mu \pm 3\sigma$) are also illustrated.

Composite Kernels Multiple kernels can also be combined to create new ones with different properties, which allows to incorporate as much high-level structure as necessary into the model. There are two main methods of combining kernels: addition and multiplication (2.5). Sampling from composite kernels was performed in Figure 2.8.

$$\begin{aligned} k_a + k_b &= k_a(\mathbf{x}_n, \mathbf{x}_m) + k_b(\mathbf{x}_n, \mathbf{x}_m) \\ k_a \times k_b &= k_a(\mathbf{x}_n, \mathbf{x}_m) \times k_b(\mathbf{x}_n, \mathbf{x}_m). \end{aligned} \tag{2.5}$$

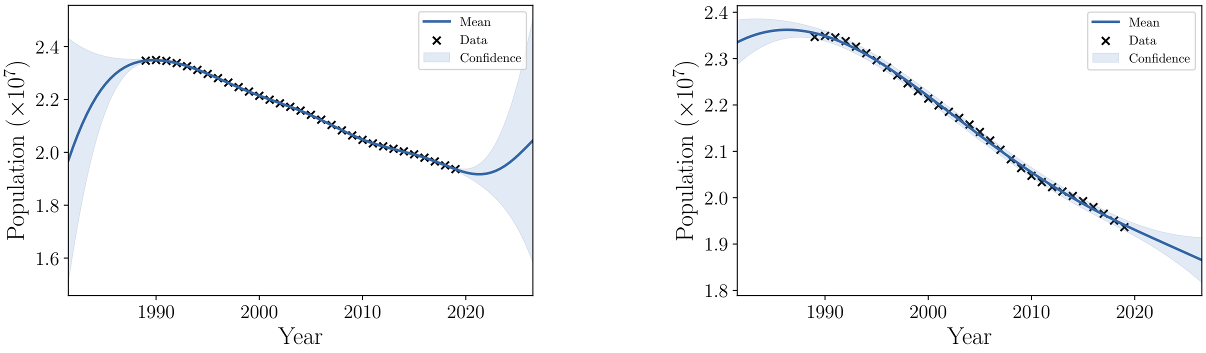


(a) (b)

(c) (d)

Figure 2.8: Sampling from the prior and posterior distributions of a GP function. In a) and c) it was sampled 5 times from a multivariate Gaussian distribution with mean 0 and kernel ($k(\mathbf{x}_n, \mathbf{x}_m)$). Kernels were represented by composite kernels obtained either through addition or multiplication of the RBF and neural network (MLP). In b) and d) it was sampled 5 times from the posterior distribution after the same sample points as in Figure 2.7 were modelled using GP. The mean of the prediction and the posterior predictive variance ($\mu \pm 3\sigma$) are also illustrated.

GP packages There are several frameworks which can be used for conducting GP regressions in different programming languages, such as scikit-learn, GPy [40] or gptk [41]. One of the drawbacks of using GPs is the size of the training data. Since the computation of GPs involves determining the inverse of a matrix, there is an increase in processing time with the increase in training data size. For the programming in Python conducted for this thesis, GPy package was used [40]. Using the population size example, the relationship between the population and year were modelled using GP regression (Figure 2.9). This allows for a posterior predictive range to be defined, extending thus the variability in future predictions.



(a) **(b)**

Figure 2.9: Modelling the population size using GP regression with different hyper-parameter values, specifically the lengthscale (γ) hyper-parameter. After optimisation of the model, $\alpha = 2.2$ and in a) $\gamma = 8.9$ and in b) $\gamma = 34.7$.

2.2.3 ML applications in the biological field

The bigger the size of the data, i.e. larger number of observations, the better the model will be trained to predict the target variable [36]. Due to the increasing data available from biological and clinically-related investigations, machine learning algorithms have become increasingly used in this area.

Machine learning has been increasingly applied in biology-related scientific research. Using the term ’machine learning’ on PubMed, a free search engine accessing primarily scientific journal literature related to biomedical and life sciences, more than 64k results were obtained. The number of ML related journal articles has dramatically increased over the last decade [42]. A large number of topic have been addressed including those of image processing, evolution, text mining and the analysis of -omics big data such as genomics, proteomics and metabolomics which will be discussed in more detail in the next section [43]. All three types of ML algorithms, unsupervised, supervised and semi-supervised learning are used in the field of biomedical sciences. For example, an analysis workflow of a clinical metabolomics study would involve first an unsupervised approach to cluster the data and determine whether the groups correlate with the different disease states investigated. This is usually done through principal component analysis (PCA). Next, using a supervised approach the features most of which determine the separation into the different groups are selected through a process called feature selection. In this thesis concepts of Bayesian inference and regression were applied to the analysis of the data, including that of the PD-CMOS output and the metabolomics data. The next section introduces basic aspects of metabolomics and how such data is processed.

2.3 Metabolomics Overview

Metabolomics is the study of the metabolome which defines the complete set of small molecule chemicals up to 1.5 kDa found within a biological sample such as a cell, an organ, a tissue or a biofluid [44, 45]. The main goal of a metabolomics experiment is to measure the changes in the metabolome of a particular system in response to a change in the system's homeostasis. Starting from the central dogma of molecular biology, it is well known that DNA gets transcribed into different types of RNA, out of which some get translated into proteins, some of which being enzymes that catalyse the generation of metabolites, such as sugars, nucleotides, amino acids and lipids, which make up the biological phenotype of the measured organism (Figure 2.10) [46]. The metabolome, thus, provides a snapshot of the physiology of the cell, tissue, organ or organism being measured. It directly reflects the underlying biochemical activity and state of cells, which is why it has the potential to provide reliable biomarkers for certain diseases [44]. Metabolomics studies have been applied in various research areas such as environmental and biological stress studies, biomarker discovery and integrative systems biology [47].

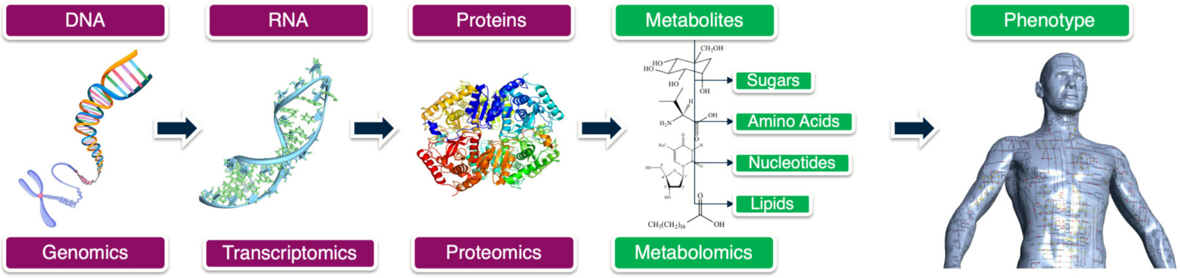


Figure 2.10: The central dogma molecular biology. The DNA is transcribed into several types of RNA. Some of them get translated into proteins which are then degraded into metabolites. The metabolites represent the end product of the cellular processes and they provide a snapshot of the physiology of the cell, tissue, organ and organism measured. The illustrations representing the DNA, RNA, protein crystal structure and the human representation were obtained from the following sources: [48–50].

Metabolomics workflow The main processing steps in a metabolomics study after the samples have been collected and the metabolite extraction performed are the following [51]:

1. Data acquisition through mass spectrometry (MS) coupled with either gas chromatography (GC) or liquid chromatography (LC) or through nuclear magnetic resonance (NMR).

2. Computational data pre-processing: peak detection, peak alignment.

3. Statistical analysis: different statistical or machine learning techniques may be applied at this stage.
4. Metabolite identification (by database search)
5. Metabolite verification (by comparing fragmentation patterns of unknown metabolites with known standards)
6. Pathway and metabolic network analysis

The first two steps related to data acquisition and data pre-processing are described in more detail in the next sections.

2.3.1 LC-MS analysis

The metabolome can be measured either through nuclear magnetic resonance, or through mass spectrometry. This overview of metabolomics, however, focused on describing data obtained from mass spectrometry analysis. The process of mass spectrometry is usually coupled with either GC or LC, with liquid chromatography coupled with mass spectrometry (LC-MS) being the most widely used analytical platform [52]. This section describes the principles of high performance LC-MS analysis, as part of this thesis dealt with the pre-processing of this type of data.

Before the extensive use of LC-MS use, GC-MS systems were preferred. This was due to the incompatibility between the liquid column and the MS, until the development of the electrospray ionisation source (ESI) [53, 54]. ESI works especially well with metabolites, xenobiotics and peptides [53]. The liquid chromatography uses a liquid as the mobile phase to transport the sample molecules through the chromatographic column until they reach the ESI source and undergo a soft ionisation process. In high performance LC, a pressurized liquid and the sample mixture are passed through a column containing an adsorbent, a granular type of material made of solid particles (e.g. silica, polymers) with sizes between 2 µm and 50 µm. This separates the sample components based on their different degree of interaction with the adsorbent particles; the interactions can be hydrophobic, ionic, dipole-dipole or a combination of these. The composition of the pressurised liquid, which is typically a mixture of solvents (water, acetonitrile, methanol, ethanol, etc.), and its temperature also influence the separation process. Once the separated compounds reach the ESI, they get ionised, which allows the creation of compounds with the gain or loss of atoms or molecules. If the resulting analyte has a greater mass than the original molecule, it is called an *adduct*, and if it has a lower mass it is called a *fragment*. ESI sources can be set up either in negative or positive polarisation mode, because, depending on their characteristics, some molecules are more easily ionised in one or another. When operated in positive ESI mode, a proton is added to the analyte $[M+H]^+$, and when operated in negative ESI mode the analyte loses a proton $[M-H]^-$. Adduction or loss of other cations or anions during ESI is also likely during different conditions. For example, when salts are present the adduction of cations such as $[M+NH_4]^+$, $[M+Na]^+$ and $[M+K]^+$ and anions $[M+formate]^-$, $[M+acetate]^-$ to the analytes is possible [53].

2.3.1.1 The fragmentation process

Under normal conditions, the ESI provides a "soft" ionisation source, causing little or no fragmentation to the compounds [53]. However, ions can also purposefully be induced to undergo fragmentation [53]. In general, this process is performed during the LC-MS analysis through collision induced dissociation (CID), when ions collide with inert gases such as nitrogen or argon [53]. During the collision some of the kinetic energy applied to the analyte is converted into internal energy resulting into bond breakage within the analyte. The CID technique which was used to obtain the fragmentation data analysed in this PhD was higher-energy C-trap dissociation (HCD) [55], a method also associated with the Q-Exactive Quadrupole Orbitrap mass spectrometer. The dissociation takes place in the HCD cell at lower energy voltages [55]; the fragmented ions are then stored in one of the mass spectrometer's components called the C-trap, where higher radiofrequency voltage is applied to retain the fragmented ions, which are next injected into the orbitrap for mass analysis [55]. Fragmentation data or tandem mass spectrometry (MS2) data is obtained at the end of this process.

The MS2 information can be acquired either through data-dependent acquisition (DDA) or data-independent acquisition (DIA). For this project, the LC-MS2 data analysed was collected using DDA strategies. During the survey scan, MS automatically selects precursor ions above a selected abundance threshold and only enables their fragmentation. DIA, on the other hand, performs fragmentation on all ions within a given m/z or retention time window [56].

MS2 data aids in the annotation process of the analytes, as molecular structures may be determined from the fragmentation pattern. The fragmentation pattern of small molecules is less straightforward than that of proteins where fragments are made following cleavage of the amino bond. Thus, data bases containing experimental data of tandem mass spectrometry from metabolomics data have been developed in order to facilitate the comparison and annotation process of tandem mass spectrometry data. Examples of generally used databases are HMDB [57], LipidMAPS [58], Massbank [59] and ChemSpider [60]. These consist of either experimentally obtained MS2 spectra or theoretical spectra obtained in an in-silico manner.

2.3.2 LC-MS data pre-processing: Peak detection

At the end of the LC-MS run, a chromatogram is obtained, in which each ion is represented by a peak characterised by its mass-to charge ratio (m/z), retention time (RT) and intensity. This type of raw data is typically quite noisy and it needs to be pre-processed by noise filtering followed by a process called peak detection or peak picking. In the end, a list of ions along with their m/z, RT and intensity is obtained. Examples of open source software which perform these steps are XCMS and MZmine [61, 62]. In MZmine, the software which was used for the analysis in this thesis, the peak detection process is comprised of several steps including: mass detection, chromatogram builder, and chromatogram deconvolution. In the mass detection step, the data for each mass spectrum (MS) scan is converted from profile to centroid data, a process also referred to as binning. During the binning process, the data is converted into pairs of m/z and intensity [62].

After binning, the extracted ion chromatogram (EIC) is then constructed by connecting the m/z values, found within a pre-specified m/z tolerance window, which span across multiple MS scans [62]. Next, chromatographic peaks are detected from the EIC based on a RT range [62]. A candidate peak is represented by multiple m/z data points belonging to the same molecular ion, because of the ^{13}C isotope, or due to multiple charge states being distributed over multiple MS scans as result of chromatographic elution [62, 63]. At this point, the signal-to-noise ratio of the candidate peak is also computed and filtered out if it is lower than the user predefined value [62]. A good chromatographic peak should have a relatively smooth Gaussian shape and the algorithm behind peak detection should be able to distinguish between the actual signal generated from a chemical analyte and irrelevant signals from chemical or electronic noise.

There are several algorithms which perform peak detection. Chromatographic peaks can be detected by directly analysing the local maximum points, matching peaks with the second derivative of the Gaussian function using a fixed window width, or by analysing the EIC's continuous wavelet transform (CWT) coefficients [64]. The CWT algorithm, also implemented by XCMS and MZmine, is the most flexible and frequently used method of detection. The other two aforementioned method have limitations such as overestimation of the number of actual peaks and detection of peaks with fixed width [64].

The peak detection process is dependent on the parameters predefined by the user such as m/z tolerance (measured in parts-per-million (ppm)), signal-to-noise ratio, minimal peak intensity and peak width. For instance, if a peak has a m/z width larger than the predefined range, then the signal generated by one analyte could be split into multiple neighbouring bins, as it drifts between scans [64]. Incorrect parameterisation could also lead to false peaks being detected or to real peaks being missed. For instance, such errors could stem from the set m/z tolerance: a larger m/z tolerance, close m/z traces can merge leading to missing peaks or incorrect quantitative

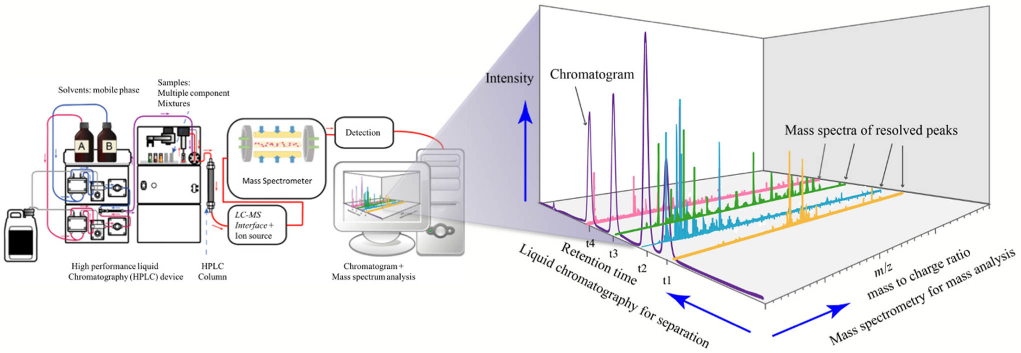


Figure 2.11: Illustration of an LC-MS platform and its data output represented in the form of chromatograms of the multiple sample run on the platform. Each peak is represented in the 3D space by its m/z, RT and intensity. Diagram constructed using figures from [67, 68].

values [65]. The m/z tolerance is also dependent on the LC-MS instrumentation. For example, a range of 5-15 ppm may generate too broad EICs for Orbitrap data [66]. Similarly, if the signal-to-noise threshold is set too high it can also lead to missing peaks. Consistency in peak width is also dependent on the LC-MS acquisition process: high efficiency LC columns, no column saturation, mass overload. Examples of bad peaks include tailing peaks, ghost peaks (column or mobile phase contamination), fronting peaks, split peaks (mobile phase pH too high) [64, 66].

Following the process of peak detection peak alignment is performed. This is described in the next section.

2.3.3 LC-MS data pre-processing: Peak alignment

During an LC-MS experiment, multiple samples are analysed and, in order to compare the peak detection results from all samples, they need to be aligned. After alignment, a list of aligned *peaksets* is obtained. The process of alignment, or correspondence, refers to the mapping of corresponding analytes in any experiment which span across multiple samples. However, this process poses several challenges. During an analytical batch of an LC-MS run, several factors can cause a systemic or component drift in m/z and, mainly, in RT. Some of these factors include: fluctuation in environmental temperature, pressure and humidity, changes in mobile phase pH, chromatographic column condition and running time, sample matrix, ion suppression and even random variation [69].

The already existing correspondence algorithms, which map peaks from one run to another, can be categorised into direct matching and warping algorithms [69]. The warping-based algorithms generally seek to model the RT drift between the different sample runs before the peak detection process. In [69], four different major warping algorithms have been identified as being used in alignment: dynamic time warping (DTW), correlation optimised warping (COW), parametric time warping (PTW) and continuous profile mode (CPM). Alignment tools using this type of warping algorithms are XCMS [61], OpenMS [70] and MZmine 2 RANSAC Aligner [62]. The direct matching method, in contrast, skips the correction of the retention time drift between runs and seeks to map the peaks across the runs directly after the peak detection process has been performed [69]. This method consists of two main stages: computing the feature similarity and using this similarity to map the peaks to an arbitrary selected reference peak list. Feature similarity between two peaks is performed by comparing their m/z and RT by using different measures such as normalised weighted absolute difference [62], cosine similarity [71], Euclidean distance [72] or Mahalanobis distance. Feature matching is then performed through greedy or combinatorial matching methods. Mzmine2 JoinAligner is an example of greedy direct matching method which maps the runs to a "master" peak list based on their m/z and RT similarity scores. Simultaneous multiple alignment (SIMA) is a combinatorial direct matching method which finds a stable matching in a graph produced by joining peaks (nodes) from one run with peaks from another run that are within certain m/z and RT tolerances [73]. As the output of direct-matching methods is the list of aligned peaksets itself, this class of methods can be used as an independent alignment method or as a second-stage process that follows a warping-based method. Once RT drift has been corrected in warping-based methods, it is often easier to establish the actual correspondence of peaks. Seen differently, if a good correspondence between peaks can be established, finding a warping function that maps the retention time from one run to another also becomes easier. The types of alignment algorithms can also be classified into profile-based and feature-based [69]. The profile-based ones perform the alignment before peak detection, as opposed to the feature-based ones where alignment is performed after the peak detection. The feature-based ones normally use reference variables as landmarks to perform retention time drift correction. These can include internal standards normally used in the quality control process of a metabolomics experiment.

2.3.4 Compound annotation

Compound annotation refers to the process of associating an ion to a chemical compound based on its m/z, RT and, when available, MS2 spectra. Annotation in untargeted metabolomics is very challenging, as a variety of atom (C,H,N,O,P,S) configurations can occur for one given molecular mass. For example, isomeric peaks have the same m/z values and the same chemical formula, but different chemical structure. Another example of peaks which could be mistaken for one another are isobaric peaks which have similar m/z values, but different chemical formula entirely. Annotation is typically performed by comparing the m/z values against a list of metabolites from publicly available databases such as the ones mentioned previously, i.e. HMDB [57], KEGG [74] or Massbank [59]. The Metabolomics Standards Initiative (MSI) was used as a guideline for reporting the compounds identified in the metabolomics analysis performed in this PhD [75]. Based on MSI, there are four levels of metabolite identifications: 1) identified compounds (which were obtained by matching the observed peaks against those generated from a set of chemical standards), 2) putatively annotated compounds (obtained by mapping the peaks to publicly available spectral libraries), 3) putatively characterized compound classes (identification only based on the chemical class) and 4) unknown compounds [75].

2.3.5 Conclusion

In conclusion, several methods, both computational and laboratory-based, which will be addressed in the next chapters of this thesis, were employed for improving the detection of fever associated diseases.

Chapter 3

Biosensor signal processing for the quantitative detection of a reaction

3.1 Introduction

The biosensor platform was developed as part of the Multicorder project and it consists of a 16x16 micro-electrode array which amounts to a total surface area of 2.56 mm^2 [4]. Developing a multiplex immunoassay without any additional physically separating components (wells, channels) on a surface area of this size presents several challenges including that of signal processing. In [5] where the same platform was used for running a multiplex immunoassay, the reaction spots were calculated by manually selecting their location using rectangle shapes which encapsulate the spots and computing the mean voltage values of the pixels contained inside the rectangles. This is exemplified in Figure 3.1 [5]. However, this type of approach is prone to introduce several types of biases which could affect the accurate identification of the reaction spot characteristics (intensity, size and location). For example, since the location of the reaction spot is not exactly known, pixels with higher intensity read-outs may be ignored when selecting the reaction area for negative controls or vice-versa, thus affecting the results. Moreover, due to their method of deposition, the reaction spots tend to have a circular shape, rather than a rectangular one. Additionally, various other sources of noise, such as the light source angle, surface of the immunoassay, chip manufacturing issues such as faulty pixels, could all interfere with the processing of the signal received from the pixel array photodiodes (PDs). Therefore, a robust method for determining or inferring the reaction spot characteristics is needed.

An impediment in the experimental analysis using the biosensor platform was constituted by the insufficient provision of chips. Thus, actual experimental data could not be obtained for the analysis performed in this chapter. In order to overcome this, synthetic data was generated for testing and comparing methods of analysis of the biosensor output. The synthetic data represented by a 16 × 16 image array was generated according to the data obtained on the same biosensor platform [5]. The data was obtained following the multiplex detection of HIV antibody gp120, rabbit anti-mouse IgG and their respective negative controls, and the reaction spots were present on four different regions of the image generated from the pixel array PDs (Figure 3.1) [5]. Based on this data, it was assumed that the reaction spot has a circular shape, and, thus, the synthetic data was generated by creating image arrays with one circle characterised by its centre location on the x- and y-axis (x, y), its radius (r) and intensity (i) determined by the PDs.

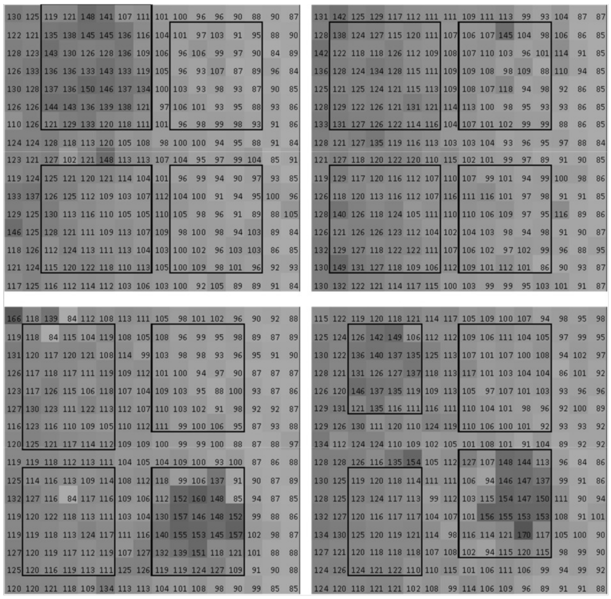


Figure 3.1: Images obtained from the PD-CMOS microchip biosensor by [5] following the multiplex detection of rabbit anti-mouse IgG, HIV antibody gp120 and their respective negative controls. Reaction spots with a circular shape are observed. The four reaction spots are manually selected in [5].

In this chapter three methods were proposed for the quantitative detection of one antigen-antibody reaction on the surface of the biosensor. These were based on either deterministic or stochastic approaches and each present a number of benefits as well as several limitations. Thus, a comparison was performed between the three methods in order to determine for which method the benefits outweigh the limitations making it, thus, more appropriate for the processing of the PDs signal output.

The first method was based on a simple deterministic approach of selecting the first n pixels with the highest intensity values and calculating their mean intensity. This would ensure that any bias caused by the manual selection of the reaction area would be removed. The main advantages of this method are represented by its simplicity and fast processing speed. However this method would not be taking into account the different types of noise which could affect either the surface of the chip or the signal registered by the PDs. Thus, a possible limitation could be characterised by the lack of accuracy in determining the intensity (i) when the signal to noise ratio (S:N) is low. Another limitation could be represented by the fact that the number of pixels for which the mean is calculated needs to be pre-selected, introducing thus a possible bias.

Although numerous machine learning (ML) methods can be tested for this analysis, the final choice was narrowed down to two flexible methods for data analysis. Thus, the second and third methods used a ML approach based on Bayesian inference which was introduced in Chapter 2. By using a Bayesian inference approach it is aimed to provide an unbiased and robust method for estimating the characteristics of the reaction spot. Bayes' theorem (Eq. 3.1) provides the means to update what it is known about an initial belief, which is represented by the distribution over parameters θ given evidence z has been observed, i.e. $p(\theta \mid z)$. All the unknown quantities are described by a distribution rather than a single value and the main goal of Bayesian inference is the computation of the posterior probability $p(\theta \mid z)$. In the present case, the samples from the prior distribution $p(\theta)$ are represented by the characteristics of the reaction spot we are interested in detecting, i.e. its centre coordinates (x,y), its centre radius (r) and intensity (i). Thus, in this case it was aimed to develop a method which accurately detects either all characteristics x,y,r,i or only the intensity, i, given the image z where $p(\theta \mid z) = p((x,y,r,i) \mid z)$.

$$p(\theta \mid z) = \frac{p(z \mid \theta)p(\theta)}{p(z)} = \frac{p(z \mid \theta)p(\theta)}{\int p(\theta)p(z \mid \theta)d\theta} \tag{3.1}$$

The first ML method which was tested in this chapter is based on a regression model that implements a Bayesian framework, i.e. Gaussian Process regression which was also introduced in Chapter 2. This was selected due to its flexibility and applicability to smaller datasets, as it performs well with little training data. In this case, a set of synthetically generated images was used as training data in order to build a regression model which predicts the characteristics of a reaction spot, focusing on the detection of the intensity (i). A limitation of this approach is represented by the need of training data, i.e. reactions performed on multiple different chip surfaces measured on the biosensor platform, and the difficulties in obtaining it. For this analysis, this limitation was bypassed by using synthetic data. This approach could also perform worse when the spot locations are different, unless the training data covers a multitude of spot locations.

The second ML method is based on Bayesian generative modelling which bypasses the need for training data. Since there is no mathematical relation linking the prior samples -the distribution over (x,y,r,i)- with the image (z), it is not possible to analytically obtain the posterior distribution of the parameters given a particular image ($p((x,y,r,i) \mid z)$). This can be instead solved through a stochastic Bayesian computation approach. Stochastic or simulation methods are based on obtaining random samples (x,y,r,i) from the distribution $p((x,y,r,i) \mid z)$ in order to obtain the posterior distribution, which, in this case, refers to the spot characteristics. Such techniques, based on random sampling, are generally known as Monte Carlo (MC) techniques. The MC approach mainly relies on generating random samples from a distribution to make an inference. MC is useful for Bayesian models which involve more than one unknown parameter, as is the case in the current situation. A flexible and easy to implement Monte Carlo approach is Sequential Monte Carlo (SMC). Sequential Monte Carlo (SMC) methods, also known as particle filtering (PF) or importance sampling methods, are generally known from their applications in robotics, particularly in tracking moving objects. However, they can also be used in simpler static settings where there is no known mathematical function linking the input and output.

The final method is, thus, a SMC generative modelling approach which aims to obtain the posterior distribution, i.e. to estimate the characteristics of interest of the reaction spot. The general algorithm starts by sampling a large set of particles from the prior distribution. In this case, the term particle is used to denote a sample drawn from the prior distribution, i.e. a set of randomly generated x, y, r, i parameters. SMC uses the prior distribution directly to draw its first samples from. With each iteration, the particles are weighted, by comparing the synthetic image generated from them with the image which is being analysed. A new sample population from the prior distribution is generated by selecting the particles with the highest likelihood probability and applying a random modification to them such as a Gaussian density. As this resampling process is repeated, the filter approaches the optimal Bayesian estimate. In the end, a sample similar to the one which would be obtained from the posterior distribution is obtained, leading thus to an estimate of the spot characteristics x, y, r, i, and, for this analysis, of i, in particular. This method provides a robust approach of detecting spots in any location even at low S:N. It also benefits of the fact that it requires no training data. One limitation, however, could be related to a possible slow processing speed.

In conclusion, three methods were developed and tested for the signal processing obtained from the pixel array PDs. Additionally, the analysis in this chapter aimed to compare the three methods described above and to determine which performs better under various noise conditions.

3.2 Method

3.2.1 Artificial images

It was assumed the image generated by the PDs array is a 16 × 16 array, each position in the array representing one pixel of the active pixel surface area. It was also assumed that the reaction spot is a circle. Hence, in order to synthetically generate an image array with one reaction spot, a Python object (*Circle*) was created. The object's main attributes are the x and y coordinates of the circle's centre (x, y), the radius length (r) and intensity (i) of the circle. The class also contains the function for generating the synthetic image (generate_circle_image) and for adding noise to it (add_noise_image) while assuming, in first instance, that the noise is described by a Gaussian distribution with zero mean and variance σ_{noise}. The circle image generation is based on the Pythagorean theorem when characterising the radius of a circle based on its centre's coordinates. If (x_1, y_1) is inside the circle, then $x_1^2 + y_1^2 \leq r^2$. Taking into account that one pixel is represented by one point of the array, each pixel is looped over for n number of steps S_{total} and the proportion of the circle present in each pixel $S_{proportion}$ is calculated. Thus, the following equation (3.2) is used for attributing the intensity for each pixel.

$$I_{pixel} = \frac{I_{circle}}{S_{total}} \times S_{proportion} + value_{baseline} \tag{3.2}$$

3.2.2 Signal processing methods

3.2.2.1 Method 1: Deterministic approach

This method was similar to the one used by [5], but instead of manually selecting the spot location and calculating the mean value of the pixels inside the rectangle, the pixels with the highest values were selected. The array representing the image was sorted in descending order and the first n=28 values were selected. This particular number of pixel values were selected as it was approximate of the surface area of a circle with radius r=3. In order to evaluate this method, the mean absolute error was calculated (MAE), i.e. the absolute difference between the mean of the obtained spot intensity values and the actual spot intensity. This method was also referred to as Max method.

3.2.2.2 Method 2: GP regression

For this method, a set of synthetic images had to be generated to be used as a training dataset for the regression model. For performing the GP regression GPy Python package was used [40]. A set of n=100 synthetic image arrays was generated by using randomly allocated $i \in [1,25]$ and $\sigma_{noise} \in [1,20]$ values and constant x, y, r values. The input variables were the image arrays and the target variable was the intensity. For comparing the accuracy of the methods in terms of the spot location identification a second set was generated by using randomly allocated x, y and r values with constant signal (i) to noise (σ_{noise}) ratio (S:N).

Model evaluation The model was evaluated using functions included in Python package scikit-learn for calculating accuracy, mean absolute error (MAE) and mean squared error (MSE). Accuracy refers to the amount of correctly predicted data points out of all the data points, MAE reflects the magnitude of error between the predicted and true value and MSE represents the squared difference between the predicted and true value of each data point.

3.2.2.3 Method 3: SMC generative modelling

The steps of the SMC based algorithm are represented in the diagram in Figure 3.2 and explained in more detail below.

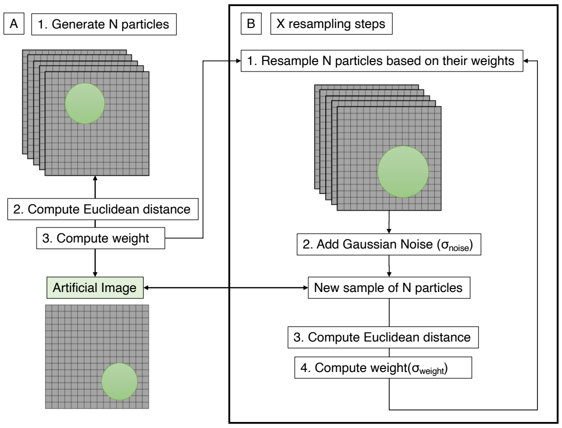

A
1. Generate N particles
2. Compute Euclidean distance
3. Compute weight
Artificial Image
B
X resampling steps
1. Resample N particles based on their weights
2. Add Gaussian Noise (σ_{noise})
New sample of N particles
3. Compute Euclidean distance
4. Compute weight(σ_{weight})

Figure 3.2: Diagram representing the steps taken in the SMC algorithm. A. N random particles are generated and each is compared with the artificial image (synthetically generated image). Based on their Euclidean distance to the image, each particle is attributed a weight. B. Based on the computed weights N equally-weighted particles are resampled from the original sample. These are then modified by adding Gaussian noise to it, creating thus a new sample of N particles which are again compared with the artificial image by calculating their weights. Resampling is performed X times.

1. Initialisation step: Generate N random particles, i.e. the samples for the prior $p(\theta)$.

$$For\, i = \{1, 2, ..., N\}\; sample\, \theta_i \sim p(\theta). \tag{3.3}$$

A particle is defined by (x, y, r, i) where $x \in [0,16]$, $y \in [0,16]$ represent the circle's centre coordinates, $r \in [0,8]$ the radius length and $i \in [0,80]$ the intensity for which a 16×16 array image is defined.

2. Importance sampling step: Compute the weight of each particle.

$$For\, i = \{1,2,...,N\}\; w_i = p(z \mid \theta_i). \tag{3.4}$$

The weight of each particle (w^i) is determined by the Euclidean distance between the synthetically generated image (q) and the particle image (p) (Eq.3.5, 3.6). It is computed based on the exponential term of the function of a normal distribution.

$$d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{3.5}$$

$$w_i = -0.5 \times \frac{d(q,p)^2}{\sigma_{weight}{}^2} \tag{3.6}$$

At each resampling step the highest weight w_{max} is subtracted from the weight of each particle, exponentialised and normalised so that $\sum_{i=1}^{N} w_i = 1$ (Eq.3.7). This is done in order to make sure that at each resampling step there is at least one weight $\neq 0$ before normalisation.

$$w_i = \frac{e^{(w_i - w_{max})}}{\sum_{n=1}^{N} w_n} \tag{3.7}$$

3. Selection and modification step: Resample with replacement N particles based on the importance weights which were slightly modified.

At this step the particles with the highest probabilities w_i are more likely to be resampled. Particles (n=N) with probability w_i were sampled from the original sample of particles and modified by adding noise described by a Gaussian distribution for which the variance is described by σ_{noise} (add_noise_to_characteristics). Steps 2 and 3 are then repeated X times. After the X resampling steps, only those particles which are consistent with the measurement survive.

Model evaluation The aim of the SMC algorithm is to obtain a distribution similar to that of the actual posterior distribution. In order to determine the convergence between the two distributions the mean and variance of the obtained distribution were calculated at end of each resampling step. Various values were tested for N, the number of randomly generated particles, X, the resampling steps, and σ_{noise}. At each resampling step the mean and variance of x, y, r, i were computed and plotted to determine convergence with the coordinates of the artificial image. Thus, the algorithm was tested for different particle sample sizes, $N \in [125, 250, 500]$ for $X = 1000$ resampling steps and constant S:N. After the suitable parameters were selected, the algorithm was tested for convergence at decreasing S:N values (50:1, 5:1, 2:1, 0.5:1). After the final resampling step, the MAE and variance obtained was used for comparison with the other methods.

3.2.3 Experiments: Comparing the three methods

Two experiments were performed for comparing the methods and determine which method obtains the characteristics closest to those of the image being analysed. The first experiment aimed to determine which method performs best in determining or estimating the intensity i. The second experiment was done to determine which ML method best identifies the location (x,y) and size (r) of the circle.

3.2.3.1 Identifying the spot intensity

The three methods were tested in triplicate on the same artificially generated image with decreasing S:N (50:1, 5:1, 2:1, 0.5:1). The image had the same reaction spot size $(r = 3)$ and location $(x = 4, y = 6)$. The results were evaluated based on the intensity deviation or MAE obtained after running the three methods on the images. MAE represents the absolute difference between the actual spot intensity and the spot intensities obtained from each method. For the first method MAE =| mean of the first n pixels - i|, for the GP method MAE = |mean of the GP model distribution output - i| and for the SMC method MAE=|mean of the distribution from the last resampling step - i|. The mean and standard variance of the three MAE of each method were computed and compared against each other.

3.2.3.2 Identifying the spot location and size

The GP method and SMC method were tested in triplicate on three image arrays with the same S:N = 2:1 and different spot locations and dimensions $((x = 4, y = 6, r = 3)$, $(x = 6, y = 12, r = 4)$, $(x = 12, y = 4, r = 2))$. The difference between the actual spot location and size and the spot location and size obtained from the two methods $(\Delta x, \Delta y, \Delta r)$ were calculated. The mean after running all replicates for each method were computed and compared.

3.2.4 Code

All code was written in Python programming language and it can be found in the publicly available Github repository https://github.com/anamaria-uofg/biosens.

3.3 Results

3.3.1 Signal processing methods

3.3.1.1 Deterministic approach

The results from finding the spot based on the pixels with maximum intensity are presented below. The mean of the first n=28 highest values from the image array were selected from the artificially generated images and are presented in Table 3.1. The mean of the first n=28 pixels with the highest value were computed for each image. The results after 100 images is presented in Table 3.1. Both the MAE and standard deviation are increasing with decreasing S:N. This signifies that this method is performing less accurately and less robustly with increasing noise.

S:N	50:1	5:1	2:1	0.5:1
MAE	0.78	0.78	0.72	0.92
STDEV	0.09	0.27	0.37	0.58

Table 3.1: The results for decreasing signal:noise ratios when using the deterministic approach on 100 synthetically generated images.

3.3.1.2 GP regression

The multilayer perceptron (MLP) kernel worked optimally for fitting regression model on a training dataset of 100 synthetically generated images. Following cross-validation where 70% data was split into training data and 30% in testing data, an accuracy score of 0.99 was obtained with a mean accuracy error of 0.52 and the mean squared error of 0.56 (Table 3.2). For spot location prediction, however, the accuracy was much lower and the error magnitude higher (Table 3.2).

	Accuracy	MAE	MSE
Intensity prediction	0.991	0.521	0.559
Location prediction	0.46	2.65	14.46

Table 3.2: The results for decreasing signal:noise ratios when using the GP regression approach on 100 synthetically generated images.

3.3.1.3 SMC algorithm

Convergence The algorithm was initially run with N=125 particles and X=1000 resampling steps (Figure B.1). Results for the mean of the particles, represented by their x,y circle centre coordinates (pixel location), radius and intensity, at each resampling step are represented. Convergence with the values of the synthetically generated image are reached within the first 50 resampling steps for the x, y circle coordinates. However, it takes more resampling steps, 400, to reach convergence in the case of the radius and intensity of the circle. The process gets quite stable quickly for the estimation of the circle centre coordinates and the radius for all three sample sizes. However, the intensity determination is less stable. The results obtained for N=250 particles and N=500 particles are presented in Figure B.2 and Figure B.3, respectively. Intensity reaches the actual value within 300 resampling steps with N=500, as opposed to the other N values, for which it takes approximately 400 resampling steps for intensity to reach the actual value. Thus, in the next section N=500 particles were used for running SMC.

Analysing the algorithm efficiency: reducing the signal-noise ratio The signal-noise ratio was gradually decreased from 50:1 to 0.5:1. With decreasing signal:noise ratios, the spot becomes less visible to the eye as illustrated in Figure 3.6 (j), but its location can still be detected by the algorithm. The images and the algorithm results are illustrated in Figure 3.6. The results of analysing the efficiency of the algorithm are presented in the figures below. As the noise variance increases and the spot intensity decreases, the number of resampling steps it takes to reach convergence increases, especially for the circle radius and intensity. Additionally, for the 5:1 signal to noise ratio, there seems to be a slight over-estimation of the circle radius and intensity (Figure 3.4). For the lower 2:1 and 0.5:1 signal-noise ratios the variance was quite high for the first resampling step, but it substantially decreased in the subsequent resampling steps (Figure 3.5). At an S:N of 0.5:1, there is a slight over-estimation of the radius of the circle, which seems to be compensated by an under-estimation of the circle intensity.

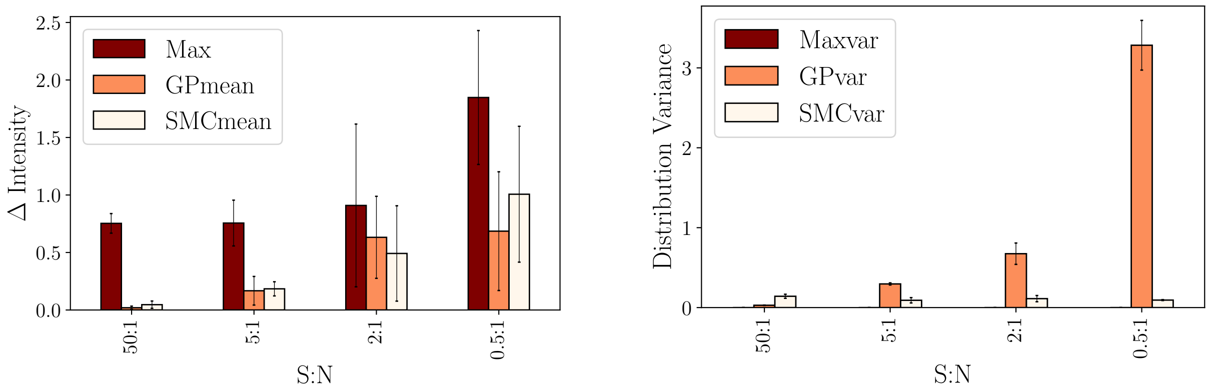
3.3.2 Comparison between the three approaches

3.3.2.1 Intensity detection

The comparison between the three approaches at decreasing S:N are presented in Table 3.3 and Figure 3.3. Based on the values obtained, the deterministic method results into the highest intensity deviations ranging from 0.7 to 2.0 which increases with higher noise levels. Also, at S:N = 0.5:1, the max method does not identify correctly the pixels inside the spot as not all are located inside the spot, but distributed across the image (Figure 3.6). Overall, in terms of determining the spot intensity, the SMC and GPR perform better (Figure 3.3). At the lowest S:N, GP performs better than SMC, based on the intensity deviation results. However, when taking into consideration the results from the variance of the posterior distributions from the ML methods, GP has a very high variance in comparison to SMC. This means that the results obtained from SMC in estimating the intensity are more robust.

S:N	50:1	5:01	2:01	0.5:1
Max_mean	0.753 ± 0.085	0.756 ± 0.199	0.909 ± 0.707	1.848 ± 0.582
GP_mean	0.020 ± 0.014	0.167 ± 0.124	0.632 ± 0.357	0.686 ± .516
GP_var	0.029 ± 0.002	0.296 ± 0.013	0.674 ± 0.134	3.283 ± 0.311
SMC_mean	0.046 ± 0.032	0.184 ± 0.061	0.492 ± 0.415	1.007 ± 0.591
SMC_var	0.141 ± 0.025	0.091 ± 0.033	0.112 ± 0.039	0.094 ± 0.006

Table 3.3: Differences in the detected spot intensity when using different detection methods. Decreasing signal:noise ratios where used. The results are reported as the mean of the tested triplicates alongside the standard deviation (±).



(a)

(b)

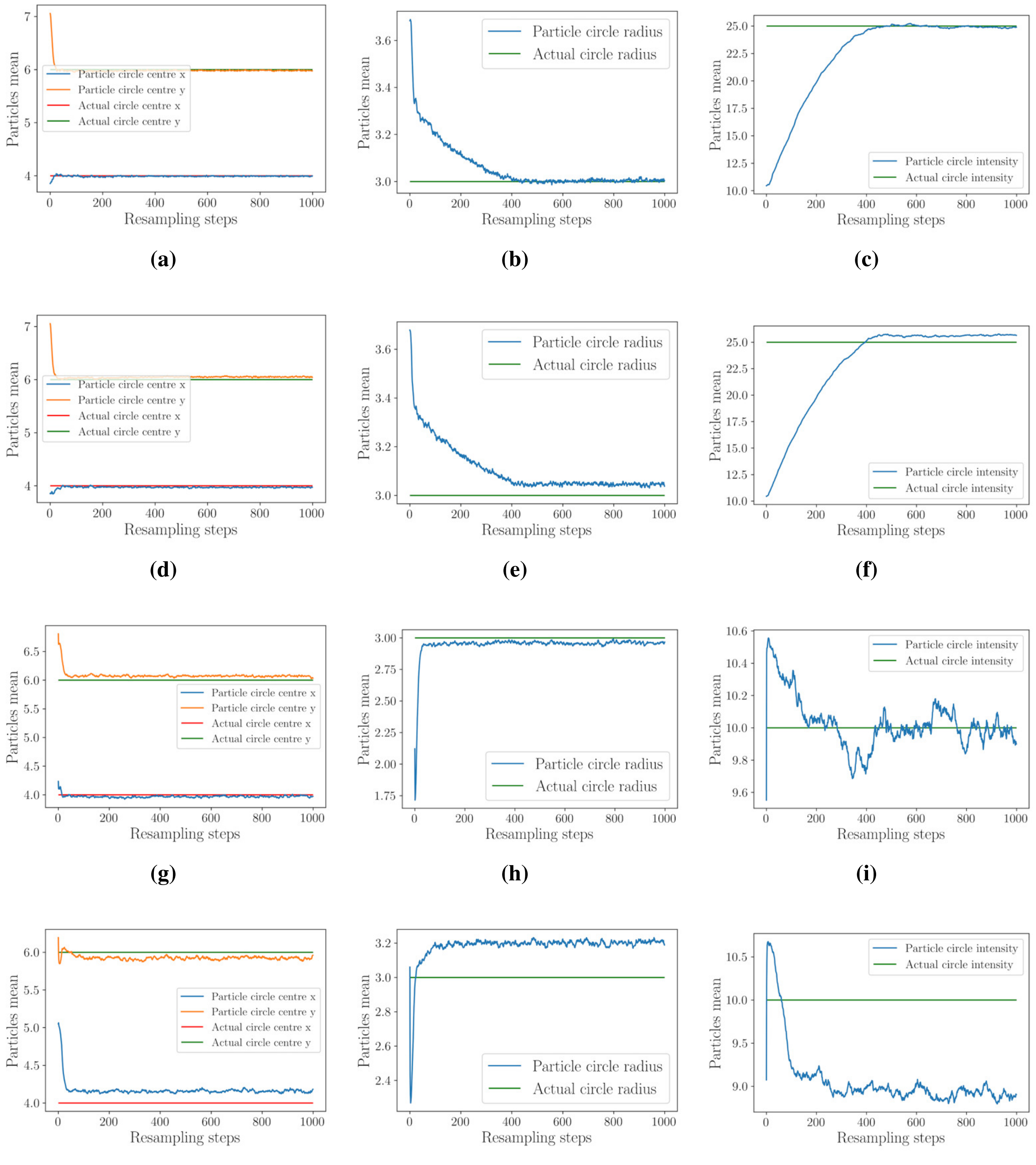
Figure 3.3: a) Comparison between the three methods at different S:N based on MAE (Δ Intensity). The error bars represent the standard deviation from the mean obtained from the measured triplicates. b) The intensity distribution variance of the image obtained following each method is also presented.

3.3.2.2 Location and size detection

GPR and SMC methods were compared against each other for determining which performs better at detecting different spot locations. In this case, SMC performs better than GPR at detecting both the size (r) and the spot location (x, y), as it results into lower MAE and STDEV (Table 3.4).

	x	y	r
		Spot 1	
GPR Mean	1.382 ± 0.323	0.479 ± 0.327	0.617 ± 0.279
SMC Mean	0.026 ± 0.008	0.049 ± 0.040	0.060 ± 0.042
		Spot 2	
GPR Mean	0.623 ± 0.711	0.813 ± 0.873	0.473 ± 0.443
SMC Mean	0.037 ± 0.023	0.185 ± 0.091	0.034 ± 0.009
		Spot 3	
GPR Mean	3.195 ± 0.686	2.708 ± 0.407	0.662 ± 0.074
SMC Mean	0.100 ± 0.127	0.052 ± 0.045	0.157 ± 0.024

Table 3.4: Detecting the spot size and location using GPR and SMC method. The methods were applied to three image arrays for which the reaction spots had different size and locations.



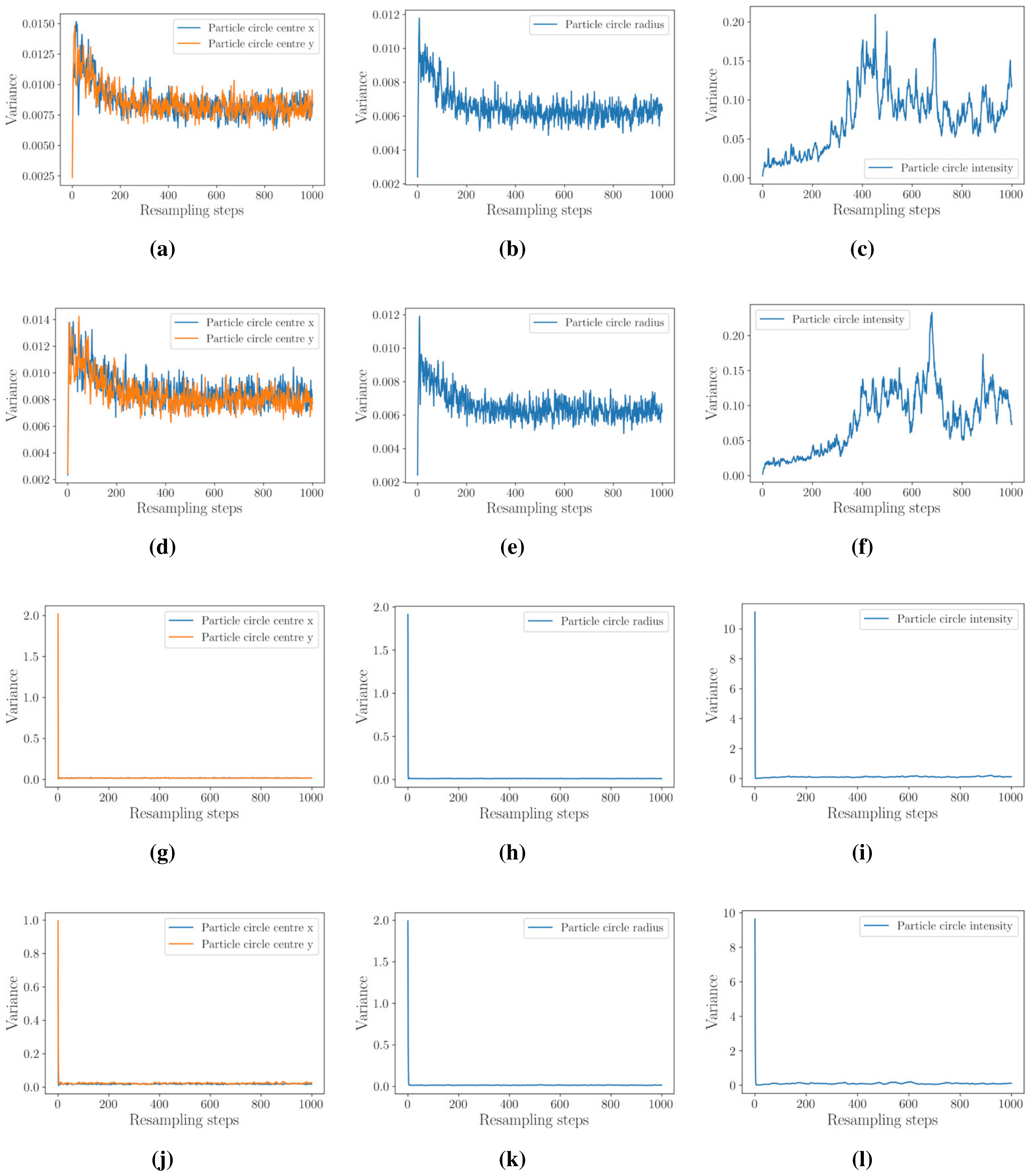
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 3.4: Results for the particle filtering algorithm for N = 500 particle, X = 1000 resampling steps. The intensity of circle was kept at i=25 and σ_{noise} of the image of interest was increased from 0.05 (a-c), to 0.5 (d-f) and 5 (g-i). The intensity was decreased to i=10 and σ_{noise} kept at 5. The mean of the particles at each resampling step were computed for the x,y circle centre coordinates (pixel location), radius and intensity.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

(j)

(k)

(l)

Figure 3.5: Results for the particle filtering algorithm for N = 500 particle, X = 1000 resampling steps. The intensity of circle was kept at i=25 and σ_{noise} of the image of interest was increased from 0.05 (a-c), to 0.5 (d-f) and 5 (g-i). The intensity was decreased to i=10 and σ_{noise} kept at 5 (j-l). The variance of the particles at each resampling step were computed for the x,y circle centre coordinates (pixel location), radius and intensity.

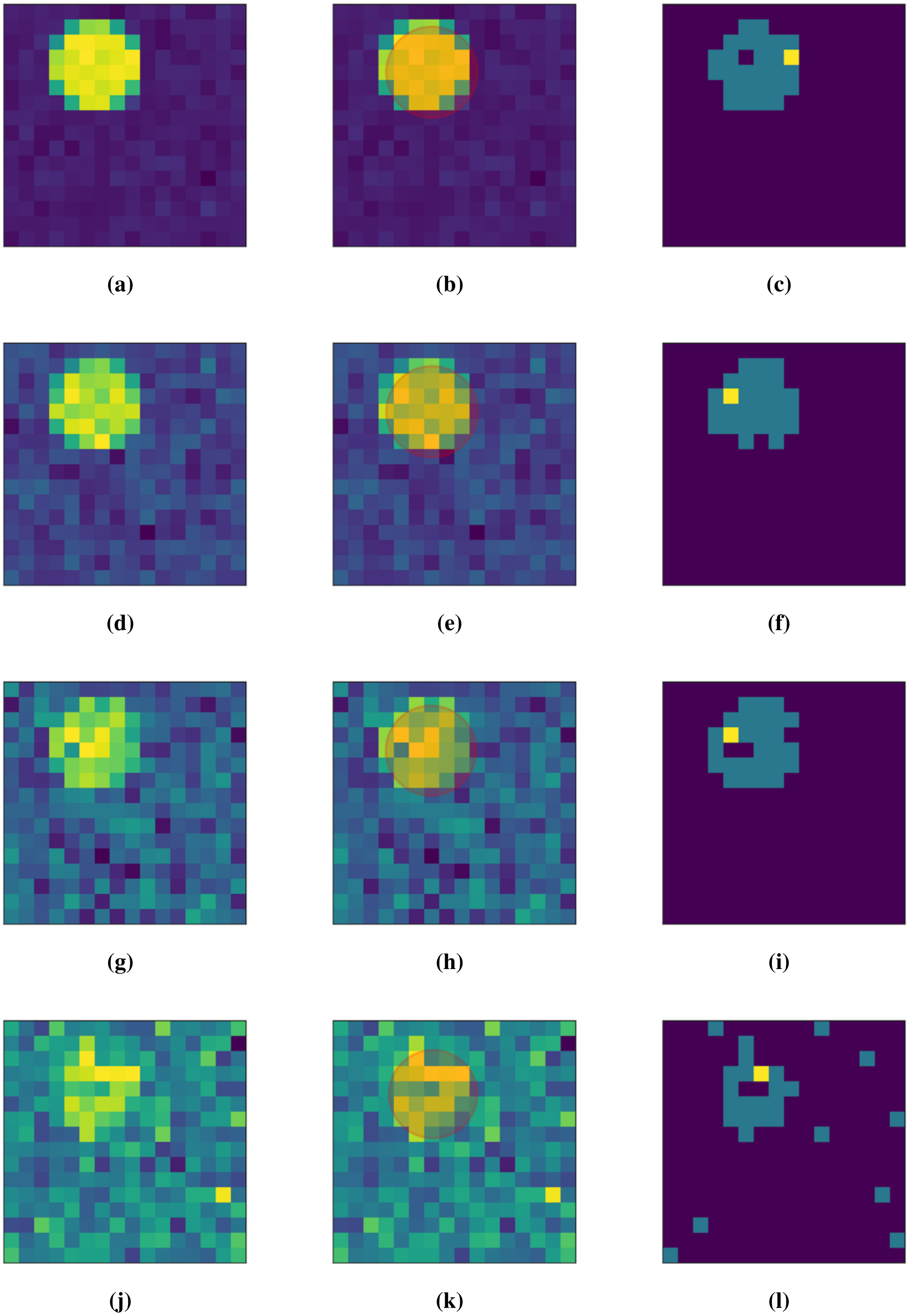


Figure 3.6: The synthetically generated images with $\sigma_{noise} \in 0.05, 0.5, 5$ and $i \in 25, 10$ (a,d,g,j) and the location of the circle as estimated by the SMC algorithm (b,e,h,k) and as obtained by the deterministic approach (c,f,i,l).

3.4 Discussion and conclusion

Based on the results obtained from comparing the three methods, the deterministic approach results in the highest intensity deviation from the actual intensity value. At the lowest S:N, based on the mean Δi results, the GPR method performed best in estimating the intensity, followed by SMC. However, the SMC intensity detection was more robust, as reflected by the lower variance.

The main advantage of using SMC over GPR is that no training data is required for building a prediction model. In the case of GPR, 100 image arrays were synthetically generated for training the model. This number was selected, as it was thought that this small sample size would be possible to be obtained in the laboratory. The size and location of the spot were hard coded in the algorithm and they were all the same for the 100 images; only the intensity and noise variance added to the images varied. In contrast to this, no training data is required for SMC, as the algorithm automatically generates random images and samples the ones closest to the image of interest. Thus, detecting spots of different sizes, location or intensities would not pose a problem to SMC. The SMC approach can also identify the correct location and intensity of a circle of varying intensities in a 16 × 16 image array with increasing levels of added Gaussian noise. Using the SMC approach in this case is useful because there is no probabilistic model to define the mathematical relationship between the model input and output. The deterministic approach, also, does not require any training data and, in contrast to SMC, is computationally very fast. However, the intensity deviations are higher than for the other two methods, increasing with decreasing S:N, and at low S:N this approach does not locate the spot accurately. Additionally, it would not perform optimally when several reaction spots would need to be detected of various intensities, or when a control reaction spot with low intensity would need to be detected.

Several limitations of the SMC algorithm were also identified. Firstly, the running time is long and it increases proportionally with the number of particles being sampled. For sampling 125 particles the algorithm takes 1.6 hours, whilst for sampling 500 particles it takes 9.6 hours. This limitation could be overcome by improving the code and replacing the loops with faster structures. Another method of overcoming this limitation is through parallelisation of the algorithm. This process has already been studied by different research groups and reviewed by [76]. One of the easiest methods of parallelisation is performed by running M independent particle filters with N particles and, at the end, average the M estimators. Several algorithms implementing and improving this concept have been developed such as DRNA, particle island and α-SMC. Another common limitation of the particle filtering algorithm is the weight degeneracy which occurs when one particle has the weight approximately 1, while the others have weights close to 0. Also, the number of particles should be high enough to avoid this problem [77].

Taking into account the advantages and limitations enumerated above, depending on the S:N actually generated by the biosensor platform and the spot number either of the three methods could be selected. In the case where the chip signal has high S:N then the deterministic method results into higher intensity deviation than the GPR and SMC methods. An advantage posed by SMC is that any type of noise can be encoded into the SMC algorithm. Another advantage of SMC over the other two methods is that it can accurately determine the location, size and intensity of the spot. Moreover, if multiple spots of varying intensities were to be detected using the deterministic approach it will be harder and increasingly biased for detecting the spot intensity.

Image processing tools for detection of spots in an image are also available. For example, the Python package scikit-image offers has a collection of various algorithms for image processing [78]. Edge detection and segmentation are some of the functions used for detecting spots of various characteristics in an image. However, these can only be applied to high resolution images. In terms of methods of processing images obtained from biological data similarly to the ones presented in this chapter, one earlier study developed a method based also on Bayesian inference for the analysis of microarray assays [79]. In this study they have incorporated the uncertainty of estimating the expression levels into their analysis of microarray assay images. However, the microarray images have a higher resolution than the images obtained from the biosensor, so this was not considered for testing on the data from this chapter.

In terms of future research directions for the work presented in this chapter, relevant modifications would be added to the algorithm once the immunoassay has been successfully developed. Improvements need to be made in order to be able to use it on a real image obtained after running an immunoassay. These are detailed next. Currently, the particle filtering algorithm is used for identifying the position and intensity of a circle. However, if needed, the shape of the blob could be changed. Also, the image produced by the chip might have a gradient background noise caused by the positioning of the LED. This could also be easily incorporated into the algorithm in order to account for the background noise. The algorithm could also be easily extended for identifying multiple spots of various sizes and intensities on a 16×16 image array. For example, one particle would represent an image with multiple circles of size r_n and intensity i_n, instead of just one.

In conclusion, three algorithms were developed for the detection of spot intensities on a biosensor image array. These were tested and compared on synthetically generated images. Out of the three, the SMC performed best in terms of accuracy and decreased variability in estimating the intensity. The next chapter describes the approach of developing and running immunoassays on the biosensor platform with the aim to obtain images similar to the ones presented in this chapter.

Chapter 4

Developing an immunosensor for the detection of a fever associated disease

4.1 Introduction

Immunoassays are a sensitive method of detection for infectious diseases, as they are based on the antigen-antibody interaction [80]. Concomitantly, the world of electronics is dominated by the low cost, mass-manufactured complementary metal oxide semiconductor (CMOS) technology, which has made a huge impact on sensing technology. By combining immunoassay techniques and CMOS technology a powerful tool for the *quantitative* detection for infectious diseases could be developed. Well-established diagnosis techniques for infectious diseases include Enzyme-Linked Immunosorbent Assay (ELISA), microscopy and microorganism culture, and nucleic acid-based assays [12]. The standard overall process of detection is, however, time-consuming, costly, labour intensive and requires complex sample preparation. Thus, the use of biosensors in this field would offer the possibility of a low-cost portable technology platform that can identify pathogens rapidly and help in predicting appropriate treatment [14]. Other advantages of using biosensors include the use of small sample volumes, high selectivity and sensitivity and rapid response [15]. In this chapter a sensitive diagnostic immunoassay tool which is based on the CMOS-based biosensor described in Chapter 2 Section 2.1.2 was developed for the detection of a fever associated disease, Human African Trypanosomiasis (HAT).

Due to its high detection sensitivity, an ELISA format was selected to be adapted and used on the CMOS-based biosensor for the detection of a fever associated disease. In this chapter, an indirect ELISA format was used for the detection of antibodies in infected serum samples (Figure 4.1). This type of ELISA delivers high flexibility since different primary antibodies can be used during the assay which require only a single type of labelled antibody to be used for the final detection step. Additionally, as the primary antibody is not labelled, maximum immunoreactivity is retained. A limitation of this type of assay could be represented by a possible background noise generated from cross-reactivity of the secondary antibody to the adsorbed antigen [81].

The antibody detection is performed in two stages after the specific antigens have been immobilised onto the surface either through passive adsorption or covalent attachment. The antigens are generally diluted at a concentration of 2-10 μg/ml in an alkaline buffer [81]. The high pH of the buffer aids solubility of proteins and ensures that they have an overall negative charge, which helps binding to a positively charged surface [82]. In the first stage of the indirect ELISA, the antibodies of interest from the infected serum sample, i.e. primary antibodies, attach to the antigens. Next, a labelled secondary antibody, often polyclonal and anti-species, binds to the primary antibody. Enzymes such as horseradish peroxidase (HRP) are generally used as labels. However, in this case, as the final result of the assay should be a precipitate which blocks the transmission of photons to the chip surface, the labels used are gold nano-particles (AuNP), as previously used on this platform to detect HIV gp120 [5]. This method of detection works on the same principle as the immunogold-silver staining (IGSS) method which is commonly used as a immuno-histochemical visualisation technique [83]. Metallic silver precipitates on heavy metals such as gold, and thus enlarges the small colloidal gold particles. The reaction is time dependent and a longer than needed enhancement time could determine a background noise due to silver precipitates formed by self-nucleation. This reaction is usually performed at room temperature.

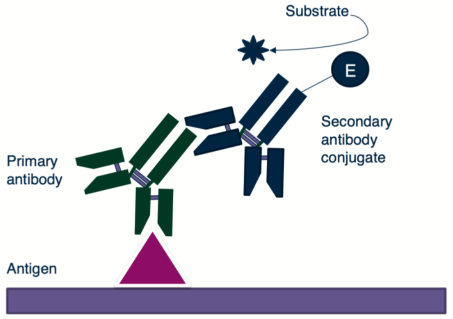


Figure 4.1: Indirect ELISA scheme. The antigens are immobilised on the surface and an unconjugated primary detection antibody added to bind with the antigen. Next, a conjugated secondary antibody directed against the host species of the primary antibody is added. The substrate then produces a signal proportional to the amount of antibody bound to the antigen. Reproduced and modified from [84]

Adapting the general principle of immunoassay to a silicon surface may involve the chip surface to be directly functionalised for the immunoassay. However, since the number of immunoassays required for the experiment exceeded that of CMOS chips provided as part of the Multicorder project, an alternative approach, using poly-methyl methacrylate (PMMA) slides, was employed. PMMA has important characteristics relevant to the development of immunoassays on their surface. These include: high optical transparency, low cost, high scratch resistance, versatility in fabrication and impact resistance. Moreover, PMMA has been previously used successfully for the immobilization of enzymes, proteins and DNA [85–87]. In this project, PMMA slides which were placed directly on the sensor chip surface were used as the surface to run the indirect immunoassay on. The PMMA slide on which the immunoassay was run can then be disposed of allowing a second sample to be tested on a fresh PMMA slide on the same detector surface.

Generally, immunoassays are performed on specially designed ELISA 96-well microplates. These are made from polystyrene or polyvinyl chloride (PVC) which are then treated using different approaches in order to ensure affinity to molecules with hydrophobic, hydrophilic or mixed characteristics [88]. Both polystyrene and PVC are hydrophobic surfaces, but this can be easily modified using various techniques which alter the chemistry of the surface. For instance, in the case of polystyrene, it is its benzene ring which gets modified in the chemical functionalisation process [88]. Through it, carboxyl and amine groups can be added to the surface.

Treating the surface with chemicals such as glutaraldehyde (GA) is usually used to prevent non-specific protein adsorption [89]. This method is efficient for antigens with high carbohydrate content since these bind poorly to plastic surface. Additionally, using hydrophobic immobilisation methods could have a denaturing effect on the biomolecules, as they unfold to expose hydrophobic regions that can interact with the surface. Therefore, covalent attachment of proteins is preferred in order to avoid the above mentioned problems. A similar protocol of PMMA functionalisation was used previously [89].

In order to test the optimal amount of antigen needed to coat the well and the optimal amount of anti-species conjugated with gold nanoparticles (AuNP) which it can detect, a checkerboard titration (CBT) method was used [90]. The process of CBT involves the dilution of two reagents against each other to examine the activities inherent at all the resulting combinations. Given the small reaction surface, lower antigen concentrations were also be tested, as density affects the binding to the surface. A high concentration of antigen may cause steric inhibition, i.e antigen molecules are too closely packed. High concentration of antigens may also increase stacking or layering, which may allow less stable interactions of subsequent reactions. Coating time and temperature are also important: times and temperature are usually inversely proportional, i.e for high temperatures (37 °C or room temperature) assays are normally left for 1-3 hours and for lower temperatures (4 °C) coating is done overnight.

For this immunoassay, it was aimed to detect one fever-associated infectious disease, Human African Trypanosomiasis (HAT), which is prevalent in the low to middle income countries, especially in regions of west, central and east Africa. This is mainly caused in humans by the pathogen *Trypanosoma brucei*. In west and central Africa, *T.b.gambiense* causes a chronic disease taking several years to evolve from stage 1 (hemolymphatic) to stage 2 (neuroencephalic).

If left untreated, it may be fatal as it affects the nervous system of the host in the later stages of development [91]. In stage 1, trypanosomiasis *T.b.gambiense* infections can be treated with pentamidine and, in stage 2 with nifurtimox and eflornithine combination therapy [92]. The surface of a trypanosome is coated with approximately 10^7 different types of the so-called molecules of variant surface glycoproteins (VSGs) that are highly immunogenic. This dense layer of VSG dimers acts as a protective layer for the invariant surface glycoproteins (ISGs) and other surface components [93]. A particular characteristic of this parasite is a mechanism of antigenic variation whereby the parasites change their variant surface coat to avoid the immune system. In this process one VSG coat is replaced by another one of a different antigenic type that is not recognized by antibodies raised to the previous type [94].

Early detection and treatment of the disease is important not only to reduce the transmission rate in the community but also because treatment for patients in the later stages is more complicated and the risk of severe side effects is significantly higher [94]. Gambiense HAT is characterized by low parasitaemia levels (<5 ng total trypanosome protein/ ml blood) and thus, screening for the presence of specific antibodies after infection with the parasite constitutes a valuable detection tool. The Card Agglutination Test for Trypanosomiasis (CATT), which uses the LiTat 1.3 VSG (L1.3) as biomarker was the first HAT detection test [95]. More recently, lateral flow rapid diagnostic tests have also been developed for HAT detection [96, 97].

In order to combat the pathogen's antigenic variation mechanism and to increase the sensitivity of an antibody detection test, multiple antigens can be used simultaneously as biomarkers for the detection test. Apart from L1.3, LiTat 1.5 VSG (L1.5) has also been identified as another suitable biomarker, since both of the VSG types are generally expressed early in a trypanosome infection (ultimately up to a thousand different VSG types are available, hence finding those that predominate in early infection has been useful) [98, 99]. It is the presence of antibodies to these markers that is used as the basis of these tests. Other antigens which are recognized by the sera of HAT infected patients have been identified through proteomic studies, in particular the invariant surface glycoprotein 65 (ISG65) [100]. Another ISG with diagnostic potential was ISG75, even though the ELISA results for this glycoprotein were weaker than those for ISG65 [100]. For developing diagnostic tools based on the aforementioned biomarkers, recombinant antigens are preferred to native antigens, in order to eliminate the infection risk for staff and the need for laboratory animals for antigen production.

As the above markers have been successfully used in the detection assays or, more recently, the lateral flow rapid diagnostic tests [95, 99, 100] they were considered as useful starting points to develop a CMOS chip based immunoassay for the detection of HAT on a platform that had also been tested for the immuno-detection of HIV [5]. Thus, this chapter aimed to firstly develop the recombinant antigens rL1.3, rL1.5, rISG65 and rISG75 in order to use them for the screening of sleeping sickness on the CMOS-based biosensor. Additionally, it also aimed to develop the immunoassay which would be run on the chemically functionalised PMMA slide on the biosensor. The initial aim of this chapter was to develop a multiplex immunoassay using all four recombinant antigens. However, due to the Covid-19 pandemic which brought upon the laboratory closure, only the recombinant antigen rISG65 was used on the immunosensor platform.

4.2 Materials and Methods

4.2.1 Chemicals

All reagents were purchased from Sigma-Aldrich (Poole, Dorset, UK) unless otherwise stated. For the DNA manipulation experiments *Escherichia coli* (*E.coli*) strain BL21(DE3) from Novagen was utilised. The bacterial vector pET-28a(+) was also purchased from Novagen; The NuPAGE 4-12% Bis-Tris gels were purchased from Invitrogen and were used for the separation and visualisation of proteins.

4.2.2 Construct engineering

The mRNA coding sequences specifically selected regions within the total protein sequence of L1.3, L1.5, ISG65, ISG75 were identified using relevant literature research [94, 98, 101, 102], GenBank and UniProtKB. The expression vectors which were optimised for being cloned into pET-28a(+) vector with the cloning sites EcoRI/XhoI were commercially obtained from GenScript Biotech. The final theoretical molecular weight was calculated using Expasy.

rL1.3 The protein sequence was obtained from GenBank: accession ID AHW98113.1 (UniProtKB X5GEX5) with 479 aa (region 24:372 trypanosomal VSG domain) with the mRNA coding sequence accession ID KJ499460.1 with 1440 bp [94]. The optimised sequence length was 1062 with the GC content of 58.53%. Based on this, the final molecular weight should be approximately 37.4 kDa (36.6 kDa +0.8 kDa (6xHis Tag)).

rL1.5 The protein sequence was obtained from GenBank: accession ID ADV15625.1 (UniProtKB E7EDN2) with 502 aa (residues 33:426) with the mRNA coding sequence HQ662603.1 with 1635 bp [98]. The final sequence length was 1197 with GC content of 58.03%. Based on this the final molecular weight should be approximately 43.2 kDa (42.4 kDa+0.8 kDa (6xHis Tag)).

The same protein regions as in [99] were selected to be purified, as they represented the native DNA sequences of the antigens.

rISG65 The protein sequence was obtained from GenBank: accession ID AAA30147.1 (UniProtKB Q26712) with 436 aa (residues 19:385) with the mRNA coding sequence M86709.1 [101]. The final sequence length was 1116 with GC content 54.62%. Based on this the final molecular weight should be approximately 41.2 kDa (40.4 kDa+0.8 kDa (6xHis Tag)).

rISG75 The protein sequence was obtained from GenBank: accession ID AAC41567.1 (UniProtKB Q26769) with 523 aa (residues 28:468) with the mRNA coding sequence L07866.1 [102]. The final sequence length was 1338 with GC content of 57.87%.Based on this the final molecular weight should be approximately 50.0 kDa (49.2 kDa+0.8 kDa (6xHis Tag)).

Specific protein regions were selected based on the study by Sullivan et al [100].

4.2.2.1 Transformation protocol into BL21(DE3) competent cells

The GenScript constructs were expression-ready constructs containing the desired DNA sequence cloned into kanamycin resistant (K+) pET-28a(+) vector at the EcoRI/XhoI cloning sites. The constructs were each prepared according to their vendor instructions and further transformed into BL21(DE3) competent *E.coli* cells.

The BL21(DE3) cells were then thawed on ice and 50 μl of cells were pipetted into a transformation tube; 1.5 μl of the construct containing the plasmid DNA was added. The tube was flicked several times to mix the DNA and cells and left on ice for 30 minutes. Heat shock of 42°C for 60 s was then applied and the mixture was placed back on ice for 5 minutes. Next, 950 μl lysogen broth (LB) were added and placed at 37°C for 1 hour (ZHWY-200D Incubator Shaker). Dilutions of the mixture were plated on LB-kanamycin agar plates. Resistant cells were selected by inoculation of individual colonies in 10 ml LB-K(+) broth (0.1% K) and incubated overnight. Glycerol stocks were then prepared using 750 μl 20% glycerol and 750 μl culture and stored at -80°C.

4.2.2.2 Protein expression protocol

The starter culture was prepared from the glycerol stock and incubated overnight. This was done by adding 10 μl kanamycin to 10 ml LB. The frozen glycerol stock was scraped with a plastic stick and mixed in the solution. The starter culture was then inoculated into a fresh culture and incubated for 2-3 hours until it reached OD_{600nm} of 0.5-0.6.

Optimisation of culture conditions for protein expression Different induction conditions, i.e. molar concentration of isopropyl β-D-1-thiogalactopyranoside (IPTG), incubation period and temperature, were tested for determining the optimal expression of the proteins. Concentrations of 0.2 mM, 0.4 mM, 1.0 mM IPTG were tested with temperatures and incubation periods of: 15°C overnight, 25°C overnight and 37°C 2-3 hours. By lowering the temperature of the induction temperature, correct folding of the molecule can be improved by avoiding the formation of inclusion bodies, as translation is slowed [103, 104]. Moreover, lower IPTG concentrations facilitate the soluble form of the target protein.

For large scale protein purification, 1 L of liquid medium (K+) was inoculated with a freshly grown colony or 10 ml of freshly grown culture and incubated at 37°C until OD_{600nm} reached 0.4–0.8. Protein was expressed using the optimal induction time/temperature determined in the small scale trial. The cells were collected after centrifugation at 10°C, 9,000 g for 15 minutes. The supernatant was discarded and cells were resuspended in 20 ml of 0.1M PBS. The resulting supernatant which was collected by centrifugation of the cell culture for 10 minutes at 4,500 g was resuspended in a binding buffer supplemented with 1 pill per 50 mL of binding buffer of protease inhibitor cocktail tablet (EDTA-free Protease Inhibitor Cocktail, Roche Diagnostics GmbH). This was then vortexed until everything was dissolved.

In order to disrupt the membrane cells, the bacterial cells were sonicated at 4° C by using 15 sonication cycles –10 seconds pulse burst + 10-20 seconds rest at 18 μm amplitude (Soniprep 150, MSE Ltd). The resulted debris was removed by performing an ultracentrifugation step at 30,000 x g for 30 minutes at 4°C.

Ni^{2+} Affinity Chromatography The proteins were then purified using His GraviTrap kit(TALON, Sigma) according to the affinity chromatography process, i.e. cell lysis, binding of the tagged protein to an affinity resin inside the column, washing off the unwanted lysate and eluting the tagged protein. For efficient binding the pH of the lysate should be between 7.5 and 8 and the buffer should not contain chelators, such as EDTA or citrate, or high imidazole concentrations (>30 mM) [105]. Thus, after lysis the resulting supernatant was loaded onto was loaded onto the GraviTrap Ni-charged affinity chromatography column which was previously equilibrated using 10 ml of binding buffer (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole, pH 7.4). At this step the recombinant protein was bound to the Ni-Sepharose column. Next, the column was washed with washing buffer (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole, pH 7.4) and the protein was eluted using elution buffer (50 mM NaH_2PO_4, 300 mM NaCl, 250 mM imidazole, pH 7.4). Imidazole competes for the Ni^{2+} and it displaces the His Tag and the tagged protein falls off allowing thus their elution. The imidazole competitor was eliminated and the protein was concentrated by centrifugation using a centrifugal filter unit with a 30 kDa cut-off membrane (Amicon Ultra-15, Millipore). The final protein was stored in solution in PBS, 5% glycerol at -80°C. This last step enhances the solubility and stability of the proteins.

At the end of the purification stage, the sample was collected, stained with Coomassie dye and loaded on SDS-PAGE using Mini Gel Tank (life technologies) and run at 150 V. The protein concentration was determined by measuring the UV absorption at A280 and calculate the concentration of the protein using the NanoDrop® 10000 spectrophotometer (Thermo Scientific).

4.2.3 Testing the recombinant antigens on human serum samples and antisera

The reactivity and specificity of the recombinant antigens were determined using sera from infected humans or uninfected controls. In order to test the reactivity of the purified proteins, 39 HAT positive, 41 control human sera and 4 rabbit anti-sera (anti-ISG65, anti-ISG75, anti-L1.3, anti-L1.5) samples were used. The human samples (n=80) originated from patients from endemic regions in the Democratic Republic of Congo [106]. The testing was performed by myself at the Institute of Tropical Medicine (Antwerp, Belgium). Native antigens for L1.3 and L1.5 (nL1.3, nL1.5) and recombinant L1.3 and L1.5 expressed in *L.tarentolae* were also provided. The recombinant L1.3 and L1.5 expressed in *L.tarentolae* were kindly donated by Dr. Barrie Rooney and the antibodies for ISG65 and ISG75 were donated by Prof Mark Carrington (University of Cambridge).

4.2.3.1 ELISA procedure

The immunoassay protocol was adapted from Lejon [107]. Microplates were coated overnight at 4°C with 100 μ l/well of purified recombinant protein at 4 μ g/ml or with native antigen at 2 μ g/ml in phosphate buffer (pH 6.5). Further manipulations were undertaken at room temperature. After coating, the wells were blocked with blocking buffer (0.01 M sodium phosphate, 0.2 M sodium chloride, 0.05% NaN_3, 1% (casein) skimmed milk powder, pH 7.4) for 1 hour. The sera was diluted at 1:150 in PBS blocking buffer. Antibody binding was visualised with goat anti-human IgG conjugated with horseradish peroxidase diluted in PBS-Tween (1:40000) (Jackson ImmunoResearch, Europe Ltd). The optical density values were read at 450 nm (Multiskan RC Version 6.0; Labsystems). The corrected optical density (OD_{corr}) values were calculated by subtracting for each serum the OD reading in the control well from the OD reading in the antigen coated well.

4.2.4 CMOS-based detection platform

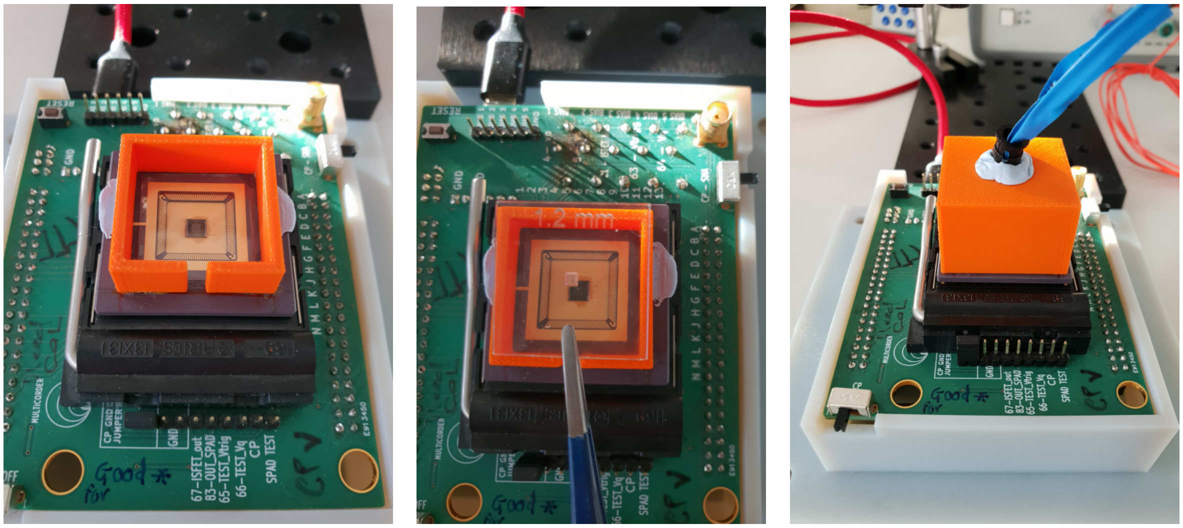
The platform had already been developed by Al-Rawhani et al [29] as part of the Multicorder project and its design and setup are thoroughly described in the respective paper and summarised in Chapter 2 Figure 2.3. The setup consists of a CMOS-based photodiode (PD) array and a light emitting diode (LED). The chip measures 3.4 x 3.6 mm with an active sensor array area of 1.6 x 1.6 mm.

4.2.4.1 Data acquisition and processing

The CMOS-chip is connected to an ARM mbed STM32 Nucleo-F334R8 board. The mbed microcontroller was programmed to provide addressing signals and to acquire the output readings from the array (MST, School of Engineering). These readings were then transferred via USB to a MATLAB program developed by Valerio Annese (MST, School of Engineering). Matlab files were further processed using Python. Due to the fact that the reaction took place on the whole surface of the APS, the spot detection methods proposed in Chapter 3 were not used in this chapter. Instead the mean of the image array obtained from the PD array was computed for each time frame (36 tf/s) and was reported in terms of the PD voltage (V).

4.2.4.2 Setup

The setup which was developed for this chapter is presented in Figure 4.2. The wires bonded on the chip surface are protected with a glass cover slip. A support (22x22 mm) was 3D printed and glued to the chip. This would ensure that the positioning of the PMMA slides (22x22 mm) and, therefore, the analysed data would be reproducible. The adapted ELISA assay is performed on PMMA slides and afterwards the slide is placed on top of the chip and the light transmittance reaching the PDs is read.



(a) (b) (c)

Figure 4.2: The chip setup. The chip is connected to the mbed and it is surrounded with the 3D printed support. The PMMA slide is then placed on top (b) and covered with a lid with 560 nm LED attached to it (c).

Platform stability The stability of the platform was tested by measuring the light transmittance over 500s. For each time frame the mean of the image array (V) was computed and results were plotted.

PMMA slides The immunoassay was only tested for one recombinant antigen, hence the PMMA slides were designed to have one well in the middle. The slides were sized at 22 x 22 mm with a 1.2 mm well in the middle which could be positioned exactly above the APS of the chip. CO_2 laser was used to cut the PMMA slides and engrave the wells. This was done by Chunxiao Hu (MST, School of Engineering). The variability between the PMMA slides was measured by reading n=18 slides on the platform and comparing the obtained image array mean.

Surface functionalisation The PMMA slides were prepared for covalent attachment of the antigens. The functionalisation protocol was adapted from [85]. PMMA slides were first hydroxylated using oxygen plasma treatment at 150 W for 40 s by Valerio Anesse (MST, School of Engineering). Following this, the slides were functionalised using 4% 3-aminopropyltriethoxysilane (APTES) solution for 1 hour at room temperature, followed by rinsing them with 70% ethanol solution once and with dH_2O twice [108]. After drying, the silanized glass slides were reacted with a 1% solution of glutaraldehyde (GA) in 0.1 M PBS for 1 hour, followed by rinsing with PBS buffer. The added GA yields aldehydes which can form an imine linkage with the primary amines on the protein (Figure 4.3). The APTES+GA functionalised PMMA surfaces were then reacted with the antigen solution in carbonate buffer (pH 9.0) (Figure 4.3).

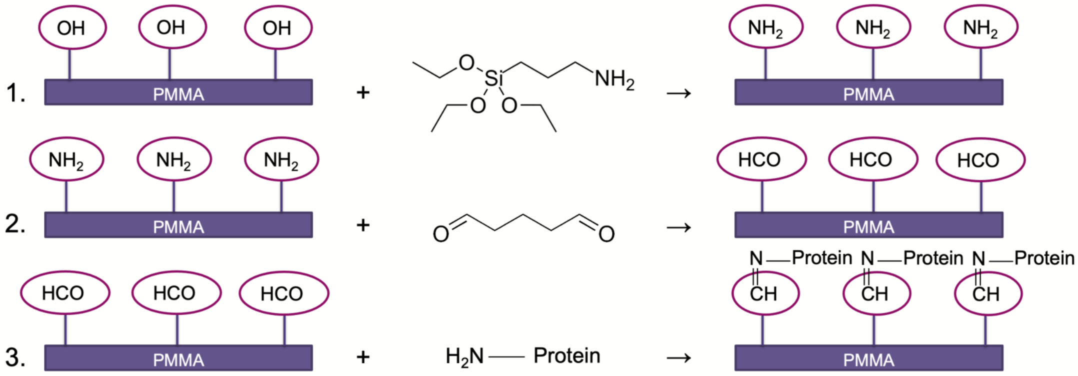


Figure 4.3: Schematic of the steps involved in the chemical functionalisation of a hydroxylsed PMMA surface. The silane ends of the molecules attach to the hydroxylsed PMMA substrate leaving the aldehyde groups to react with the amine groups on the proteins, as represented below. Glutaraldehyde was subsequently used to modify the surface,yielding an aldehyde that can form an imine linkage with the primary amines on the protein.

4.2.4.3 Biosensor immunoassay protocol

The antigens were immobilized on the functionalized surface and left for drying overnight at room temperature. Immobilisation of the recombinant antigen diluted in carbonate buffer (pH=9). As a negative control for the titrations, fetal bovine serum (FBS) 1:50 was used. In order to determine the optimal reagent concentrations, titrations with different recombinant antigen and antibody concentrations were made. Antigen concentrations ranging from 0 to 10 μg/ml were coated on to the surface of the PMMA slides and each of these concentrations were tested against 3 serum antibody concentrations: 0 μg/ml, 10 μg/ml and 20 μg/ml.

ELISA The antigens were first rehydrated in PBS for 1 min. The rest of the protocol was adapted from the one presented in Section 4.2.3.1 by changing the incubation times, reagents quantity (1 μl of solution for each step) and the substrates. The schematic representation of the immunoassay is illustrated in Figure 4.4. The slides were then blocked with Blocking-PBS (1% milk powder, 0.05% NaN_3) for 15 minutes at room temperature. After washing with PBS-Tween20 (0.05% Tween20), the primary Ab from rabbit serum was added and left for incubation for 15 minutes at room temperature. The secondary Ab, goat anti-rabbit IgG conjugated to 12 nm AuNP (Jackson ImmunoResearch), was added and left for incubation for 15 minutes. After washing with PBS-Tween20 the silver enhancer solution was added. The silver solution is made up of 2 solutions: A: silver nitrate and B: initiator solution. The two need to be mixed 1:1 just before adding it to the slide in order to prevent the background noise caused by the quick silver auto-nucleation. Each reaction was run in duplicate.

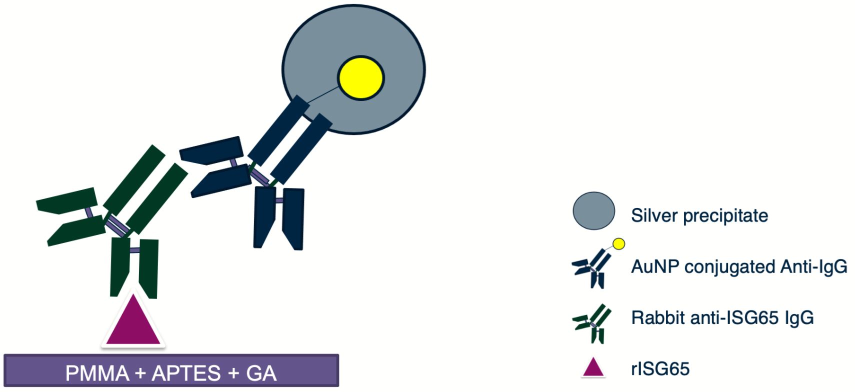


Figure 4.4: Schematic representing the immunoassay. The PMMA surface on which the immunoassay is performed is functionalized with (3-aminopropyl)triethoxysilane (APTES) and glutaraldehyde (GA). The precipitate sinks to the surface.

Determining the optimal LED wavelength The ELISA procedure was first tested using IgG and anti-IgG conjugated to AuNP. Different dilutions of the detection antibody, i.e. anti-IgG conjugated to AuNP, were added to the slides (1:5, 1:10, 1:20, 1:40, 1:80, 1:160, 1:320). The results (% of light transmission) were measured using a Raman microspectrometer in order to determine the optimal wavelength to be used on the platform.

ELISA output processing The PMMA slide was read on the platform before the reaction (control slide) and during/after the reaction (reacted slide). The results were processed and presented as the difference ΔV between the control slide and the reacted slide at different time points during the reaction. Afterwards, the slide was washed and the remaining silver precipitate was read once again. Linear regression was used to fit a straight line through the results. This was achieved using Python scipy package (stats.linregress). The limit of detection (LOD) and limit of quantitation (LOQ) in each case were calculated as following:

$$LOD = 3.3 * \frac{\sigma}{S} \tag{4.1}$$

$$LOQ = 10 * \frac{\sigma}{S} \tag{4.2}$$

where: S= slope of calibration curve and σ = residual standard deviation of the regression line

4.3 Results

4.3.1 Protein expression and purification

All four recombinant proteins were initially induced with 0.4 mM IPTG for 3-5 hours at 37°C. Total protein (both soluble fraction and insoluble fraction) bands for the four antigens were clearly observed in the SDS-PAGE (4.5 A). However, the soluble fraction for the recombinant proteins, especially rL1.3, rL1.5 and rISG75, had very low to zero expression level. Thus different induction conditions were tested. Induction at 22°C overnight with 0.4 mM IPTG had the best outcome in terms of protein expression levels. The purified protein was run on the SDS-PAGE and a clear improvement in the expression of the soluble protein was observed especially for rISG65 and rISG75 (4.5 B). For the variant surface glycoproteins, however, after lowering the induction temperature, there was very small improvement in terms of the expression of the soluble protein.

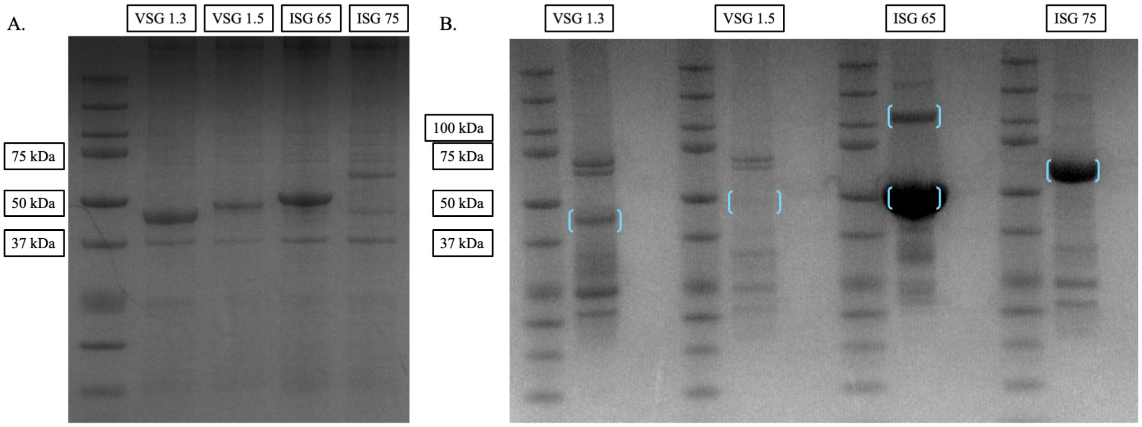


Figure 4.5: Coomassie stained SDS-polyacrylamide 12 % gel of the recombinant proteins. The molecular weights of the protein ladder are marked on the left side. A) Expression of total (soluble+insoluble fractions) recombinant proteins with induction with 0.4 mM IPTG for 3-5 hours at 37°shows some expression for the 4 recombinant antigens. Soluble proteins were not expressed for rL1.3 and rL1.5 B) Expression of the soluble fraction of the rISG65 and rISG75 was improved with induction with 0.4 mM IPTG overnight at 22°C. It is likely that for rISG65 its dimer was also expressed at around 100 kDa

Recombinant protein	Theoretical mass(kDa)	Concentration (mg/ml)	A260/ A280
LiTat 1.3 VSG	37.4	2.77	0.7
LiTat 1.5 VSG	43.2	0.40	0.69
ISG65	41.2	11.85	0.73
ISG75	50	4.96	0.62

Table 4.1: Protein A280 measurements: Protein yield after purification with protein induction at 22°C. ISG65 and ISG75 are smaller than the total antigen as only partial antigens of known immunogenicity were expressed. Evidence was used to choose regions of antigenicity [99, 100].

Based on the SDS-PAGE in Figure 4.5 B, rISG65 dimer might have also been expressed at around 100 kDa. Also, there are indications in the SDS-PAGE that the purification process might not have worked optimally, as there are still some possible *E.coli* contaminants left in the samples both at 37°C and 22°C induction (Figure 4.5). It is known that the main contaminants during purification of proteins expressed in *E.coli* are the GroEl and DnaK chaperones which appear at around 70 kDa as a doublet [105]. Although at a low concentration, the doublets can still be spotted in the SDS-PAGE and they became more evident when induction was performed at a lower temperature. DnaJ is another *E.coli* chaperone which appears at around 40 kDa and a protein of this mass is evident in Figure 4.5 A (at 37 kDa).

The protein concentration measured for the purified proteins is presented in Table 4.1. Based on the A280 results the protein yield was lowest for rL1.5 at 0.4 mg/ml and the highest for rISG65 at 11.85 mg/ml.

4.3.2 Testing the recombinant antigens on human serum samples and antisera

A bar chart was used to represent the results of the ELISA tests (Figure 4.7). Two-tailed heteroscedastic t-test was performed to check whether there was a significant difference between the positive and control samples in terms of absorbance levels. A p-value lower than 0.05 was obtained in all of the 6 cases represented in the figure below, which suggests that the ELISA coated with the recombinant proteins could differentiate between control and infected patients. The p-values for each case were: 8.01e-29 (nL1.3), 8.95e-26 (nL1.5), 1.15e-10 (rL1.3), 2.44e-5 (rL1.5), 3.57e-18 (rISG65), 4.83e-8 (rISG75). As it may be observed, both native VSGs (nL1.3, nL1.5) separate the positive and control sera very well both in terms of specificity and sensitivity.

In comparison to these native antigens, the number of false positives is higher using the recombinant antigens (rL1.3, rL1.5, rISG65, rISG75) which might be explained by the remaining *E.coli* contaminants in the samples as observed in the SDS-PAGE (Figure 4.5 A & B). The mean ODs for infected patients were 2.7 (rL1.3), 2.03 (rL1.5), 3.05 (rISG65), 2.78 (rISG75) and for control 0.94, 1.27, 0.91, 1.4 respectively. Based on these values and those aforementioned obtained from the t-test, the purification of the rISG65 worked the best out of the four recombinant antigens, with a lower number of false positives and higher number of true positives.

ELISA was also performed for anti-sera for each of the recombinant antigens (Figure 4.6). Each antisera was made using rabbit sera by injecting purified antigens. For anti-L1.3 antibody the highest reactivity was obtained with the recombinant antigen. There was, however, some reactivity with the other recombinant and native antigens excluding L1.3, especially with rL1.5, but not statistically significant. The p-value obtained after comparing the reactivity of L1.3 antigens with anti-L1.3 antibodies and the reactivity of the other antigens was 0.001. For anti-L1.5 the highest reactivity was obtained with the native L1.5. Very low reactivity was also obtained with the other recombinant antigens. The reactivity of anti-L1.5 with L1.5 antigens was significantly higher than with the other antigens (p-val =0.002). In the case of the antibodies for the invariant surface glycoproteins, however, it seems there was a high degree of cross-reactivity, as binding was high with all of the four recombinant antigens, including the variable surface glycoproteins.

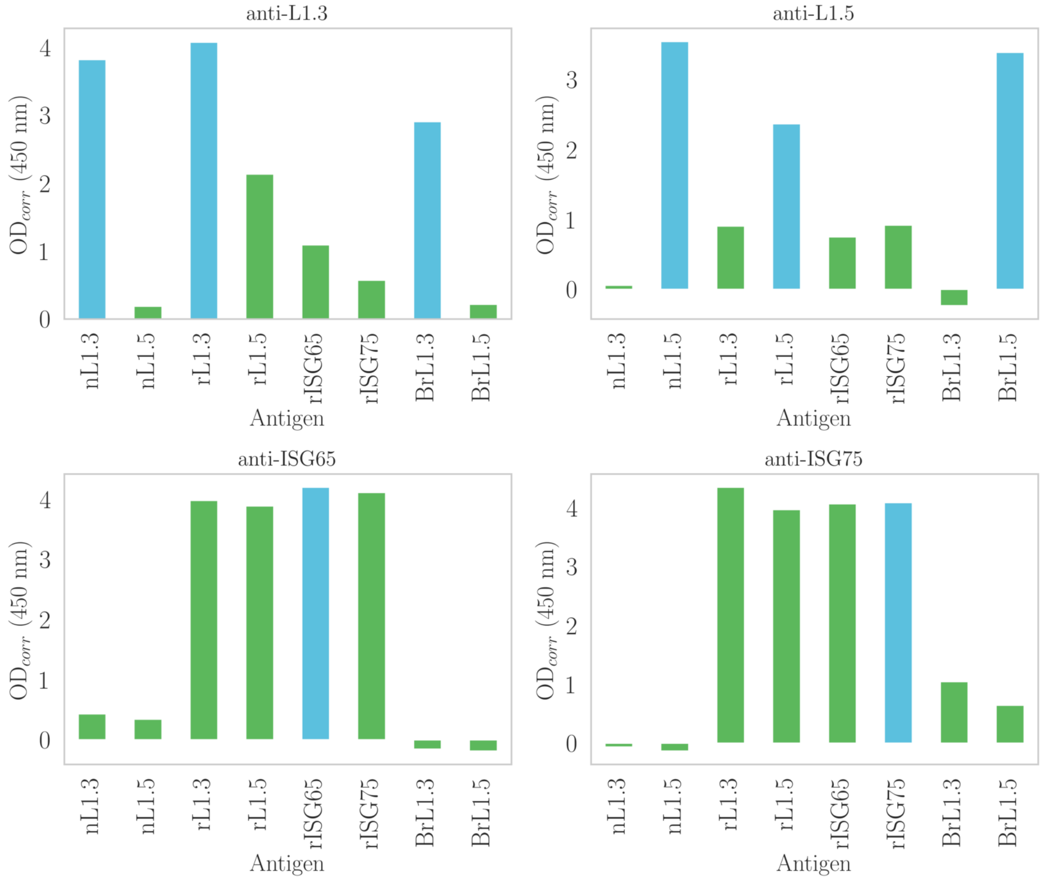


Figure 4.6: ELISA results against each antibody for the antigens. Y-Axis: The mean absorbance value of the samples. X-Axis: The antigens in the following order: native antigens (nL1.3, nL1.5), recombinant antigens (rL1.3, rL1.5, rISG65, rISG75), recombinant antigens expressed in *L.tarentolae* (B_rL1.3, B_rL1.5). Anti-VSG 1.3 reacts best with rL1.3, nL1.3 and B_rL1.3, but there is some reaction with the other antigens as well.

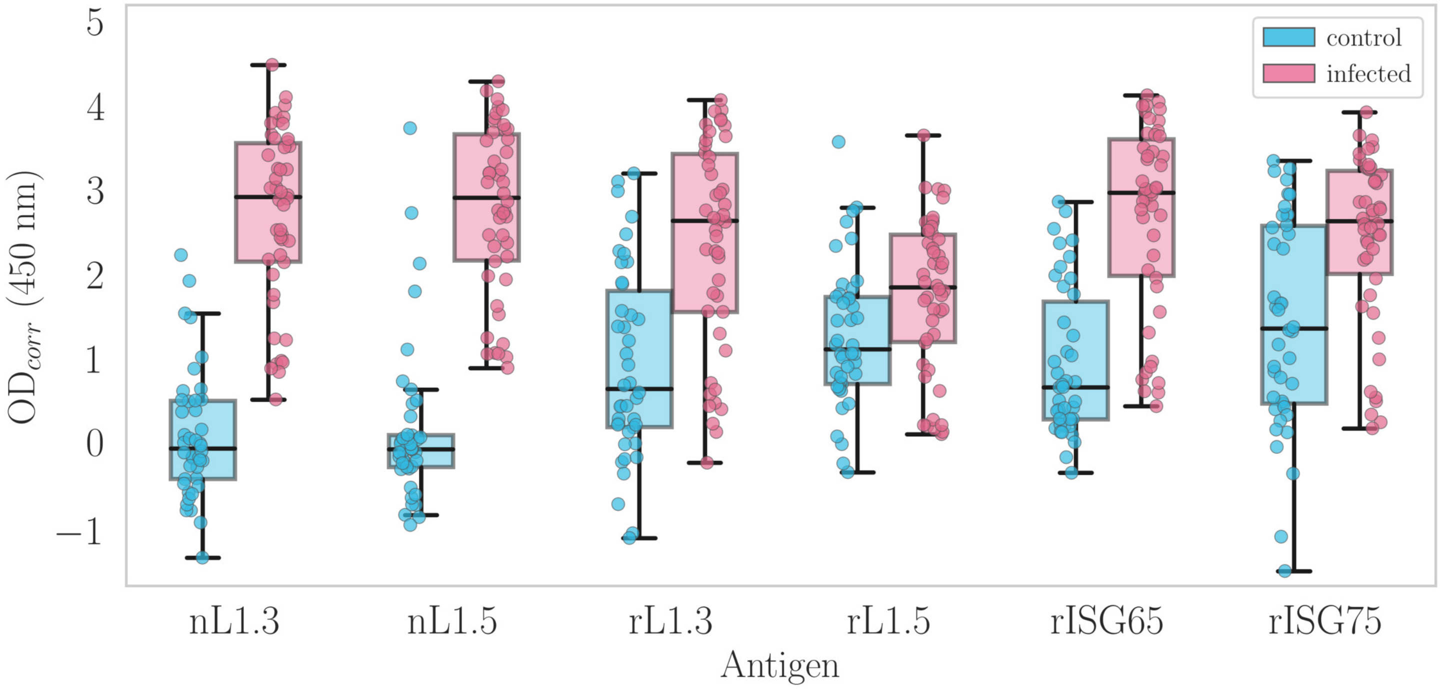


Figure 4.7: ELISA results against the human sera samples for the antigens.Human sera samples were diluted 1:1000 and used in duplicate on ELISA plates coated with the recombinant and native antigens. The mean ELISA signals are plotted against the recombinant proteins used. Y-Axis: The mean absorbance value of the samples. X-Axis: The antigens in the following order: native antigens (nL1.3, nL1.5), recombinant antigens (rL1.3, rL1.5, rISG65, rISG75). All of the antigens present high sensitivity, however the *E.coli* recombinant antigens present higher numbers of false positives, indicating low specificity. The difference between infected and control is statistically significant for all antigens used for detection (p-val<0.05).

4.3.3 CMOS-based antibody detection

Optimal LED wavelength detection Based on the results obtained from the microspectrometer, the quantity of silver precipitate was correlated with the quantity of detection Ab. For all of the conditions it can be observed that less light is transmitted between 450-600 nm (Figure 4.8). A closer inspection of the results indicate that 560 nm the lowest transmittance was registered for all concentrations: 98.7% for the blank well, 77.98% for 3.1 µg/ml antibody, 68% for 6.25 µg/ml antibody, 58.8% for 12.5 µg/ml antibody, 52.08% for 25 µg/ml antibody, 48.1% for 50 µg/ml antibody, 43.21% for 100 µg/ml antibody and 35.98% for 200 µg/ml antibody. Therefore, a LED with a 560 nm wavelength was used for reading the immunoassay results on the biosensor. The LED was connected to a power source which permits adjusting the intensity of the LED. The voltage which produces the best output was 2.7 V.

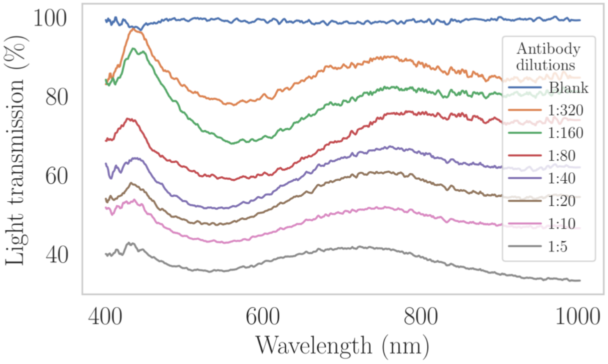
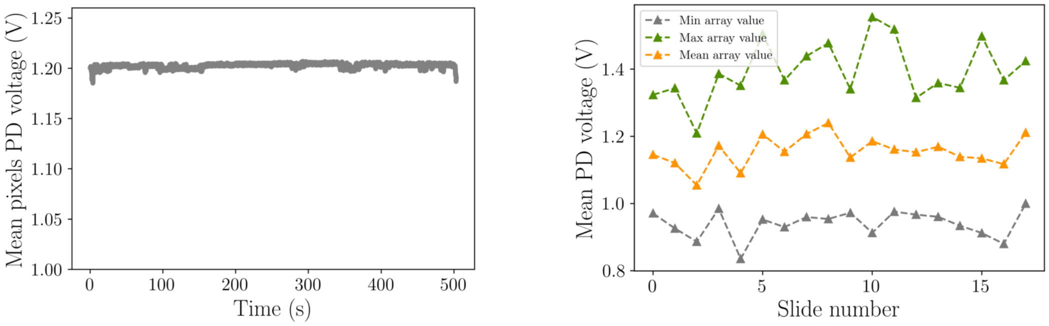


Figure 4.8: Transmission spectrum of the silver spots (developed after immunoassay with different gold conjugated antibody solutions) developed on a APTES-GA treated glass slide.

Platform stability The stability of the measuring platform is confirmed by the constant light transmittance values for a dry PMMA slide over the course of 8 minutes, as observed in Figure 4.9. The mean of voltage of the pixel array PD output over 500 s is 1.2 V with a standard deviation of 0.002.

PMMA slide variability Prior to surface functionalisation, the light transmission through 18 blank slides was read. The power for the LED was set at 2.7 V for all of the slides and the LED was positioned at the same distance in each case. In Figure 4.9b there is some variability between the readings of the light transmittance of each slide (min: 1.09 V, max: 1.24 V, mean: 1.15 V ± 0.04 V). Therefore, each slide should be read on the platform before the ELISA reaction as control.

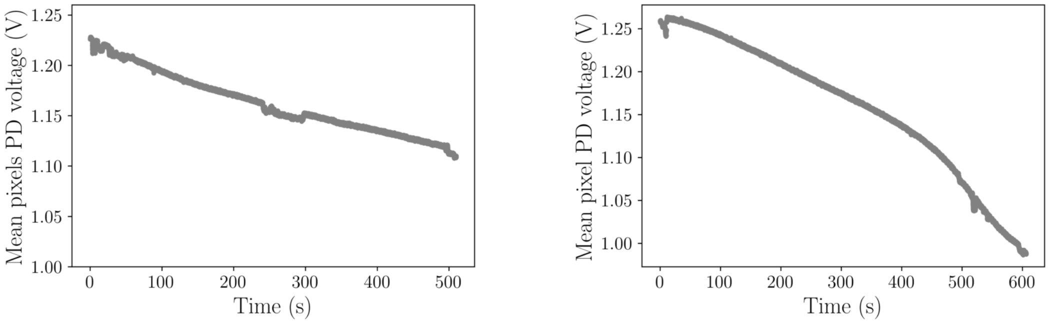
Silver auto-nucleation Two control slides consisting of one PMMA slide filled with 1 μl water and one PMMA slide filled with silver solution were used to detect silver auto-nucleation



(a) **(b)**

Figure 4.9: a) Mean pixel PD voltage measured for a dry PMMA slide over 500 s. b) Light transmission measured through PMMA slides prior to their chemical functionalisation. A high variability between them can be observed.

(Figure 4.10). Self-nucleation is noticeable from 300 s onwards. The rate of the decreasing the voltage for the first 300 s (-10 s) for both slides is R = -0.24 mV/s. For the slide with silver solution the rate (600s and 300s) is -0.61 mV/s and for the water slide (500-300s) R = -0.2.

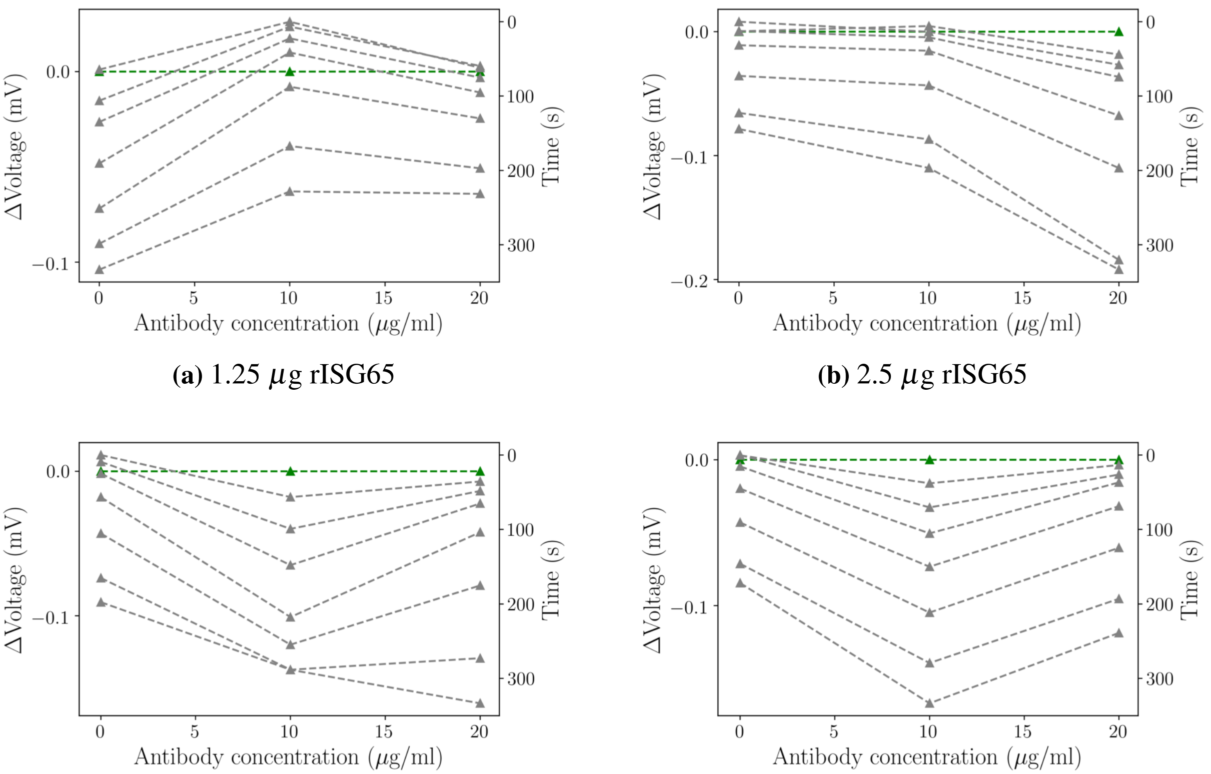


(a) **(b)**

Figure 4.10: a) Mean pixel PD voltage measured for (left) water and (right) water with silver solution

4.3.3.1 ELISA titration results

The developed recombinant rISG65 was used in this ELISA, as it had the best specificity and reactivity out of the four developed recombinant antigens. Antigen concentrations of 1.25 μg, 2.5μg, 5μg and 10μg were used to coat the functionalised PMMA slides well. As for the primary antibody, concentrations of 5μl, 10μl and 20μl were used. As a negative control, FBS solution was used. The chip results read during the final stage of the immunoassay are presented below in Figure 4.11. From these graphs it appears that the PMMA slide coated with 2.5 μl rISG65 showed a decrease in voltage correlated with the antibody concentration.



(a) 1.25 μg rISG65

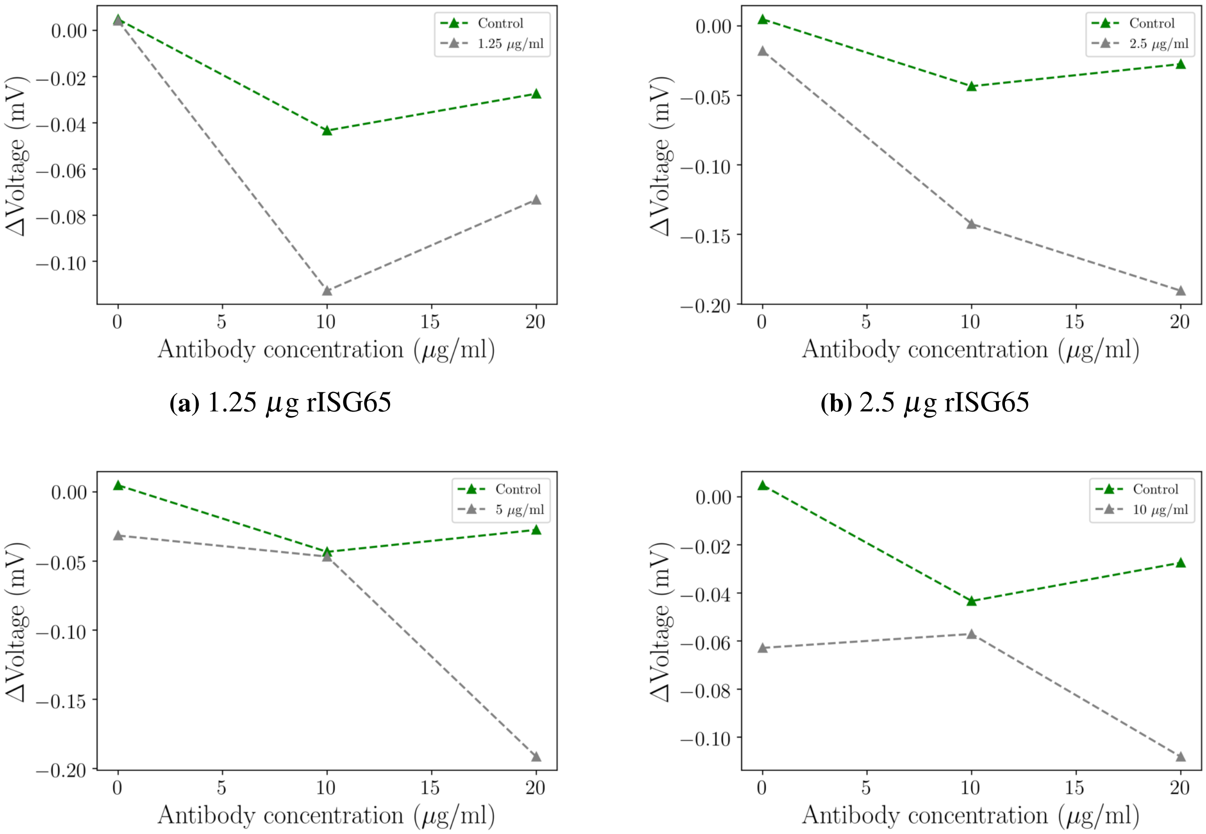
(b) 2.5 μg rISG65

(c) 5 μg rISG65

(d) 10 μg rISG65

Figure 4.11: Immunoassay at various antigen and primary antibody concentrations. The readings are recorded during the last stage of the immunoassay.

The final ELISA results after the final stage of the immunoassay was completed and the slides washed with dH_2O are presented in Figure 4.12. The R^2 values of the fitted linear model, LOD and LOQ were computed for each rISG65 concentration (Table 4.2). The highest R^2 value was obtained for the assays with 2.5 µg/ml rISG65 (Table 4.2). The limit of detection obtained for this rISG65 concentration from the slope of the calibration curve was 0.84 µg/ml and the limit of quantitation was 2.56 µg/ml. These were the smallest LOD and LOQ values obtained out of the four antigen concentrations tested. Based on these results, the developed immunoassay worked optimally only for the antigen concentration of 2.5 μg/ml.



(a) 1.25 μg rISG65

(b) 2.5 μg rISG65

(c) 5 μg rISG65

(d) 10 μg rISG65

Figure 4.12: Immunoassay at various antigen and primary antibody concentrations. The readings are recorded after the slides were washed and dried following the final immunoassay stage.

rISG65 concentration (μg/ml)	1.25	2.5	5	10
R^2	0.65	0.97	0.91	0.81
LOD (μg/ml anti-ISG65)	3.84	0.84	1.54	2.39
LOQ (μg/ml anti-ISG65)	11.64	2.56	4.68	7.25

Table 4.2: Determining the LOD and LOQ based on the calibration curves.

4.4 Discussion

In this section the results obtained following the development of the recombinant proteins and those obtained after the running the developed immunoassay for the biosensor are discussed, whilst taking into consideration the limitations of the experiments performed. Overall, the recombinant proteins of interest were expressed in *E.coli*, but *E.coli* specific contaminants were also co-purified. Taking into consideration that the biosensor platform was a proof-of-concept experiment and that samples tested on the platform were unlikely to contain antibodies to *E.coli*, the analysis was continued using the purified recombinant protein with the highest expression level, i.e. rISG65. Although, the experiment was initially aiming to use all four recombinant antigens on the platform for multiplex detecting, the developed immmunoassay protocol was first tested using only one recombinant protein. The immunoassay protocol proved to be successful when the surface was coated with lower antigen concentrations (2.5 µg/ml). In comparison to [5] where a LOD of 10 µg/ml was obtained, the immunoassay developed in this chapter performed better as suggested by the obtained LOD of 0.84 µg/ml. This could stem from several factors such as the different type of chemical surface functionalisation and coating antigen concentration. However, several impediments were also identified, which mainly stemmed from working with a small surface area. These could be remedied by using, for example, inkjet printing or microcontact printing method coupled with a microfluidics systems when multiplexing [109–111].

4.4.1 Protein expression and purification

For this protein purification experiment pET-28(+) was chosen as it provides kanamycin resistance (kanamycin works by blocking the protein synthesis at mRNA level) and it also provides an N-terminal His Tag/ thrombin/T7 Tag configuration (N-terminally 6xHis-tagged proteins with a thrombin site) to recombinant proteins. The hexa-histidine tagged proteins can be purified using a relatively simple protocol using immobilized metal affinity chromatography. Also, the hexa-histidine tags are small and usually do not affect the solubility characteristics of the protein.

Regarding the competent cells used for transformation, *E.coli* B21(DE3) cells were selected mainly due to their high levels of protein expression caused by the T7 RNA polymerase-IPTG induction system. Addition of IPTG in the culture medium leads to the release of the lac gene repressor and subsequent expression of T7 RNA Polymerase. This initiates the transcription from the T7 promoter present in the vector, thus allowing the expression of any foreign gene cloned downstream to this promoter [112].

One feature of the T7 system is that many recombinant proteins precipitate or fold incorrectly exposing them to protease degradation when expressed at 37°C, but are soluble and fold correctly when the temperature during induction is 15–25°C. This might occur due to the fact that slower rates of protein production allow more time for the translated recombinant proteins to fold properly. This was seen in my work where solubility for the recombinant antigens (mainly ISG75 and VSG1.5) was adversely affected at 37°C induction. Thus, induction at lower temperatures was used in this project. Previous studies used induction at 22°C for 2 days for ISG65 and ISG75 which were also expressed in *E.coli* vector [100].

Overall, the protein purification experiment led to partially impure recombinant proteins (rISG65, rISG75, rL1.3, rL1.5) being expressed. The possible factors affecting this are explained next. The protein purification experiment may be affected by several factors such as poor bacterial cell lysis, failure of the tagged protein to bind to the chromatography column, the wrong protein being expressed or co-purification with the bacterial proteins [105]. In this study, the main problem with the protein purification stems from the co-purification with the bacterial proteins. This generally happens with proteins expressed in *E. coli* when the expression level of the recombinant protein is low [105]. Contaminants usually include proteins with multiple histidine residues or molecular chaperones that bind directly to the resin or to the recombinant protein. Such molecular chaperones include GroEl, DnaK, DnaJ as described in the results above are candidates for contaminants here. In such cases, either additional chromatography is required or expression in an alternative expression system, i.e different bacterial vector or different expression host. For example, in the case of the variant surface glycoproteins successful protein expression was achieved in *L. mexicana* and *P. pastoris* [94, 99].

Because of the contaminated background of the purified proteins, the ELISA performed on human serum samples from endemic region resulted in a high number of false positives, in comparison to the native antigens. The recombinant variant surface glycoproteins also showed lower reactivity than the invariant surface glycoproteins.

Since it was anticipated that the human serum from infected patients was considered unlikely to have antibodies to *E. coli* proteins, testing the chip based format and assessing the utility of the imperfectly pure proteins was still carried out. Moreover, based on the results which indicated a better purification of rISG65, testing the chip based format was done using the obtained rISG65.

4.4.2 Bionsensor immunoassay

The aim of this chapter was initially to develop a multiplex immunoassay for the detection of fever-associated diseases. However, due to limited resources and issues which were encountered during the experimental laboratory work, the biosensor was developed only using one recombinant antigen. The results obtained using the immunosensor developed using rISG65 are discussed below and a possible reason for the the optimal results obtained with 2.5 µg/ml rISG65 are explained.

Based on the obtained results, the lowest LOD was obtained at 2.5 µg/ml. This could be explained by the different factors affecting antigen immobilisation, one such factor being the number of rISG65 molecules which could be optimally immobilised on to the surface of the well. As previously mentioned, high concentration of antigens could increase stacking or layering, which may allow less stable interactions of subsequent reactions with antibody. Although some proteins are elongated, most proteins, including ISG65, fold into globular domains, their interior consisting of protein subunits and domains with closely packed atoms, meaning there are no substantial holes and almost no water molecules in the interior of the protein [113]. Consequently, proteins are quite rigid structures. Also, proteins have a similar density of 1.37 g/cm^3 [114]. Assuming the rISG65 protein is a sphere, the radius size of a 50kDa molecule is approximately 2.4 nm with a surface are of 72.38 nm^2 ($A_{sphere} = 4\pi r^2$) [113].

Therefore, on a surface area of 2.56 mm^2, i.e. the surface area of the PMMA well, a maximum of 3.54×10^{10} protein molecules could be immobilised on one layer with no space in-between the molecules. Next, it was calculated how many protein molecules would be in the different concentrations of rISG65 solution used for the assays. It is known that in 1 mole of a substance there are $6 * 10^{23}$ molecules (Avogadro's number). Therefore the number of molecules in a solution with a known volume and concentration is:

$$n_{molecules} = n_{moles} \times 6 \times 10^{23} = \frac{m}{\mu} \times 6 \times 10^{23} = \frac{V \times c}{\mu} \times 6 \times 10^{23} \tag{4.3}$$

Where: $n_{molecules}$ = number of molecules; n_{moles} = number of moles; µ= molecular mass of the substance of interest (rISG65); m = mass of rISG65; c = concentration of rISG in solution with volume V ($c = \frac{m}{V}$).

Since the rISG65 molecule is approximately 50 kDa because I used a truncated version of the protein with optimal antigenicity, this equals to a molecular mass of 5×10^4 g/mol. A solution with a concentration of 10 µg/ml would therefore have 12×10^{10} molecules in 1 µl, which is approximately 3.4 times more than the number of molecules which would fit in one layer on the surface. A solution of 5 µg/ml would have 6×10^{10} molecules in 1 µl which is around 1.7 times more than the number of molecules which would fit in one layer on the surface. Finally for a concentration of 2.5 µg/ml which presented the highest R^2 value after linear fitting, there would be 3×10^{10} molecules, which is by 1.1 times less than the number of molecules which would fit in one layer on the surface. This could be an explanation for the better results obtained for this concentration. For a concentration of 1.25 µg/ml there would be 1.5×10^{10} molecules, by 2.4 times less than the maximum molecule occupancy, which means reactivity of antigen was not satisfied at its maximum capacity. The results in terms of theoretical percentage of molecular occupancy are presented in Table 4.3 below. In conclusion, these results confirm the fact that a concentration of 2.5 µg/ml rISG65 used for the coating step of the assay worked the best out of the 4 concentrations, due to the fact molecules could be stacked in one layer on the surface of the PMMA well at a molecular occupancy of 85%. For future improvements of the assay, further antibody concentrations should be tested.

ISG65 Concentration (µg/ml)	Molecule occupancy (%)
1.25	42.37
2.5	84.75
5	169.49
10	338.98

Table 4.3: The molecule occupancy of the well for the different concentrations solutions of ISG65.

4.4.2.1 Factors affecting the immunoassay and suggestions for improvement

Several factors affecting the immunoassay were identified. First of all, the surface of the PMMA well may have affected the assay because of its rough structure caused by the CO_2 laser. Since the surface was not uniform and smooth this could have resulted into variable buffer deposition. Additional factors which might have affected the immunoassay are the type of surface functionalisation, incubation times and temperature. The chemical surface functionalisation might affect the reproducibility of the immunoassays. Wet chemical functionalisation of PMMA slides is a cost-effective and relatively simple procedure. However, it does present certain limitations. First of all, it could cause irregular surface etching and, secondly, stability of chemically modified surfaces could be compromised. Plasma treatment is an alternative to functionalisation of PMMA surfaces [115]. It is recommended, however, that if further chemical treatment is required this should be done within 1 h of the plasma treatment [85]. Alternatively, other chemicals such as polyethylene imine (PEI) could also be used instead of APTES [89]. The repeated washing steps with varying pressure levels involved in the ELISA protocol might also affect the immunoassay. These limitations could be overcome by integrating the chip with a microfluidics system together with microcontact printing techniques for multiplex coating of antigens [116].

Antigen immobilisation techniques Another aspect of the immunoassay which could be improved is the antigen immobilisation step. For this study, the antigen was air-dried on the PMMA well. The antigen solution contained glycerol, a chemical additive which prevents protein denaturation and chemical instabilities [117]. Air drying techniques have been previously tested and proven to be successful [118–120]. Other additives could be better in improving protein stability, such as the sugars sucrose and trehalose. The removal of water during the drying process should lead to a tight contact of the antigen to the surface. For an improved immunoassay, antigen immobilisation on smooth PMMA surface is recommended. Moreover, since the original aim of the assay was to multiplex it, inkjet or microcontact printing techniques should be taken into consideration. The principle behind the multiplex printing would be similar to that underlying microarrays for which inkjet printing is generally used. Inkjet printing techniques could be tried for a more accurate immobilisation of multiple antigens/antibodies [110]. Inkjet printing permits the deposition of tiny droplets ($\geq$ 1 pl) onto a substrate (glass, plastic, etc.). The advantages of inkjet printing are the absence of physical contact between the printhead and the printed substrate and the absence of a mask for patterning. To develop a printable ink, the viscosity and surface tension of the liquid should also be taken into consideration. For increasing the surface tension, different compounds such as surfactant Tween 80 can be added to the ink. Proteins, single-stranded DNA oligomers and human cells have been inkjet printed [111]. For example, in [33] for the printing process on a CMOS-based immunosensor, they used a Sciflex S5 printer and the print media were 3.3 μM antibody solution in 20 mM phosphate buffered saline solution. Another technique used as an alternative to inkjet printing for depositing proteins is microcontact printing [121]. This technique is more flexible in terms of the rheology requirements and it provides higher resolution than inkjet printing. For this method a stamp (usually made from PDMS) is dipped into the “ink” containing the proteins and it is then brought into contact with the silicon surface [111].

Implementing a microfluidics system Another method for improving the immunoassay procedure on the chip would be to integrate the chip with a microfluidics system and, instead of using silver staining method for the detection of the antigen-antibody reaction, other more sensitive reactions such as colorimetric reaction with horseradish peroxidase substrate could be used for detection [116]. A microfluidics system would also improve the delivery of the reagents during the assay and compactness of the entire system. This would ensure reproducibility and would decrease the time needed to perform the assay. The fluids are transported through the microchannels via laminar flow which ensures that molecules are transported in a relatively predictable manner without any turbuluence. Several studies used microfluidics techniques to develop a faster and more efficient assay. For examples, [109] used PMMA surface combined with microfluidics system to develop and a 2-hours immunoassay.

4.4.3 Conclusion

In conclusion, the immunosensor approach presented in this chapter could be used for the detection of HAT. Although there is a clear difference between the negative controls and positive controls, further optimisation of the immunoassay protocol would be required. Based on the LOD results, the immunoassay with the PMMA surface coated with 2.5 µg/ml rISG65 worked best at detecting HAT specific antibodies. However, due to the experimental issues encountered during the laboratory work and delays in chip provision from external source, we chose to suspend the biosensor work, which became terminal with the onset of the first Covid-19 pandemic related lockdown where laboratories were closed. Instead, my focus was directed towards metabolomics approaches to fever-associated diseases that aimed to seek potentially diagnostic biomarkers and also molecular mechanisms underlying disease. These are described in the next chapter.

Chapter 5

Developing a method for the alignment of multiple disparate metabolomics datasets

5.1 Introduction

Recently, increasing numbers of studies on fever-associated diseases using mass spectrometry coupled with untargeted high performance liquid chromatography (LC-MS) have emerged [122–126]. As presented in Chapter 2 Section 3, LC-MS is one of the most sensitive methods for identifying metabolite markers and for providing a comprehensive coverage of the metabolome, as it enables the separation and measurement of thousands of discrete chemical compounds [44]. When analysis of LC-MS data is performed on datasets studying individual disease states, however, it is not possible to distinguish between metabolites associated generically with fever, and others specific for particular diseases. Thus, by integrating multiple disparate LC-MS datasets associated with fever, it was aimed to simultaneously search for common perturbations to compounds across a set of fever associated diseases, in order to identify metabolites generically associated with fever or disease severity. In order to achieve this, an algorithm for the integration of multiple LC-MS datasets through peakset alignment was proposed.

LC-MS peak alignment challenges The alignment process, also referred to as correspondence, for which algorithms are categorised into either direct matching or warping algorithms, has been extensively studied, mainly in the context of multiple injections within the same experiment [69]; these were detailed in Chapter 2 Section 3. This is due to the fact that although the LC-MS instrumentation and methodology are robust and well established, measurement variability can still appear, resulting in non-linear shifts especially in retention time (RT). Improper alignment could lead to mis-alignment or cross-alignment issues. Mis-alignment occurs when peaks fail to align together forming *split* peaks, i.e. one peakset is being treated as different peaksets (features), when the peaks being aligned have a RT drift larger than a set RT tolerance window. Whereas, cross-alignment occurs when peaks from different compounds align [127]. For example, isobaric peaks could cross-align and erroneously form one feature. In this case, isobaric compounds refer to compounds with the same nominal mass, i.e. the sum of integer masses of protons and neutrons of a chemical species [128], but with different exact mass and chemical formula. The risk of mis-alignment or cross-alignment of isobaric features increases with increasing RT drifts between samples and a fixed RT tolerance window. Additionally, some metabolites are less stable than others, becoming more prone than others to exhibit intra or interbatch shifts in RT and exhibiting a higher risk of mis-alignment [127].

There are two types of variability sources which can appear in an LC-MS experiment and cause mis-alignment: system variation and component level variation [69]. The system variation is usually consistent throughout the whole run and may be caused by factors such as the apparatus itself, the column, system stability and temperature [129]. Whereas, component level variation, by contrast, is specific to a single analyte or a group of analytes, so it cannot be modelled using monotonic functions. In order to reduce the RT variability within an experiment, especially the systematic ones, a set of known metabolites, or standard reference mixture (SRM), is run at various points during an LC-MS experiment as part of the quality control process [129, 130]. Using SRMs enables the drift in compound intensity and especially RT to be tracked throughout every LC-MS run. Therefore, the SRM metabolites can be used as landmarks for retention time correction as the m/z for each compound detected using LC-MS is constant.

In this chapter, a method for the integration of three disparate LC-MS datasets is proposed. The information provided by the SRM runs was used to determine the RT drift between injections of different LC-MS experiments. This was then modelled using Gaussian Process (GP) regressions [37] which were described in detail in Chapter 2 Section 2. GPs are a non-parametric approach to modelling data and they differ from standard regression models in that they do not require any assumptions about a particular parametric form for the function being modelled. The fitted GPs were used to perform a high-level correction of retention times between experiments, after which standard direct matching alignment can be performed.

GP regressions are flexible non-parametric modelling methods which can be optimally applied to small sized data-sets. In summary, the prior distribution of a GP regression is defined by a mean function, which is usually set to 0, and a covariance function, also known as a kernel (2.3). The kernel is the one which describes the relationship between the output functions of the model. There are multiple types of kernels, such as stationary kernels (RBF) and non-stationary (MLP). Composite kernels can also be used and these are obtained either through addition or multiplication of single kernels; this allows to incorporate as much high-level structure as necessary into the model. The shape of the kernel is also determined by its hyper-parameters (variance, length-scale) which can be optimised using methods such as Bayesian optimisation. In this chapter both single and composite kernels were used for the GP models which were tested and selected using cross-validation. These were implemented and optimised using GPy python package [40]. The optimisation of the hyper-parameters in the GPy package is done via maximisation of the marginal likelihood [40].

The algorithm was developed here specifically to determine whether particular metabolites could be identified that changed in abundance in similar ways across a series of distinct fever-associated diseases. These include Zika virus infection in patients from Ecuador [131], Leishmaniasis patients from Spain [132] and uncomplicated malaria infected volunteers from the UK [133]. The samples had all been run previously using the same LC-MS platform (Glasgow Polyomics, University of Glasgow, UK). By seeking metabolites whose variation in abundance followed common trends in different datasets we aimed to determine disease-generic metabolites that could both assist in understanding the pathophysiology of infectious disease, and also highlight metabolites that were found to have a change in abundance in individual studies that may be fever rather than specific disease related.

In conclusion, this chapter aimed to develop a method for integrating multiple disparate metabolomics LC-MS datasets and identify common and specific disease perturbation across data sets by using GP regressions for correcting the RT drift between datasets.

5.2 Materials and methods

5.2.1 Datasets

Three LC-MS datasets (D_Z, D_M, D_{VL}) analysed at Glasgow Polyomics metabolomics facility (University of Glasgow, UK) which studied different infections with pathogens associated with febrile disease were used for the cross-experimental integration in this study. These are listed below:

1. D_Z: In this dataset the blood serum from patients with Zika virus disease and healthy controls was analysed as part of a case-control experiment [131].
2. D_M: In this dataset the blood serum from patients with malaria and healthy controls was analysed. This was an intervention study, where healthy controls were infected with malaria and thus disease-specific symptoms were closely controlled [133].
3. D_{VL}: In this dataset the blood serum from patients with Visceral Leishmaniasis and healthy controls was analysed as part of a case-control experiment [132].

All three experiments were designed for detecting the differences between the serum metabolic profiles of healthy controls and infected patients diagnosed by gold-standard methods. In total, there were 74 samples (37 controls and 37 disease samples). Detailed information about each LC-MS experiment is presented in Table 5.1.

Dataset	D_Z	D_M	D_{VL}
Infectious Disease	Zika virus	Malaria	Visceral Leishmaniasis
Study Type	Case-control	Intervention	Case-control
LC Column/MS Platform	pHILIC/Q-Exactive	pHILIC/Q-Exactive	pHILIC/Q-Exactive
LC-MS Run Length (min)	26	26	46
Date Analysed	2018	2016	2018
Healthy Controls	10	7	20
Infected Patients	10	7	20
SRM Sets	3	3	3
MS2 data	Yes	Yes	Yes

Table 5.1: LC-MS datasets details: Information about each LC-MS experiment including the disease studied, type of study, number of controls and patients, LC-MS platform used and date when the samples from the LC-MS experiment were run. Fragmentation (MS2) data was also available for all datasets.

5.2.1.1 Data curation

For D$_{VL}$, plasma was taken from 20 adult patients from Fuenlabrada (Madrid, Spain) diagnosed with VL at the Hospital Universitario de Fuenlabrada between January 2013 and June 2015.

Blood samples were collected during the period of active disease and infection by *Leishmania infantum* was confirmed by Leishmania specific nested PCR; the presence of Leishmania-specific plasma antibodies were determined by rK39 immunochromatographic test (Inbios, USA) and indirect immunofluorecence test. Data obtained were compared with those of 20 matched healthy controls obtained from volunteers at the Blood Bank of the Hospital Universitario de Fuenlabrada [131]. For D_Z, patients included in the Zika virus group had presented to hospital seeking assistance for febrile symptoms confirmed as Zika virus infection by PCR (n=10). The healthy control group was composed of women attending their routine prenatal care (n=10). Samples from all participants were collected by using a red cap vacuum blood tube of 4 mL, without clot activator. After blood collection, the tubes were maintained in the rack, at room temperature for 30 minutes. After this time, samples were maintained at 4°C until centrifugation (which was performed in less than 5 hours). Centrifugation was carried out at 2000 x g for 10 minutes. Serum samples were then placed in 1.5 microtubes and stored at -80°C until metabolites extraction [132]. Regarding D_M, the malaria patients are described in detail in [133].

5.2.1.2 LC-MS platform

The experiments were performed at different time points (Table 5.1) using the same LC-MS platform: Thermo Orbitrap QExactive (Thermo Fisher Scientific) mass spectrometer coupled with a Dionex UltiMate 3000 RSLC system (Thermo Fisher Scientific, Hemel Hempstead, UK) using a ZIC-pHILIC column. In all three experiments, the column was maintained at 30°C and samples were eluted with a linear gradient at a flow rate of 0.3 ml/min. While the same flow rate was used for all three datasets, the length of the run differed for D_{VL}, which lasted longer than the other two datasets. For the MS analysis, the Orbitrap Q-Exactive mass spectrometer was operated using the following settings: resolution 70,000, m/z range 70-1050, sheath gas 40, auxiliary gas 5, sweep gas 1, probe temperature 150°C, capillary temperature 320°C. At the end of the mass spectrometry analysis .mzXML files were obtained.

5.2.1.3 Tandem mass spectrometry data

Fragmentation (MS2) analysis was also performed for all three experiments using higher energy C-trap dissociation (HCD) fragmentation of the pooled samples at a normalised collision energy (NCE) of 60. At the end of the MS2 analysis .mzXML files were obtained.

5.2.1.4 Standard reference mixtures

As part of the quality control across each experiment, three sets of standard reference mixtures (SRM) which include compounds that cover a broad range of metabolic pathways such as amino acid metabolism, central carbon metabolism and nucleotide metabolism were run twice, before and after the cohort of samples was run. Two types of SRM files were included in the dataset. The first type of files are the raw .mzXML files obtained from the MS analysis. The second ones are .csv files generated with ToxID software from the raw files where, for each SRM metabolite the following information is given: compound name, chemical formula, polarity, detected mass-to-charge ratio, delta (ppm), expected retention time and actual retention time. The total number of SRM metabolites detected in +ve ESI mode in each set is provided in Table 5.2 below.

	Set 1		Set 2		Set 3	
Dataset	Rep.1	Rep.2	Rep.1	Rep.2	Rep.1	Rep.2
D_Z	37	38	44	42	13	13
D_M	47	37	50	45	15	14
D_{VL}	35	36	44	43	14	14

Table 5.2: Total number of SRM metabolites (+ve ESI mode) identified in the ToxID generated files.

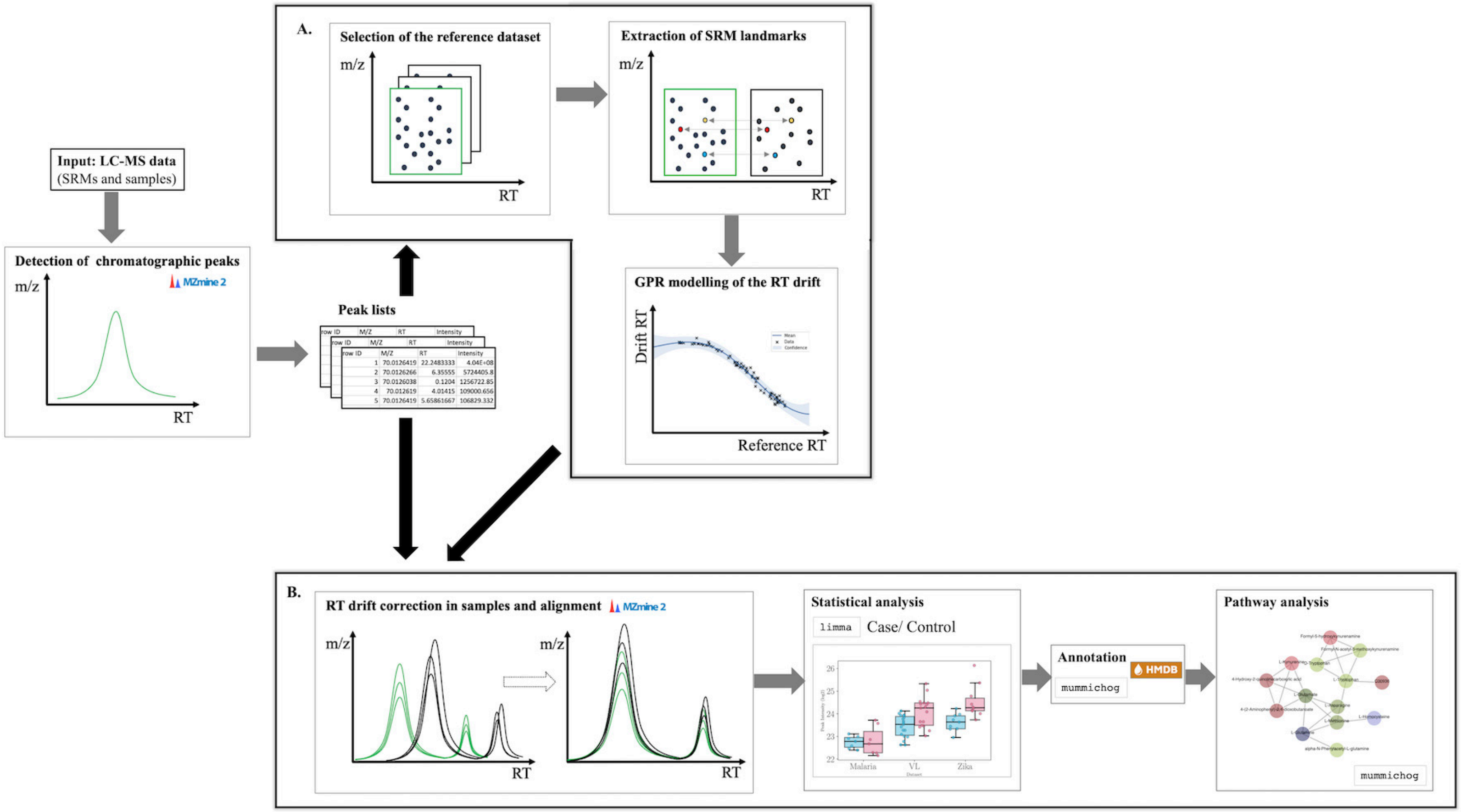
5.2.2 Study workflow

The data analysis process is outlined in the diagram in Figure 5.1.

Quality Control The total intensity of all ions at each time point is known as the total ion current chromatogram (TIC). The TICs were used as a quality control for each of the samples in the dataset. They were compared and checked in order to determine whether any of the samples needed to be removed from the analysis. The samples with a flat chromatogram throughout the entire run were deemed faulty, and thus, removed from further analysis. Three samples were removed from D_{VL} (VL6, VL6_r, VL6_r2) and one sample was removed from D_Z (C6).

5.2.3 Peak detection

The processing of the spectral data begins with the detection of the chromatographic peaks from the input SRM and samples LC-MS data. Peak detection was performed using the wavelet transform method from MZmine2 v2.40.1 [62] in batch mode. MZmine2 is an open-source Java software developed for mass spectrometry data processing. Some of the processes it performs include: raw input file manipulation, peak detection, chromatographic alignment, normalisation, visualization and data export [62]. The usual workflow for peak detection for both MS and MS2 data begins with mass detection followed by chromatogram building and deconvolution as detailed in Chapter 2. The parameter values used for each step of MZmine peak detection software are detailed in the list below. It is to be noted that for the chromatogram deconvolution step different values were used for each separate dataset by using the preview mode.



row ID	M/Z	RT	Intensity
1	70.0126419	22.2483333	4.04E+08
2	70.0126266	6.35555	5724405.8
3	70.0126038	0.1204	1256722.85
4	70.012619	4.01415	109000.656
5	70.0126419	5.65861667	106829.332

Figure 5.1: Diagram representing the study workflow.Peaks are detected from input LC-MS data including the SRMs and samples using MZmine2 and peakset lists containing ion information (m/z, RT, intensity) are obtained. A. SRM analysis. A reference dataset is selected and the RT drift in the other datasets is calculated and modelled using GP regression. B. Samples analysis. Based on the GPR models obtained for each dataset the RT is corrected in each peakset list and alignment is done using MZmine2. Afterwards, statistical analysis focused on the intensity differences between the control and infected samples is performed using limma R package. This is followed by annotation and pathway analysis using *mummichog*.

1. Mass detection:
    - RT: 0.00 – 26.02 min (46.00 min for D_{VL})
    - MS level: 1/2 (for MS2 data)
    - Polarity: +
    - Spectrum type: centroided
    - Mass detector: Centroid (Noise level: 1.0E4)
2. ADAP chromatogram builder:
    - Group intensity threshold: 1E4
    - Min. highest intensity: Group intensity threshold $\times 10 = 1E5$
    - Min group size (number of scans): 5
    - m/z tolerance (depends on the instruments used): 3ppm or 0.001 m/z
3. Chromatogram deconvolution (peak detection):
    - S/N threshold: 3 (limit of detection) - 5 (limit of quantitation)
    - Coefficient area: 10 (using preview mode)
    - Peak duration range: 0.03 - 1.00 (1.5, for detecting more peaks towards the end)
    - RT wavelet range (v. sensitive): 0.01 - 0.40
4. File export:
    - The peak picked files were exported to both csv and mzTab formats.

Some parameters are instrument specific, such as the m/z tolerance used for the ADAP chromatogram builder, whereas others where chosen by using the preview mode available in MZmine2 and the raw .mzXML standards files by comparing the number of total SRM metabolites obtained after peak detection with the ones in Table 5.2 obtained from ToxID. With the values presented in Table 5.2 the closest match in total number of standards was obtained (Table 5.3). The peak list files obtained following peak detection contained information about the retention time (RT), mass-to-charge ratio (m/z) and intensity for each peak. Additionally for MS2 data, an .mgf file was generated containing information on the fragments intensity and m/z and their MS precursor.

	Set 1		Set 2		Set 3	
Dataset	Rep.1	Rep.2	Rep.1	Rep.2	Rep.1	Rep.2
D_Z	37	38	42	35	10	11
D_M	47	37	46	44	13	14
D_{VL}	32	35	40	33	12	12

Table 5.3: Total number of SRM metabolites identified in the MZmine2 processed files.

5.2.4 Alignment MZmine2 workflow

MZmine2 was also used for peak lists alignment. Any aligned peak across the samples is defined as a *peakset*. The simplest alignment algorithm employed by MZmine2 is the Join Aligner algorithm which works as follows: the first peak list file from the list of peak list files (S_i) is set to be the master of the peak list files which will be matched against every other peak list file (S_j). For each sample file S_j, the algorithm iterates through the rows of S_i and for each row it looks for peaks which are within the pre-set retention time alignment window ($RTWindow$) and m/z alignment window ($MZWindow$). For each match, a score is computed using Eq. 5.1 and the pair getting the highest score is aligned to S_i.

$$score = (1 - \frac{MZDifference}{MZWindow}) \times MZWeight + (1 - \frac{RTDifference}{RTWindow}) \times RTWeight \tag{5.1}$$

$MZWindow$, $RTWindow$, $MZWeight$ and $RTWeight$ were all manually set. $MZWeight$ was set to 75, $RTWeight$ was set to 25 and $MZWindow$ was set to 0.01, or 10 ppm. The value for $RTWindow$ was selected later in the analysis, after the retention time for each dataset was processed as described in the next section.

5.2.5 SRMs analysis

First, the reference dataset –D_Z– was randomly selected out of the two datasets with the shorter run length. Next, the SRM metabolites were extracted from the peak lists for each dataset, creating a profile for each dataset characterised by (m/z, RT) of each of the SRMs. The profiles were then mapped to the reference dataset profile and for each metabolite the RT drift from the mapped SRMs was determined and modelled using GP regression.

5.2.5.1 GPR modelling of the RT drift

The RTs from the (m/z, RT) profile created for each non-reference dataset (input observations) were regressed against their respective RT drift from the reference dataset values (observed values). In order to obtain a closer fit to the data but still maintain the variability, SRM metabolites outliers were removed based on their RT drift from the reference profile using a z-score cutoff value of 2. Model hyperparameter optimisation was done using multiple restarts (n=10) with the GPy optimiser in order to avoid local minima. In order to determine which covariance function aids in fitting the data best in each case, cross-validation was performed, by stratifying and splitting the SRM data in half for training and testing the model. Scikit python package was used to calculate the prediction accuracy score, mean absolute error (MAE) and mean squared error (MSE). For implementing the Gaussian Process models, the GPy python package, version 1.9.9 was used ([40]).

5.2.5.2 GPR corrected data

The GPR corrected RT times were obtained by adding the GPR predicted variables (posterior mean) to the initial RT values of each compound from the non-reference dataset.

5.2.5.3 Detection of alignment RT window parameter

The alignment of the peakset lists was performed using the JoinAligner module from MZmine2. The optimal RTWindow parameter was determined by aligning the SRM peak lists RTWindow values ranging from 0.01 min to 2 min. For each of the three SRM sets the total number of peaks aligned and the total number of SRM metabolites that align for each RTWindow value across the datasets is calculated. In order to determine the optimal RTWindow for all datasets, the value for which the alignment results in the lowest total number of peaks aligned and highest number of SRM metabolites is chosen for further analysis.

5.2.6 Samples analysis

The GPR models obtained in the previous stage were applied to each sample peak list by correcting the RT values for each peakset as detailed above. The lists were then aligned using the RTWindow value previously obtained. In order to reduce the data sparcity, the final peak list obtained was first processed by filtering out the peaks based on the percentage of missing values from each dataset. An arbitrary cut-off value of 50% was used, i.e. if a feature contains more than 50% missing values in one or more datasets, the respective feature is eliminated from further analysis.. Data imputation using k-nearest neighbours algorithm (KNN), where k=3, was performed solely for visualisation purposes [134]. KNN is an algorithm used for matching a point with its closest k-neighbours in a multi-dimensional space. It can be used for both discrete and continuous data. The assumption behind using KNN for missing values is that a value can be approximated by the values of the points that are closest to it, based on other variables.

5.2.7 Statistical analysis

The statistical analysis focused on the intensity differences between the sample peak lists belonging to the control and infected groups from all three datasets. The intensities were log2 normalised and modelled using linear regression included in the limma R package, where blocking was used to adjust for the intensity variability between the different datasets [135]. The output of this analysis is a list containing all the peaksets and their respective p-value, Benjamini-Hochberg (BH) adjusted p-value and logarithmic fold change (logFC) between the two conditions. The formula used for the linear regression is given below. The logFC values were calculated both for all samples from the meta-dataset, i.e. integrated logFC (logFC$_m$), and for the samples from each individual dataset, i.e. individual logFC (logFC$_i$).

$$Y_{ijk} = \alpha_j + x_i\beta_j + z_k\gamma_j + \varepsilon_{ijk} \tag{5.2}$$

Where: Y_{ijk}= response variable (intensity level of metabolite j in condition i and dataset k), α_j = intercept for metabolite j, x_i= first predictor variable: condition (infected/ control), β_j= estimated difference for metabolite j for each condition, z_k= second predictor variable: dataset, γ_j = the dataset effect for metabolite j, ε_{ijk} = error stochastic component, within group variation.

Based on the logFC$_i$ of each dataset obtained for each peakset it was determined whether the perturbation was common to all datasets, i.e. whether the logFC$_i$ were all either positive or negative, or specific to one dataset, i.e. logFC$_i$ for the dataset was opposite to the other datasets.

5.2.8 Feature annotation and pathway analysis

Fragmentation spectra from each dataset was aligned to the final filtered peak list. A profile characterized by (m/z, RT, ms2spec) was created for the possible adducts/fragments for each peakset (Table S2) and several methods were used for feature annotation. First, annotation was performed using the SRM information by mapping the peaks profile against the SRM (m/z, RT) profiles with an absolute m/z tolerance of 0.01. The SRM profiles were obtained prior to this study following analysis with ToxID software by Glasgow Polyomics; samples are identified by matching the mass spectra and retention times against entries in a library.

Next, the features were mapped against metabolite information extracted from the Human Metabolome Database (HMDB). HMDB contains comprehensive information on a high number of metabolites found in human sera. The metabolites and LC-MS experimental MS2 spectra database were downloaded from HMDB. Where one or more spectra was aligned to a peakset, the attached spectrum/spectra were compared against the experimental LC-MS spectra from HMDB and the match with the highest cosine similarity score was used to annotate the peak.

Where possible, the annotation obtained from the matched spectra were compared with the SRM annotation from ToxID in order to determine the cosine similarity threshold for separating a good annotation from a bad one.

For pathway and activity network analysis *mummichog* version 2.3.3. was used [136]. Mummichog is a software which implements a set of statistical algorithms that predict functional activity directly from measurements considered significant when compared to those in a reference sample [136]. It uses Kyoto Encyclopedia of Genes and Genomes (KEGG) to map the metabolic pathways. The parameters used for *mummichog* analysis were: -m positive/negative (ESI mode), -u 3 (instrument ppm tolerance), -c 0.05 (cutoff p-value used to select the significant list of features). The network modules obtained from the *mummichog* analysis were inputted into Cytoscape [137] to display the activity network.

5.2.9 Results visualisation

When comparing two different experimental groups, visualisation for metabolomics data results is usually performed using volcano plots, heatmaps, scatter plots, bar plots or boxplots [51]. In this case, results were represented using a box plots approach combined with scatter plots in order to visualise the mean level and standard deviation of the samples in each group from each dataset. This was due to the fact that there were 'batch' differences in the peakset intensities, and the analysis mainly focused on determining the direction of change in metabolites abundance for each of the dataset. Therefore, for each peakset a box plot was done for both conditions for each dataset, after imputing any missing values using KNN imputation.

5.2.10 Code

All of the analysis was performed in python programming language (https://github.com/anamaria-uofg/mma). The information about any given peak was stored in an object (peakinfo) with the following attributes: id, m/z, RT, p-val, t-val, logFC, *mummichog* annotation, *mummichog* pathway, *mummichog* kegg id, std annotation, std kegg id, spectra, adducts, best_ms2_match_adduct, ms2_annotation, ms2_kegg_id, intensities (from each sample). For the data processing performed in this project Python 3.7 was mainly used and only occasionally R 3.5. For the data pre-processing, the raw metabolomics .mzXML files were processed using MZmine 2.

5.2.11 Methodology evaluation

The peakset lists with no GPR correction were aligned and the same workflow was applied to their analysis. The results obtained were compared with the results of the GPR modified data.

Also, the datasets were individually analysed and their filtered peakset lists were then intersected to determine whether any commonality can be found in this way. Additionally, the alignment process was evaluated using the available MS2 data. If two compounds with similar m/z and RT break down into the same fragments (during LC-MS analysis), then it is highly likely they represent the same compound. Therefore, if a peakset has multiple highly similar MS2 spectra from different datasets, it is likely that the peaks were aligned correctly. In order to measure the similarity between two MS2 spectra, cosine similarity score implemented in mass-spec-utils was used [138]. In order to evaluate the alignment process using MS2 data the spectral similarity between the MS2 profiles was computed when more than one spectrum was aligned to one peakset. For each of the spectral similarity scores (good spectral similarity score), the mean of the corresponding distribution of random spectral similarity scores (bad spectral similarity score) was calculated, which were obtained from spectra of peaksets with similar m/z (absolute tolerance = 0.01), but different RT (difference larger than 40 s).

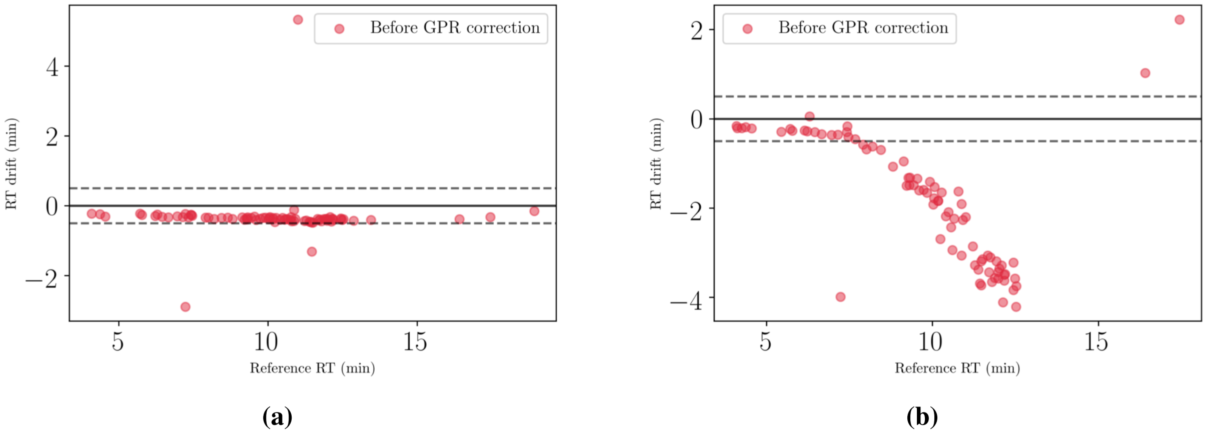
5.3 Results

5.3.1 SRM Analysis

Between D_Z and D_M, 77 SRM metabolites were found in common: 37 in SRM Set 1, 32 in SRM Set 2 and 8 in SRM Set 3. Between D_Z and D_{VL} 68 SRM metabolites were found in common: 32 in SRM Set 1, 27 in SRM Set 2 and 9 in SRM Set 3. In each case the mean retention time drift along with other statistics between each dataset and the reference dataset was calculated (Table 5.4). The reference RT was plotted against the RT drift for D_M and D_{VL} (Figure 5.2). For D_M the mean retention time drift in comparison to the reference dataset was 19.85 s and the highest retention time drift between two peaks belonging to the same ion was 319.74 s. In the case of D_{VL}, the mean retention time drift was 112.03 s and the maximum retention time drift was 252.54 s. Additionally, it may be observed that after 6 minutes the drift starts increasing exponentially with time, which might prove more challenging to model than the drift observed in D_M.

Dataset	Mean RT shift (s)	STD	Max. RT shift (s)
D_Z vs. D_M	-19.85	41.40	319.74
D_Z vs. D_{VL}	-112.03	84.78	252.54

Table 5.4: RT drift statistics between the reference dataset and D_M and D_{VL}



(a)

(b)

Figure 5.2: The RT drift (min) before the GPR correction in D_M (a) and D_{VL} (b) in comparison to the reference dataset D_Z. The straight line represents the point where no drift occurs between the datasets. The dotted lines represent a ±30 s interval allowed for the RT drift.

5.3.1.1 Correlation between the RT drift and other variables

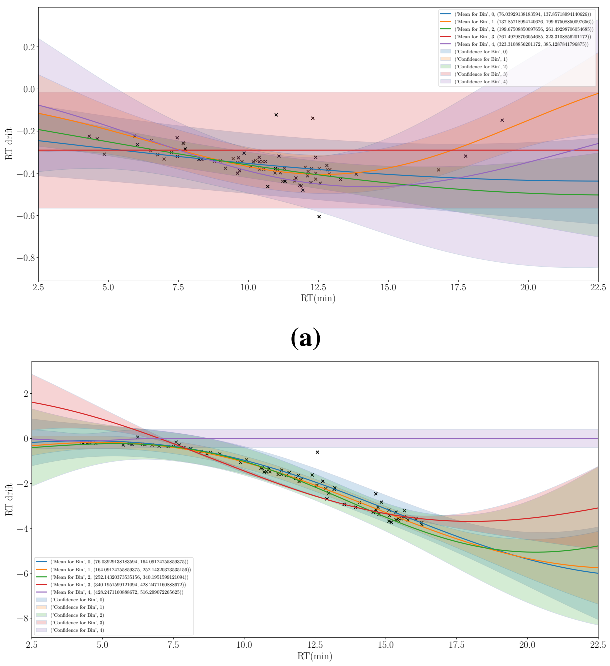
The SRM analysis for +ve ESI mode data is presented below. Correlation between the RT drift and the characteristics of the dataset profiles was checked and the highest correlation was found

	D_Z m/z	D_Z RT	D_M m/z	D_M RT	D_{VL} m/z	D_{VL} RT
RT Drift D_Z-D_M	-0.01	-0.04	-0.01	0.22	/	/
RT Drift D_Z-D_{VL}	-0.08	0.55	/	/	-0.08	0.796

Table 5.5: Correlation scores between the RT drift and (m/z, RT) of each dataset profile.

with the RT (Table 5.5). Based on this information, the training (70%) and test (30%) data were split and stratified in 4 equal length bins based on the RT. Following cross-validation, the kernel with the highest accuracy and lowest MAE was chosen. The final model was fitted using the selected kernel and optimised using multi-start in order to deal with possible bad local minimum.

The regression was also fitted to individual m/z based bins to determine whether there was any influence of m/z on the regression model (Figure 5.3). However, no difference between the models for specific m/z bins was observed.



(a)

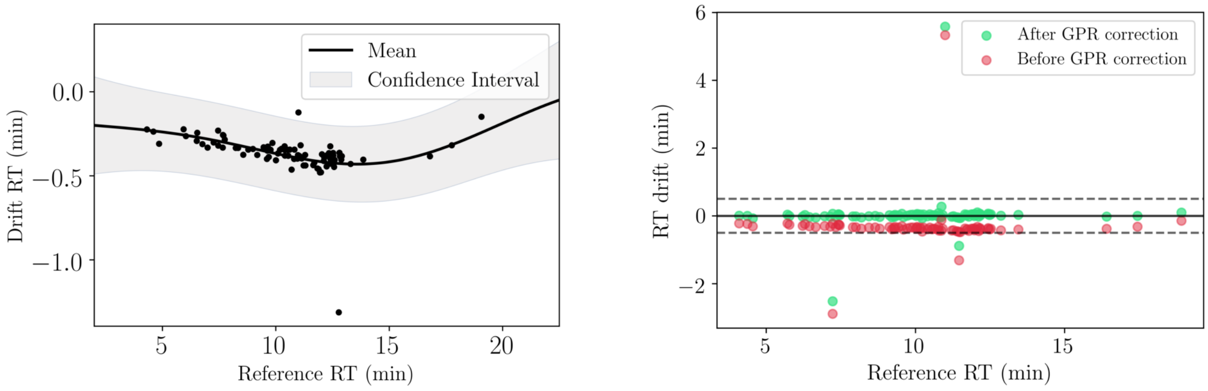
(b)

Figure 5.3: GP regression models fitted to individual m/z based bins for D_M (a) and D_{VL} (b).

5.3.1.2 GPR modelling of the RT drift

Several kernels, among which RBF, neural network and cosine kernels, were tested to determine which ones best fit the data. Composite kernels were also tested on the data. Following cross-validation an RBF kernel was selected as the best for fitting the RT drift in D_M with an accuracy of 0.99, MAE = 0.06 and MSE = 0.03. When fitted to the whole data (except the outliers) the final model had an accuracy score 0.93, MAE=0.14 and MSE = 0.46. Because of the small data size the confidence interval below 5 minutes and above 20 minutes starts increasing. The two hyperparameters of the RBF function, i.e. variance and lengthscale, were optimised to 0.07 and 6.84 respectively (Gaussian noise variance = 0.01) (Figure 3).

For the drift in D_{VL} a composite kernel RBF+MLP was chosen for fitting the data with an accuracy score of 0.995, MAE=0.14 and MSE = 0.03 (Figure 4). When fitted to the whole data (except the outliers) the final model had an accuracy score 0.89, MAE=0.27 and MSE = 0.74. The hyperparameters after 10 optimisation restarts: MLP variance = 3.99, MLP weight variance = 2.02e7, MLP bias variance = 5.56e-309, RBF variance = 7.44, RBF lengthscale = 9.41 (Gaussian noise variance = 0.1).



(a)

(b)

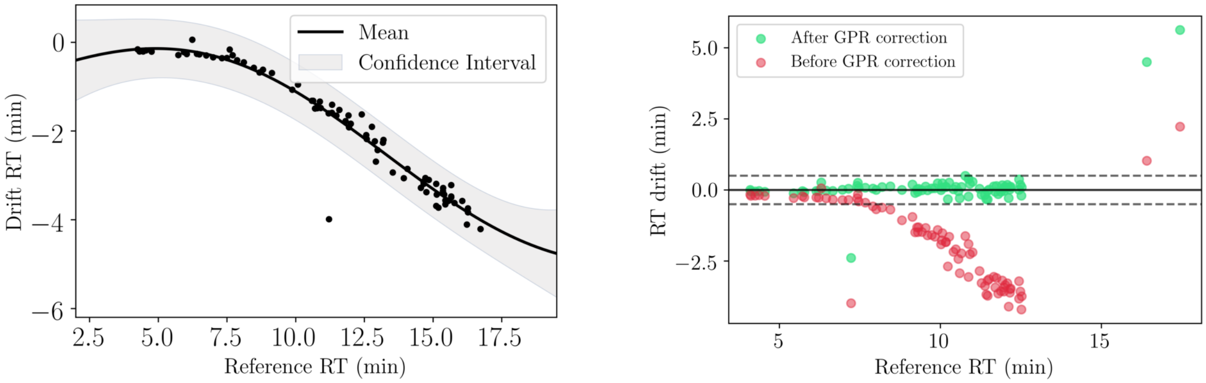
Figure 5.4: Modelling the drift in D_M using an RBF kernel. Mean and posterior predictive variance (confidence interval) of the GPR model with optimised hyperparameters. The plot on the right illustrates the RT drift in D_M before and after correction of the retention times using the GPR model.

RBF	Value	Constraints	Priors
Variance	0.07	+ve.	
Lengthscale	6.84	+ve	
Gaussian Noise Variance	0.01	+ve	

Table 5.6: Hyperparameter values of the GPR model using an RBF kernel which was used for modelling the RT drift in D_M

Sum	Value	Constraints	Priors
RBF Variance	7.44	+ve.	
RBF Lengthscale	9.41	+ve	
MLP Variance	3.99	+ve.	
MLP Weight Variance	2.02e7	+ve	
MLP Bias Variance	5.56e-309	+ve.	
Gaussian Noise Variance	0.1	+ve.	

Table 5.7: Hyperparameter values of the GPR model using a composite RBF+MLP kernel which was used for modelling the RT drift in D_{VL}



(a) (b)

Figure 5.5: Modelling the drift in D$_{VL}$ using a composite RBF + MLP kernel. The graph on the left: Mean and posterior predictive variance (confidence interval) of the GPR model with optimised hyperparameters: MLP variance = 3.99, MLP weight variance = 2.02e7, MLP bias variance = 5.56e-309, RBF variance = 7.44, RBF lengthscale = 9.41. The plot on the right illustrates the RT drift in D$_{VL}$ before and after correction of the retention times using the GPR model.

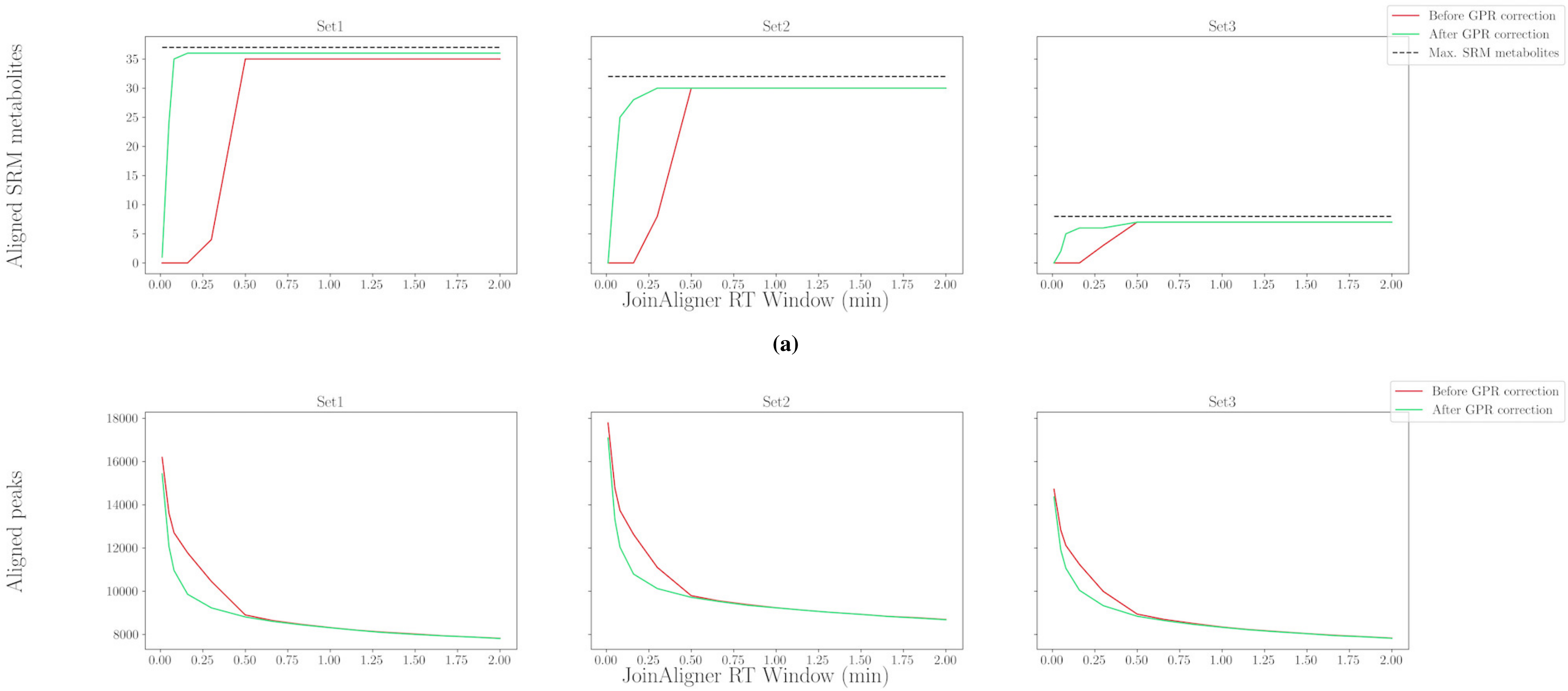
5.3.2 RTWindow value choice for JoinAligner module

For D$_M$, 94.8% of the maximum number of SRM metabolites in common with the reference dataset are aligned at RTWindow = 0.25 min after RT drift correction, as opposed to 19.48% before drift correction. Whereas, for D$_{VL}$ the maximum number of metabolites which are aligned after drift correction are obtained at RTWindow = 0.5 min for which 92.67% of the metabolites align, as opposed to 17.64% before correction. At a RTWindow value of 0.25 min 83.82% of the metabolites align (Figure 5.6, 5.7). Based on these results, the optimal RTWindow parameter for further alignment of the samples was selected RTWindow = 0.5 min. After correction, the number of total peaksets obtained following alignment also decreases, signifying better aligned peaks. Choosing the correct parameter for the RT window is important, because if the window is too small, then false negatives are introduced and, vice-versa, if the window is too big, then false positives are introduced.

5.3.3 Methodology evaluation

5.3.3.1 GPR corrected data vs original data

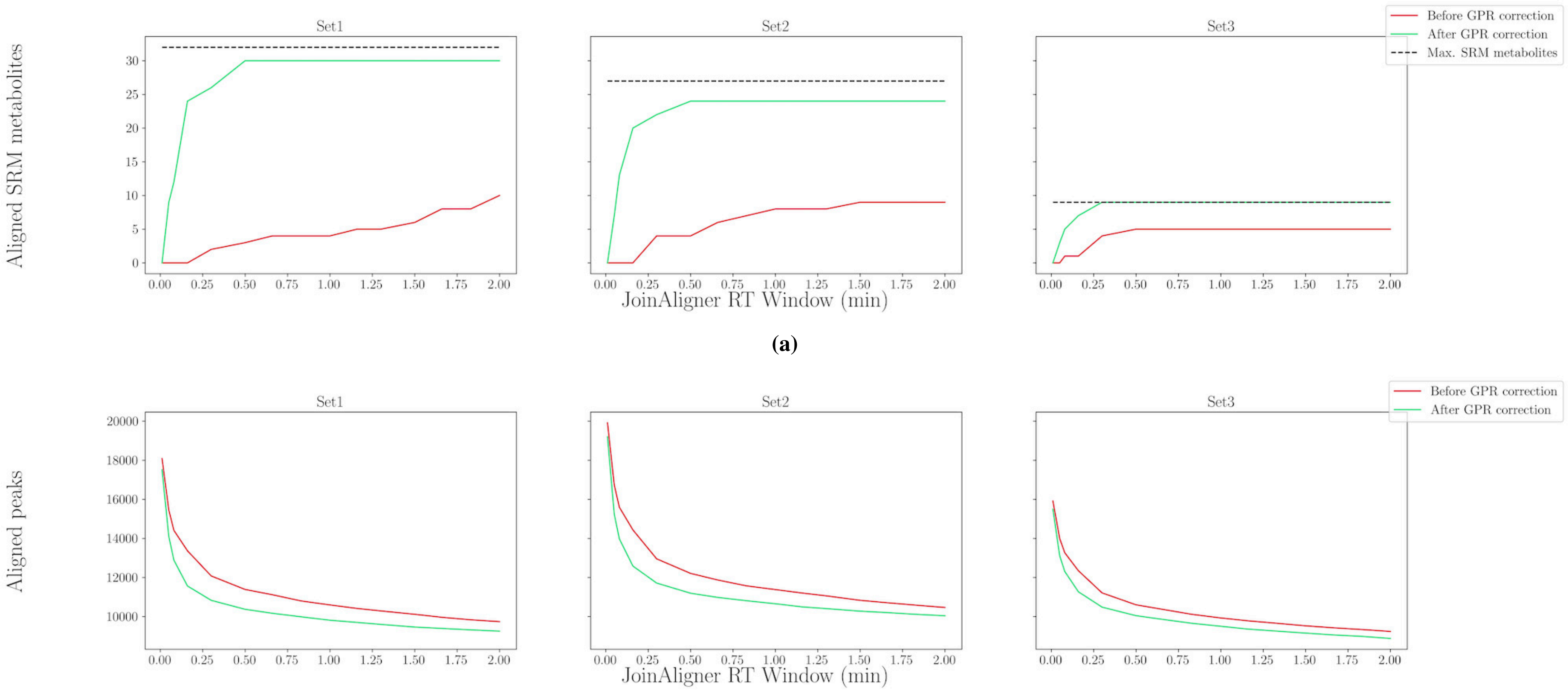
Results obtained for the +ve mode data are presented. In total, 625 peaksets remained after filtering of the GPR corrected data, as opposed to 344 peaksets in the original data (Table 5.8). Most of the peaks which are aligned only in the GPR modified data have the retention time in the range of [7,13] minutes (Figure 5.8).



(a)

(b)

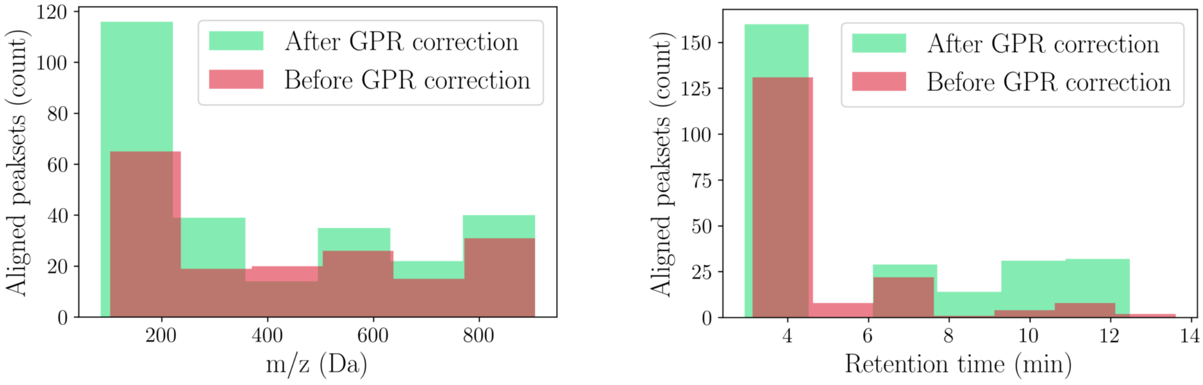
Figure 5.6: Aligned D_M standards before and after GP modification: The three SRM sample sets aligned between and before and after GP modification of SRM samples RT: a)Total number of SRM metabolites aligned at various values ranging from 0 to 2 minutes. The aim is to identify the lowest value for which the most SRM metabolites are aligned. In this case alignment, since the drift between the two datasets was somehow minimal when compared with the drift between non-modified and samples, and thus, the maximum number of SRMs is found after correction and alignment at a lower value of 0.25 min. However, after GP correction the maximum number of metabolites in common between the two datasets is found at 0.50 min. b)Total number of peaks aligned at various values ranging from 0 to 2 minutes.



(a)

(b)

Figure 5.7: Aligned D_{VL} standards before and after GP modification: The three SRM sample sets aligned between and before and after GP modification of SRM samples RT : orange = samples after GP modification, blue = samples before modification. a)Total number of SRM metabolites aligned at various values ranging from 0 to 2 minutes. The aim is to identify the lowest value for which the most SRM metabolites are aligned. In this case alignment between non-modified and samples containing the SRMs cannot find more than 10 metabolites in common for each mixture at any time point. However, after GP correction the maximum number of metabolites in common between the two datasets is found at 0.50 min. b)Total number of peaks aligned at various values ranging from 0 to 2 minutes.



(a) **(b)**

Figure 5.8: Peaks identified after alignment before (red) and after (green) GP modification. Peaks are being represented based on their m/z (a) and RT (b).

	Without GP	With GP	Difference
Sig. modified peak values	176	275	1.56
Total number of filtered peaks	344	604	1.98
Total number of peaks	38197	37220	0.97

Table 5.8: Results obtained from the un-modified and GP modified aligned data.

5.3.3.2 Individual datasets alignment

For D$_Z$, 2305 peaksets remain after filtering, out of which 3 are significant. For D$_{VL}$, 2392 peaksets remain after filtering, out of which 775 are significant. For D$_M$, 2152 peaksets remain after filtering, out of which 5 are significantly different. When intersecting the significantly different peaksets, there is no peakset found to be in common between all three datasets. Therefore, aligning the peaksets together increases the number of samples, increasing statistical significance.

5.3.3.3 MS2 data

Due to the fragmentation strategy employed in each experiment, MS2 spectra were available only for a small percentage of the data. From the filtered peaksets there were 217 peaksets with MS2 spectra aligned, out of which 141 had an MS2 spectrum from one dataset attached to it, 65 peaks had MS2 spectra from 2 datasets attached and 11 peaks had MS2 spectra from all 3 datasets. The majority of peaksets were fragmented only in one dataset. Out of the 141 peaksets with an MS2 spectrum from one dataset, 18.4% peaksets had MS2 data only from D_M, 45.4% from D$_Z$ and 36.2% from D_{VL}. Based on Figure 5.9, the bad spectral similarity scores attached to peaksets with the same m/z mainly have a lower similarity score than the good scores (87%), which in general shows that the alignment worked.

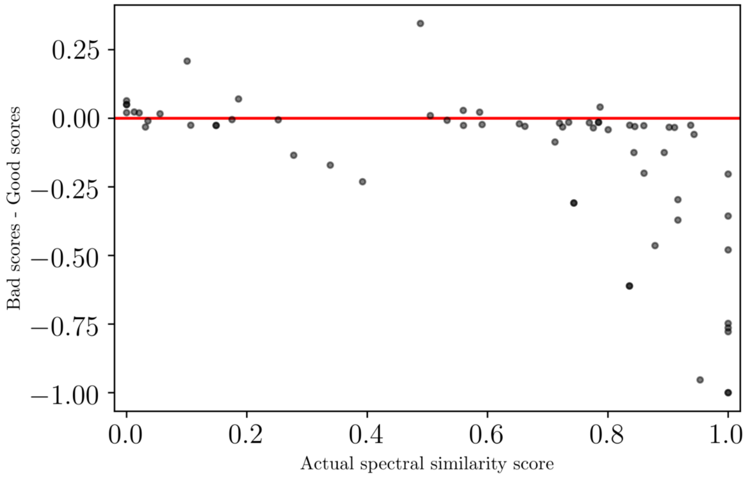


Figure 5.9: Spectral similarity scores of each peakset plotted against the difference between a random spectral similarity score and the actual spectral similarity score. The points below the red line represent the peaksets for which the actual spectral similarity score is higher than the score from randomly matched spectra with similar m/z.

5.3.4 Sample analysis

5.3.4.1 Modifying samples RT based on the obtained GP regression models

Next, the retention times in the samples files were modified based on the previously obtained GP regression models and aligned using the previously obtained *RTWindow* value.

For visualisation purposes, a particular peakset with m/z = 209.092 (L-Kynurenine) was extracted from each dataset and the chromatogram of the alignment before and after the GPR correction is shown in Figure 5.10. As it may be observed, after the GP correction, all peaks from the 3 datasets align properly.

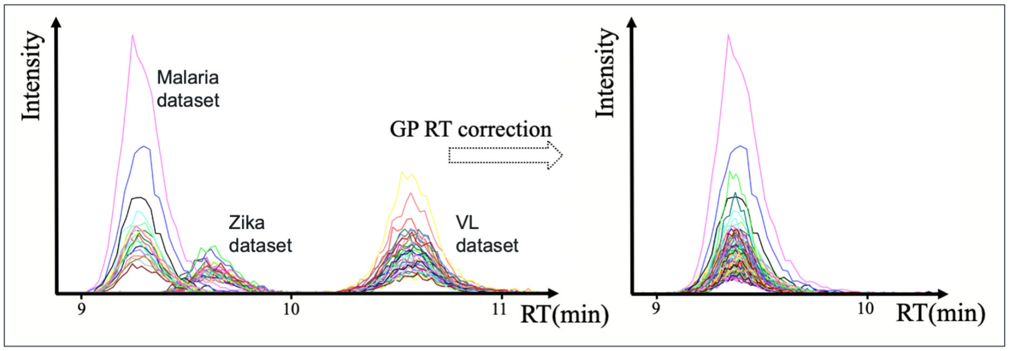


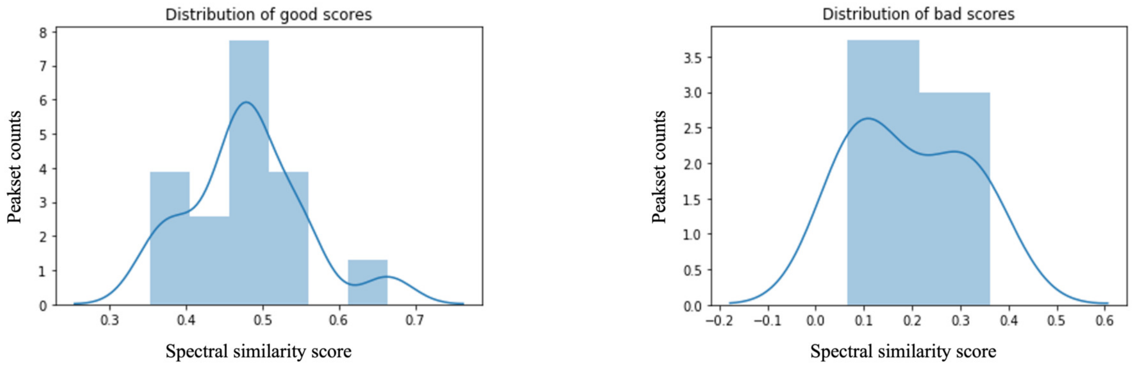
Figure 5.10: TIC for m/z = 209.092 (L-Kynurenine) before and after GPR correction

5.3.4.2 Samples alignment

For the samples alignment, the -ve ESI mode data was processed in the same way as the +ve mode data which was presented above. The JoinAligner module was run with RTWindow = 0.5 min in order to align all 74 samples across the three datasets. Following alignment there were 37220 peaksets in +ve mode and 24729 in -ve mode. After filtering out the peaksets with more than 50% values missing in any one dataset, only 1.68% of the total number of peaksets, i.e. 625, remained. A similar percentage was obtained in the case of the negative mode data where 1.85% (459) of the total number of peaksets remained. The differential expression analysis resulted in 207 significantly different (BH adjusted p-value < 0.05) features and 159 features in the positive and negative mode, respectively. The results are attached in the following Google document: Integrated LC-MS analysis results.

5.3.4.3 Metabolite annotation using HMDB and fragmentation spectra

Based on Figure 5.11 it can be clearly noticed that cosine similarity score of 0.35 marks a good threshold point to indicate a good annotation. To be noted, however, that scores of around 0.2 could signify a related molecule of the analysed molecule. Peaks with a MS2 spectral match of 0.35 or more were annotated with their respective HMDB annotation.



(a) **(b)**

Figure 5.11: Histograms of the counts of actual (good) spectral similarity scores (a) and theoretical bad spectral similarity scores (b). Most of the bad scores are distributed at a spectral similarity score = 0.1, with less of them being distributed at a similarity score = 0.35. In contrast to this, most of the good scores peak at a similarity score = 0.5.

Verifying the identity of tryptophan and kynurenine peaks using MS2 spectral information

Several statistically significant features which were annotated using the fragmentation spectra and HMDB experimental spectra are presented to illustrate the similarity after they have been matched. Their fragmentation spectra was graphically matched against the experimental Mass-

Bank fragmentation spectra using the metabolomics spectrum resolver (https://metabolomics-usi.ucsd.edu/) [139]. Particular focus was offered to a set of annotated features which are relevant to the fever mechanisms discussed in more detail in Chapter 6. Among these tryptophan, kynurenine and niacinamide were of particular significance.

As both the tryptophan and kynurenine peaks had spectral information from one out of the three datasets, it was used to verify the identity using the cosine similarity score with experimental LC-MS/MS spectral information from HMDB. Metabolomics spectrum resolver was also used to further check the similarity between the dataset spectrum and spectra obtained from MassBank. Similarity scores of 0.35 and 0.54, respectively, were obtained for kynurenine and 0.31 and 0.55, respectively, for the tryptophan fragment (with loss of ammonia) (Figure 5.12, Figure 5.13).

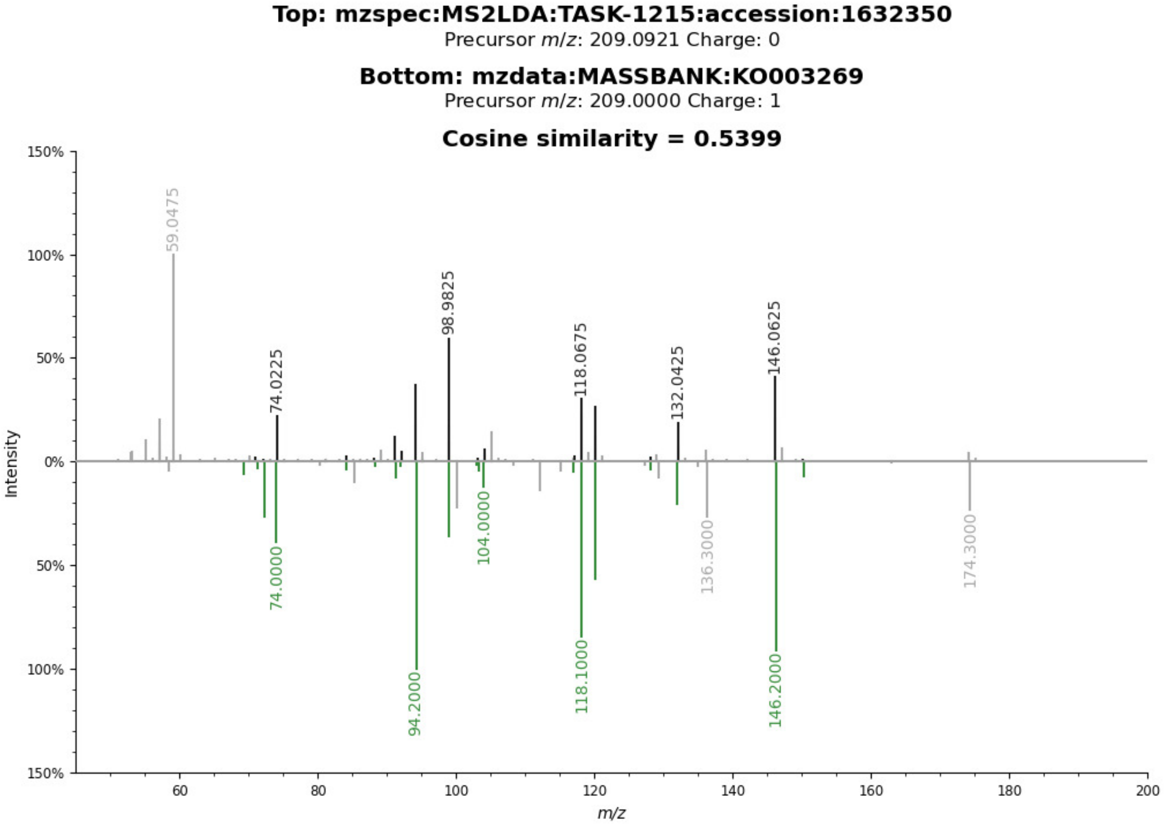


Figure 5.12: Peakset annotated as Kynurenine (M+H[1+]) using SRM matching method, *mummichog* and HMDB matching method. The spectrum belongs to D_{VL} (resolver obtained from ms2lda.org) and it was matched against experimental LC-MS MS2 information from MassBank compound KO003269 with a cosine similarity score of 0.54 (fragment tolerance =0.2) (Metabolomics spectrum resolver plot).

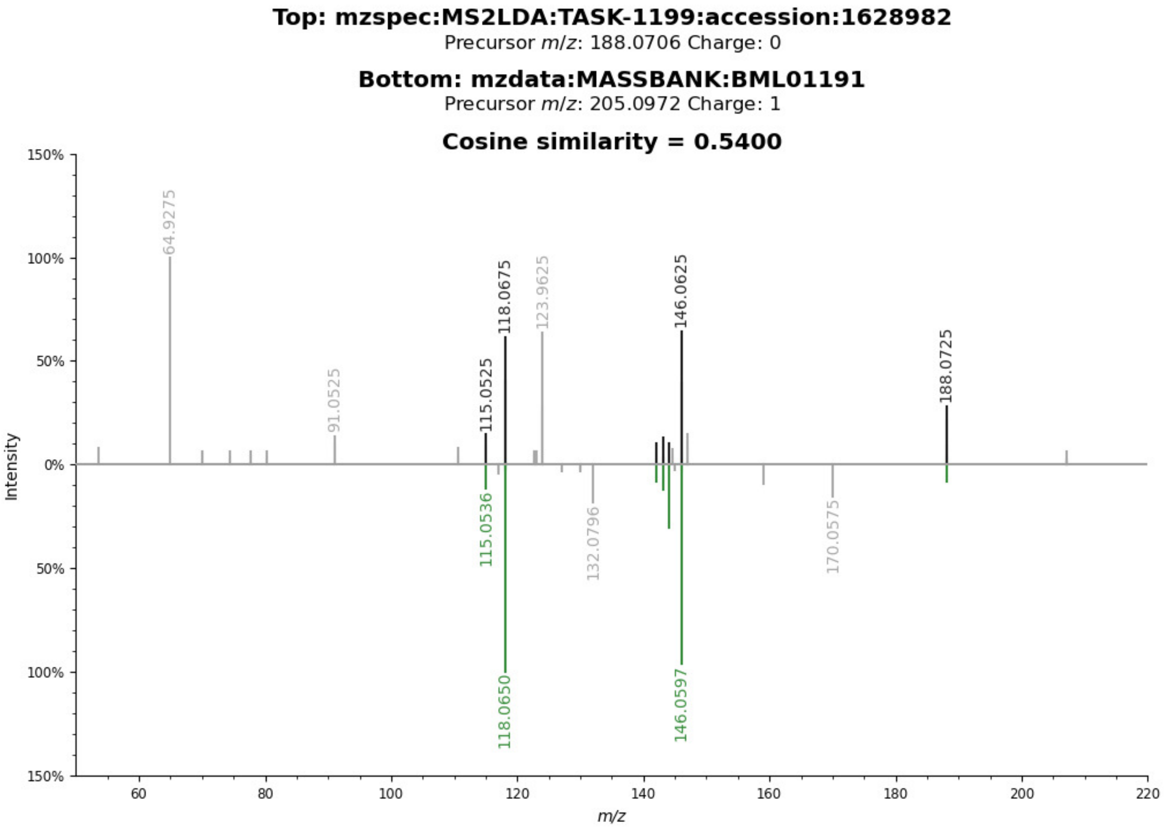


Figure 5.13: Peak annotated as L-Tryptophan fragment with loss of ammonia (M-NH3+H[1+]) using SRM matching method, *mummichog* and HMDB matching method. The loss of ammonia from protonated tryptophan was observed as the primary fragmentation pathway in gas-phase reactions (Lioe et al., 2004). The spectrum belongs to D_Z and it was matched against experimental LC-MS MS2 information from MassBank BML01191 compound with a cosine similarity score of 0.55 (fragment tolerance =0.2) (Metabolomics spectrum resolver plot).

Verifying the alignment in the case of niacinamide Both niacinamide and niacin levels are significantly lower in infected patients. The peak annotated as niacinamide in + ESI mode has MS2 spectra from two datasets and in this case the similarity between the two spectra is illustrated in Figure 5.14. A cosine similarity score of 1 was obtained, demonstrating that in this case the alignment between the datasets was optimal.

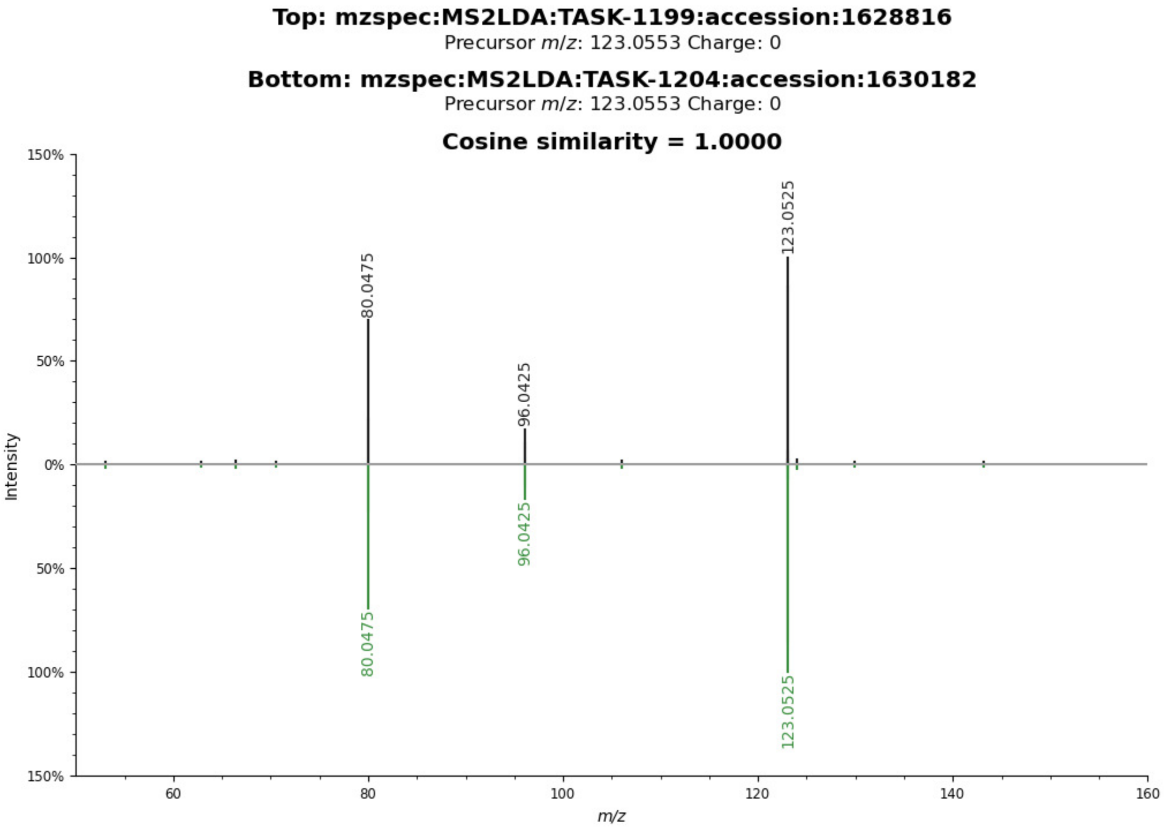


Figure 5.14: Peak annotated as Niacinamide (M+H[1+]) using SRM matching method, *mummichog* and HMDB matching method. Top spectrum belongs to the MS2 information obtained from D_M and the bottom spectrum belongs to the MS2 information obtained from D_Z. The cosine similarity obtained using metabolomics spectrum resolver is 1, which signifies a perfect match and also that the alignment was accurate in this case (Metabolomics spectrum resolver plot).

5.4 Discussion

The algorithm behind the integration of the three datasets consisted of correcting the peaksets retention time drift between the peak detected datasets based on GP models fitted to the RT drift in SRM metabolites and, subsequently, aligning them. Up until recently, retention time drift occurring in LC-MS experiments has only been studied in the context of the same experiment, rather than between different experiments. Recent studies have begun exploring the retention time drift problem in the context of large sample sizes and proposed several algorithms for correcting it. One study in which alignment between samples from large-scale datasets is addressed, proposed a profile-based alignment algorithm which uses a graphical time warping method to correct the retention times for the mis-aligned features previously detected [140]. In this case, retention time drift was corrected across different m/z bins. Two large-scale LC-MS datasets –Rotterdam dataset (N=1000 samples) and MESA dataset (N=1977 samples)- were used for developing and testing the algorithm. Both of them consisted also of internal quality control samples which are aliquots of pools of all study samples. Misalignment was observed between the samples in each dataset, which could also have been caused by the difference in acquisition instrumentation. In this case, mis-aligned features referred to those which had a non-random set of sample indices. Their reasoning was that neighbouring samples tend to be aligned together due to the similar RT shifts. Thus, a feature that contains a continuous run of samples is deemed to be incorrect, due to the fact that sample orders should be random in a well designed experiment [140]. They then apply warping to the mis-aligned features as a function of m/z. In this chapter, it was also tried to stratify the peaks according to m/z and apply different warping functions. However, no difference between the different bins was noticed. This could also be due to the smaller sample size available. In [140], they use a modified version of graphical time warping (GTW) to the XIC profiles of the mis-aligned profiles.

A few studies used endogenous reference peaks as landmarks to model the RT shift between sets of samples. Watrous et al [127] presented a feature-based method for aligning the samples at a population scale (N=2895 human plasma samples) by correcting the non-linear retention time shift inside the raw files, in a manner that is similar to this study. In order to determine the retention shift between samples, they have used internal standards which were isotope labelled and allowed modelling of the shift to correct raw files for further peak detection and alignment. Afterwards, alignment was employed using MZmine's JoinAligner alignment algorithm. Another study which used a feature-based alignment and RT drift correction method was by Li et al [141]. They used adjacent tandem mass spectrometry information to select suitable endogenous reference compounds which would act as landmarks for modelling the RT drift. However, this was used on a small scale metabolomics dataset.

A recently published study [142] performed integration of two metabolomics datasets. Similar to [141] they have also used internal reference compounds which were selected based on their m/z and intensity. In contrast to our approach, in [142] each dataset was aligned separately and annotated. They used reference metabolites generally found in metabolomics datasets (eg. creatinine in urine), which were identified within both datasets and were mapped against each other based on their RTs. A generalised additive model (GAM) was then fitted on the data and, based on it, a predicted RT was calculated for all the peaks in the non-reference dataset. For each feature pair in each m/z bin a score was calculated to rank the alignments on m/z, rt and relative abundance (Q). This is quite similar to the score used by MZmine JoinAligner algorithm, apart from the fact that they take into account the abundance parameter.

The advantage of the approach presented in this chapter of aligning multiple LC-MS datasets at a retention time level is the identification of both putatively annotated and unannotated compounds which act in the same way or uniquely to each disease. GP regression models provide a robust framework to correct the RT drift between datasets. Additionally, as the overall sample size is increased, statistical robustness of the analysis is enhanced, provided assumptions on the underlying similarity in responses of disparate datasets, e.g. in this case, the separate pathogen related infections here, are robust. A common limitation in many LC-MS biomarker discovery studies introduced by the small sample size is, thus, overcome. Annotation using MS2 information is also improved, as some datasets contain MS2 spectra for peaksets which are absent in other datasets. This could be advantageous for datasets which have limited fragmentation data available, as it improves the chance of a peakset having MS2 spectra aligned, and thus the possibility for better annotation. In this sense, the datasets become complementary to each other.

One of the limitations of this study was that the algorithm was only tested with datasets run in the same laboratory, on the same LC-MS platform, and at the moment it is not known whether this method could be applied to metabolomics datasets run on different platforms. Annotation was also a limitation, as it is the case for metabolomics studies in general [143, 144]. It is to be noted that for some of the metabolites, the difference between the control group and the infected group was larger in one of the datasets than in the other datasets and in some cases this could have contributed to the statistical significance of the difference in the metabolites abundance in control as opposed to infected. This could be either due to the disease itself or its severity. In this case it is to be noted that D_M was an intervention study and the infection was controlled and less severe which might explain relatively lower logFC in the dataset.

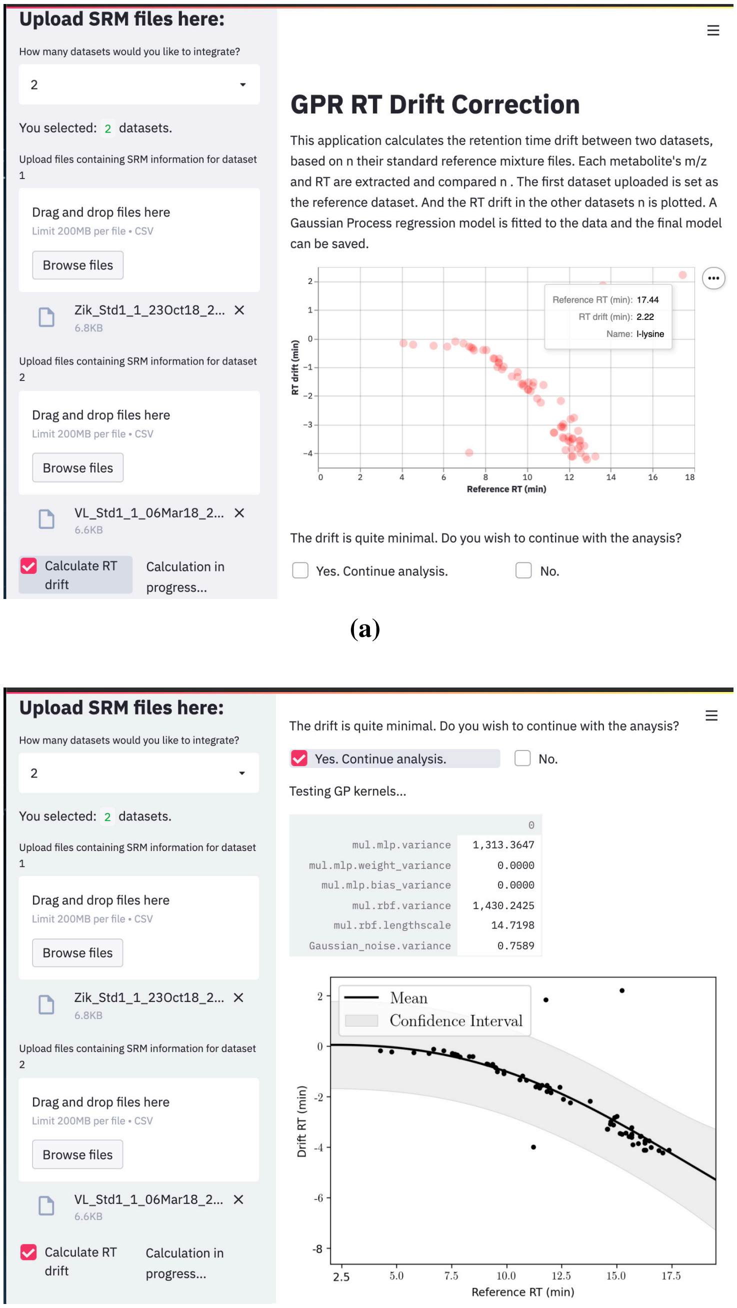
5.4.1 Future directions

Regarding future research directions, improvements to the evaluation of the algorithm could be made. At the moment, the evaluation was performed by comparing the GPR corrected aligned datasets with the aligned datasets which were not drift corrected and more aligned peaksets were found in the GPR corrected aligned datasets. Additionally, fragmentation data from individual datasets was compared for each peakset with more than one MS2 spectrum aligned to it. This was done using cosine score to determine the similarity between the aligned fragmentation spectra. This provided a more quantitative approach in evaluating the algorithm. An even better evaluation method would be by using a control dataset which is run twice on the LC-MS platform using different elution times, e.g. 20 min and 30 min, similarly to [142]. The detected peaksets from the resulting two datasets would be aligned and compared with the detected peaksets found in one of the dataset. The two resulting peaksets should then interpose with each other. Once this form of evaluation is performed, the algorithm could be tested on other metabolomics datasets. Another suggestion both for improving the evaluation algorithm and for enabling molecular networking between the datasets is related to specific MS2 data acquisition. This experiment used TopN MS2 data acquisition, where the first N ions with the highest intensities were selected for fragmentation, but MS2 data acquisition specifically for the ions of interest which were detected following alignment would improve the evaluation.

The integrating method presented in this chapter could be further developed into an web application which could be used with ease by any researcher. As a proof-of-concept, a simple web application was developed using Streamlit to compute the RT drift between 2 datasets and model it using GP regression. This uses the ToxID files obtained following the SRM run to extract the SRM metabolites information (m/z and RT). The files belong to datasets which the user would like to integrate. The graphical user interface (GUI) is illustrated in Figure 5.15. The first introduced files are considered the reference datasets. The RT drift is then computed and plotted inside the GUI. If the mean of the drift is less or equal to 30 s then the user is asked whether he wishes to continue with the analysis, i.e. fitting a GP model to it (Figure 5.15).

5.4.2 Conclusion

In conclusion three LC-MS datasets investigating metabolic changes in Zika, malaria and VL infected patients were successfully aligned together using fitted GP models for correcting the RT drift between them, determined by the RTs of the SRM metabolites within each dataset. Proper peakset alignment across multiple disparate untargeted metabolomics datasets led to the identification of compounds changing in abundance in similar ways across the different infectious diseases. Moreover this sort of approach was easily integrated with already existing alignment software such as JoinAligner. Following compound annotation and statistical analysis, both common and specific dysregulation patterns were observed in metabolic pathways. These are presented and discussed in detail in the next chapter.



Upload SRM files here:

How many datasets would you like to integrate?

2

You selected: 2 datasets.

Upload files containing SRM information for dataset 1

Drag and drop files here
Limit 200MB per file • CSV

Browse files

Zik_Std1_1_23Oct18_2... ✕
6.8KB

Upload files containing SRM information for dataset 2

Drag and drop files here
Limit 200MB per file • CSV

Browse files

VL_Std1_1_06Mar18_2... ✕
6.6KB

☑ Calculate RT drift

Calculation in progress...

GPR RT Drift Correction

This application calculates the retention time drift between two datasets, based on n their standard reference mixture files. Each metabolite's m/z and RT are extracted and compared n . The first dataset uploaded is set as the reference dataset. And the RT drift in the other datasets n is plotted. A Gaussian Process regression model is fitted to the data and the final model can be saved.

The drift is quite minimal. Do you wish to continue with the anaysis?

□ Yes. Continue analysis. □ No.

(a)

Upload SRM files here:

How many datasets would you like to integrate?

2

You selected: 2 datasets.

Upload files containing SRM information for dataset 1

Drag and drop files here
Limit 200MB per file • CSV

Browse files

Zik_Std1_1_23Oct18_2... ✕
6.8KB

Upload files containing SRM information for dataset 2

Drag and drop files here
Limit 200MB per file • CSV

Browse files

VL_Std1_1_06Mar18_2... ✕
6.6KB

☑ Calculate RT drift

Calculation in progress...

The drift is quite minimal. Do you wish to continue with the anaysis?

☑ Yes. Continue analysis. □ No.

Testing GP kernels...

	0
mul.mlp.variance	1,313.3647
mul.mlp.weight_variance	0.0000
mul.mlp.bias_variance	0.0000
mul.rbf.variance	1,430.2425
mul.rbf.lengthscale	14.7198
Gaussian_noise.variance	0.7589

(b)

Figure 5.15: Graphical user interface for the web application developed with Streamlit for detection and modelling the RT drift between sets of SRM metabolites from disparate metabolomics datasets: a) SRM files are uploaded, and the RT information of each SRM is extracted for two of the datasets and plotted against each other; b) the GP regression model is applied to the identified RT drift.

Chapter 6

Identifying metabolites common to fever-associated diseases and specific to individual diseases

This chapter aims to further explore the molecular mechanism of febrile illnesses and discuss the putatively annotated features identified in the analysis of the previous chapter. These were investigated based on the logFC values obtained following the statistical analysis performed with limma. These denoted the log transformed fold change between the infected and control samples. For each feature, the logFC value was computed both for the integrated datasets, logFC$_m$, and also for each individual dataset, logFC$_i$. Based on the logFC$_i$ obtained for each individual dataset, features were categorised as being downregulated or upregulated in a similar manner across the datasets, or uniquely for each dataset. These features are investigated in the next sections of this chapter. The pathway analysis results obtained using *mummichog* and KEGG are also discussed.

Commonly regulated features could be indicative of disease severity or infectious disease in general, whereas individually regulated features could aid in the diagnosis of a specific febrile infectious disease. It should be taken into account, however, that the interpretation of the results is highly dependent on the annotation of the features and the accuracy of the peak detection process. In this case annotation was performed either by mapping the features on to the SRM list (srm), by using the MS2 data and matching the spectra against experimental spectra from HMDB (ms2) or as a result of *mummichog* analysis (mm). Overall, for the +ve ESI mode data, out of the total number of peaksets, 15% were annotated with srm, 8% with ms2 and 39.8% with mm, and for the -ve ESI mode data 5.7% were annotated using srm, 4.4% with ms2 and 35.5% with mm. Whenever a feature had annotation from multiple sources which did not match, the srm annotation was the one taken into further consideration. The annotation process also took into consideration possible adducts and fragments which could result during the LC-MS analysis.

Therefore, multiple peaksets could be annotated as being the same metabolite. Additionally, following the peak detection process, there is a possibility of split peaks or false peaks being detected, as a consequence of static parameterization [66]. This also, in turn, influences the annotation process and could result into erroneously annotated peaksets. Identical annotation associated with multiple peaks could also be due to the presence of isomers, i.e. compounds with the same m/z, but different chemical structure.

6.1 Metabolites in common between the fever-associated diseases

In this section, the set of metabolic compounds which present common perturbations across the studied datasets are discussed. The putatively annotated metabolites which were found to be either upregulated or downregulated in all three datasets are presented, and their importance in the fever molecular mechanism is discussed based on the pathway analysis findings.

6.1.1 Overview of significantly modified annotated metabolites

The significant features which have been putatively annotated and present a common trend based on the $logFC_i$ values between the infected and control groups are presented in Figure 6.1, Figure 6.2 and Figure 6.3. Based on this, a general trend over the three datasets was established. For the +ve ESI mode data, 30 peaksets presented a general upward trend (all datasets have $logFC_i>0$) out of which 11 were statistically significant and 150 peaksets presented a general downward trend ($logFC_i<0$) out of which 69 were significant. For the -ve ESI mode data, 46 peaksets presented a general upward trend out of which 25 were statistically significant and 115 peaksets presented a general downward trend out of which 74 were significant. Figure 6.1 presents the upregulated metabolites in the infected group. These include metabolites from different molecule classes such as amino acids (asparagine, aspartic acid, kynurenine and kynurenic acid), sugars (glucose), nucleic acids (cytosine) and lipids (chenodeoxyglycocholate). Figures 6.2 and 6.3 contain boxplots of the downregulated metabolites in the infected group. In this case, more metabolites were identified, which are discussed in more detail in the next section.

It is to be noted that the boxplots were plotted after imputing the missing values. For some of the peaksets, some samples had registered an intensity value of 0. A reason for this might be that no peak was identified for that particular sample. For illustrative purposes, mainly to aid with scaling issues when plotting the values into boxplots, the missing values were imputed using KNN algorithm as described in Chapter 5 Section 2.6. Sometimes this might have an impact on the boxplots, and might skew the actual $logFC_i$ values, i.e. if the actual $logFC_i$ value is positive, after imputing missing values the resulting $logFC_i$ value could be negative.

A table containing the results following the statistical analysis with annotation is included in the linked table. It was observed that a large proportion of the features elutes in clusters of similar retention time, indicating that they might be related, i.e.fragment or adduct of the parent ion. The list of ions used to calculate the possible adducts/fragments of each peak are included in the linked table. Common adducts include gain of ^{13}C, ^{34}S, H_2O, Na or K and common fragmentation patterns include loss of CO, H_2O, HCOOH, NH_3, $C_3H_4O_2$, H_4O_2.

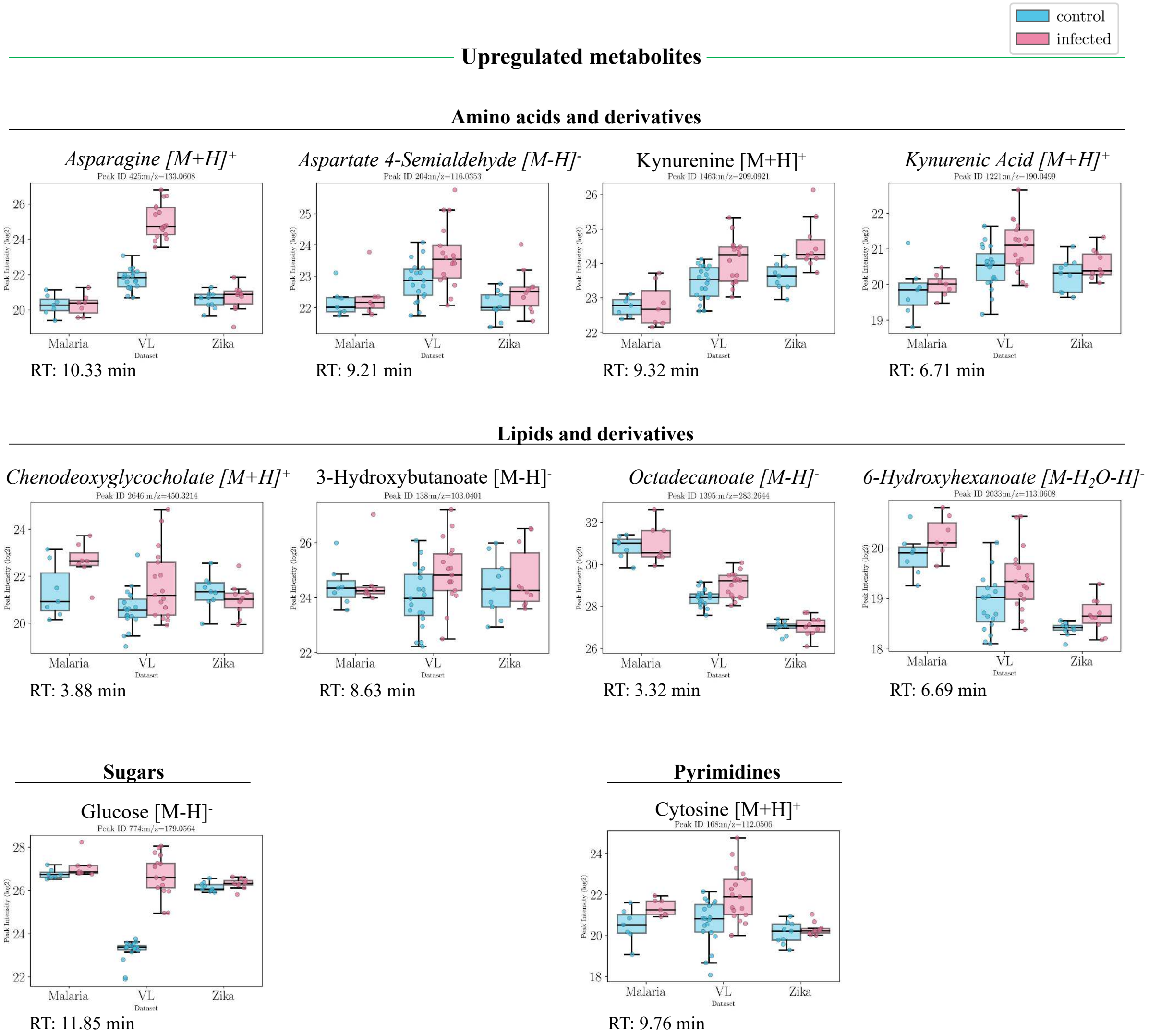


Figure 6.1: Boxplots of annotated compounds for both conditions in each dataset. Overview of annotated metabolites which are statistically significant (p-val<0.05) and present a general upward trend in all three datasets, i.e higher intensities in infected patients. Values from both positive and negative ionisation mode are presented from left to right in ascending order of their p-value. Metabolites in italic font are only annotated using *mummichog*.

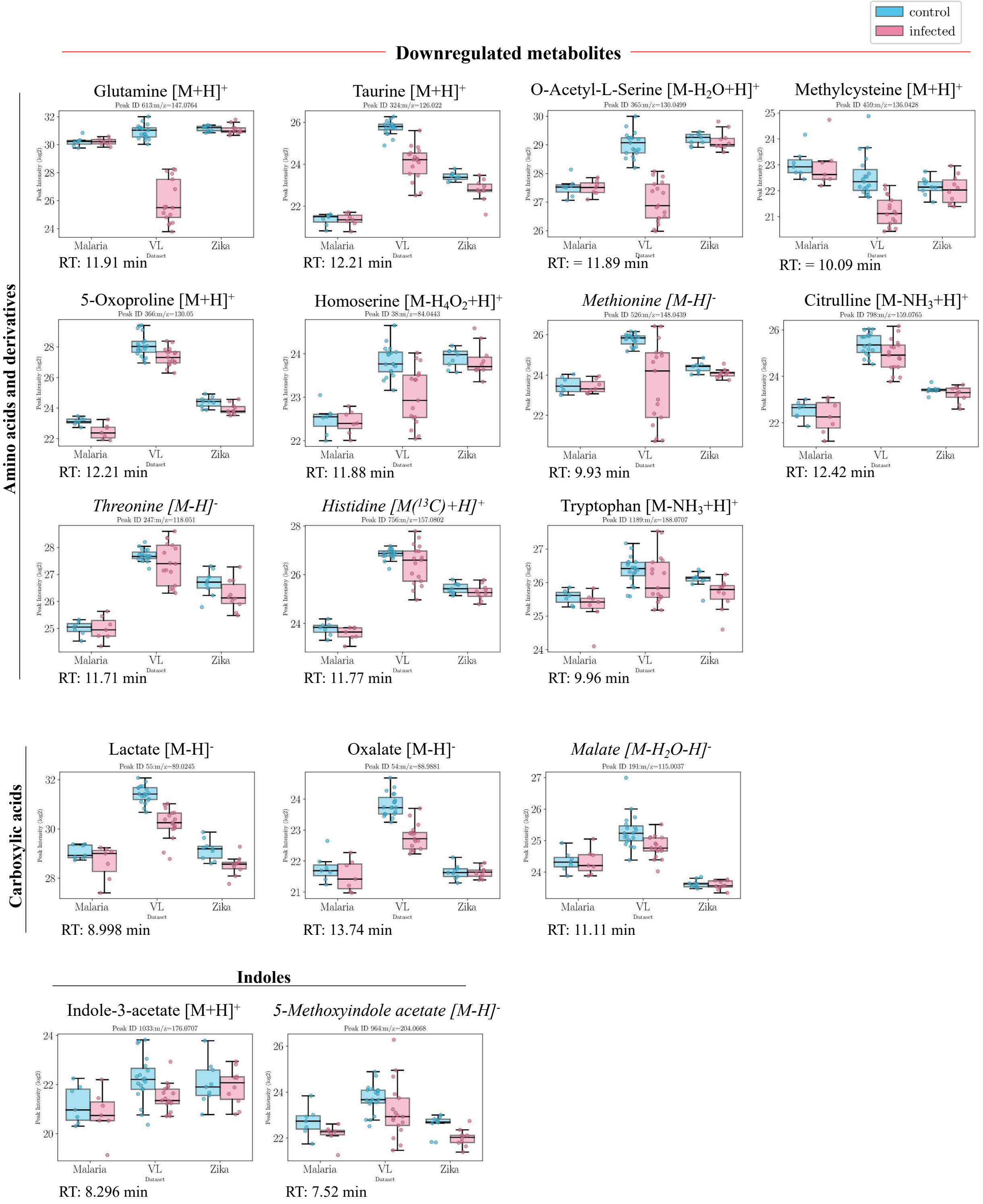


Figure 6.2: Boxplots of annotated compounds for both conditions in each dataset. Overview of annotated metabolites which are statistically significant (p-val<0.05) and present a general downward trend in all three datasets, i.e. lower intensities in infected patients (with p-value <0.05). Values from both positive and negative ionisation mode are presented from left to right in descending order of their p-value in each group. Metabolites in italic font are only annotated using *mummichog*.

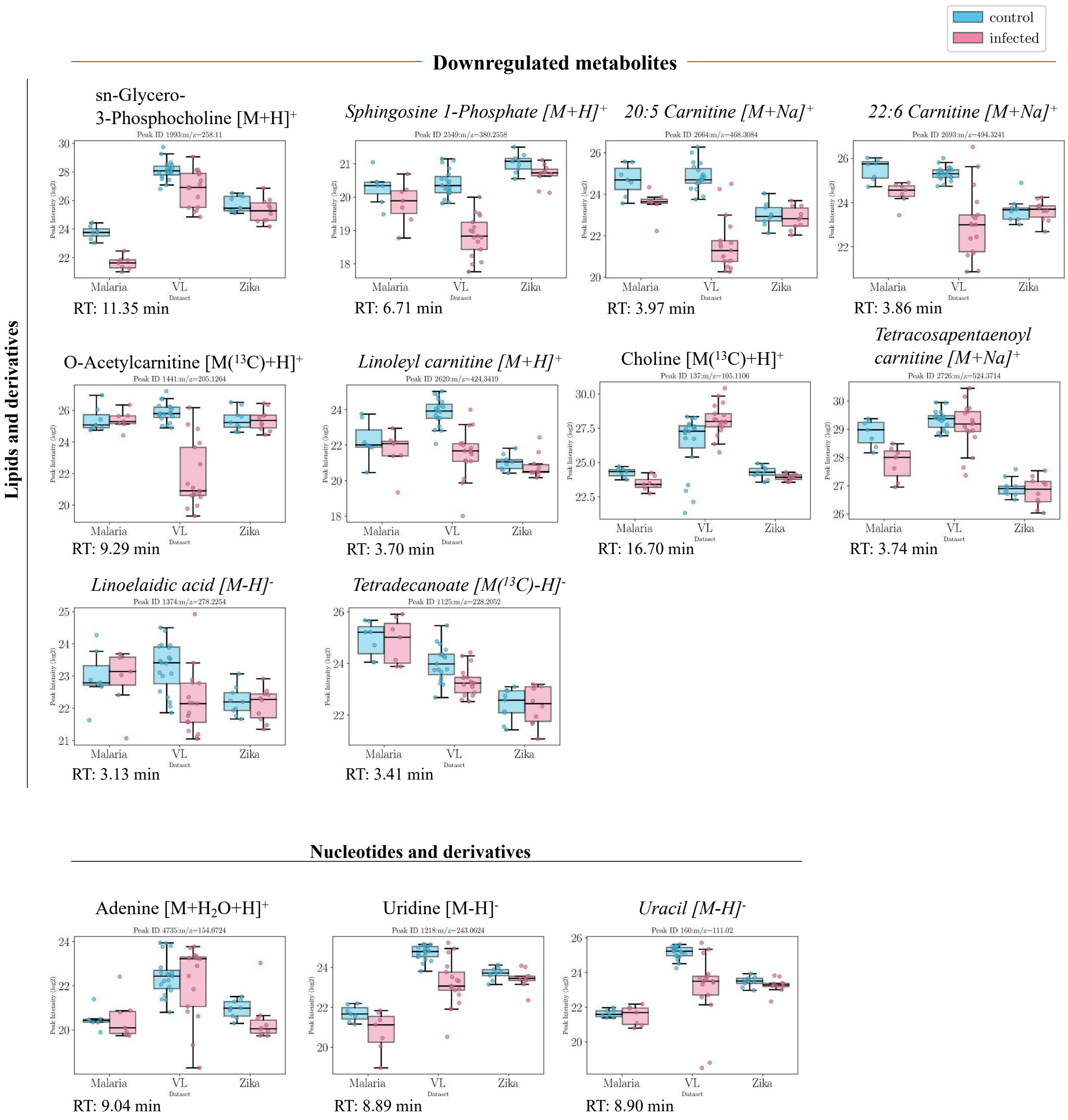


Figure 6.3: Boxplots of annotated compounds for both conditions in each dataset. Overview of annotated metabolites which are statistically significant (p-val<0.05) and present a general downward trend in all three datasets, i.e. lower intensities in infected patients (with p-value <0.05). Values from both positive and negative ionisation mode are presented from left to right in descending order of their p-value in each group. Metabolites in italic font are only annotated using *mummichog*.

6.1.2 Pathway analysis results

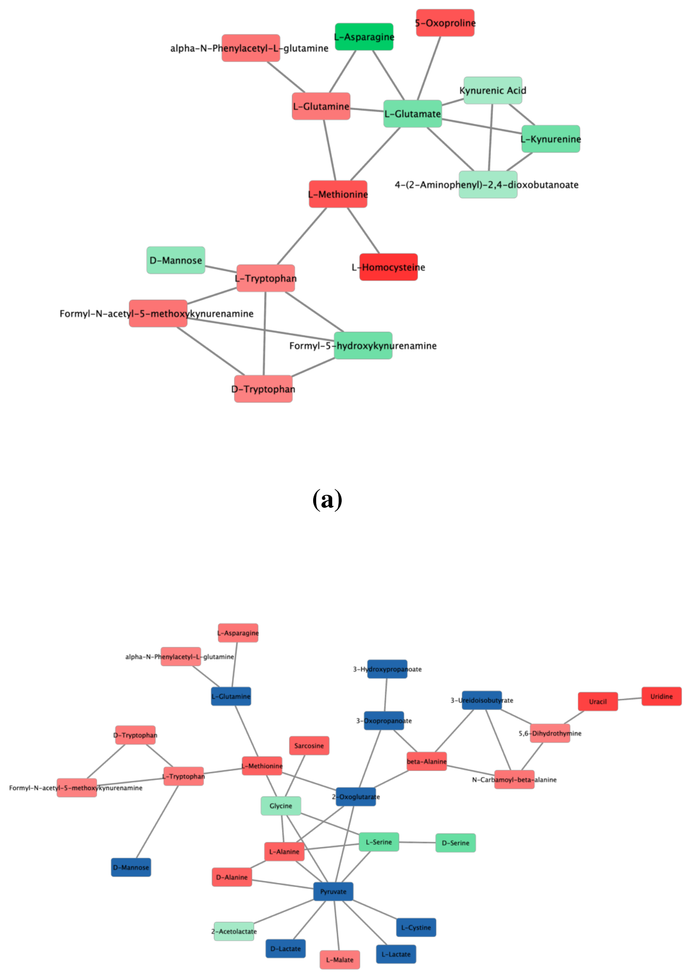
Following *mummichog* pathway analysis for the data obtained in both the positive and negative ionisation modes, 20 KEGG pathways were found to be statistically significant. The algorithm behind *mummichog* is constituted of two main complementary parts: pathway analysis and module analysis [136]. The module or activity network can be within a pathway or between several pathways which show more internal connections than expected randomly in the whole network [136]. This tends to be less biased than the usual biological pathway. The pathway and network analysis in *mummichog* are performed by mapping the data to KEGG pathways and BioCyc metabolic networks, respectively. Statistical significance of the networks is calculated based on a variation of Fisher's exact test (FET). In order to distinguish between real signals and random data, a distribution of random data is generated and mapped to the database pathways [136]. By contrasting the pathway enrichment result obtained from statistically significant features against random data, the likelihood of finding the correct pathways and metabolites can be quantified [136].

The pathway analysis revealed a significant impact of the studied infectious diseases primarily on nitrogen metabolism with a focus on tryptophan metabolism (Table 6.1). Following the modular analysis, the activity network for the +ve ESI mode data is also centered around tryptophan metabolism, specifically the kynurenine pathway, which is discussed in more detail next (Figure6.4). For the -ve ESI mode data, the activity network includes tryptophan metabolites and, additionally, citric acid cycle metabolites.

KEGG Pathway	p-value	ESI mode
Nitrogen metabolism	0.000672	+
Tryptophan metabolism	0.00084	+
Alanine and Aspartate Metabolism	0.002017	+
Pyruvate Metabolism	0.00521	-
Vitamin B3 (nicotinate and nicotinamide) metabolism	0.005546	+
Pyrimidine metabolism	0.007226	+
Glycolysis and Gluconeogenesis	0.008151	-
Carnitine shuttle	0.009243	+
Glycerophospholipid metabolism	0.013024	-
Glycosphingolipid metabolism	0.013024	-
Methionine and cysteine metabolism	0.015965	+
Aminosugars metabolism	0.022939	+
Purine metabolism	0.022939	+
Bile acid biosynthesis	0.031846	+

Fatty Acid Metabolism	0.031846	-
Putative anti-Inflammatory metabolites formation from EPA	0.031846	+
Androgen and estrogen biosynthesis and metabolism	0.03924	-
C21-steroid hormone biosynthesis and metabolism	0.03924	-
Vitamin B1 (thiamin) metabolism	0.03924	-
Glutathione Metabolism	0.04445	+

Table 6.1: Significantly altered metabolic pathways (p-val<0.05) following *mummichog* analysis of the negative and positive ionisation mode data.



(a)

(b)

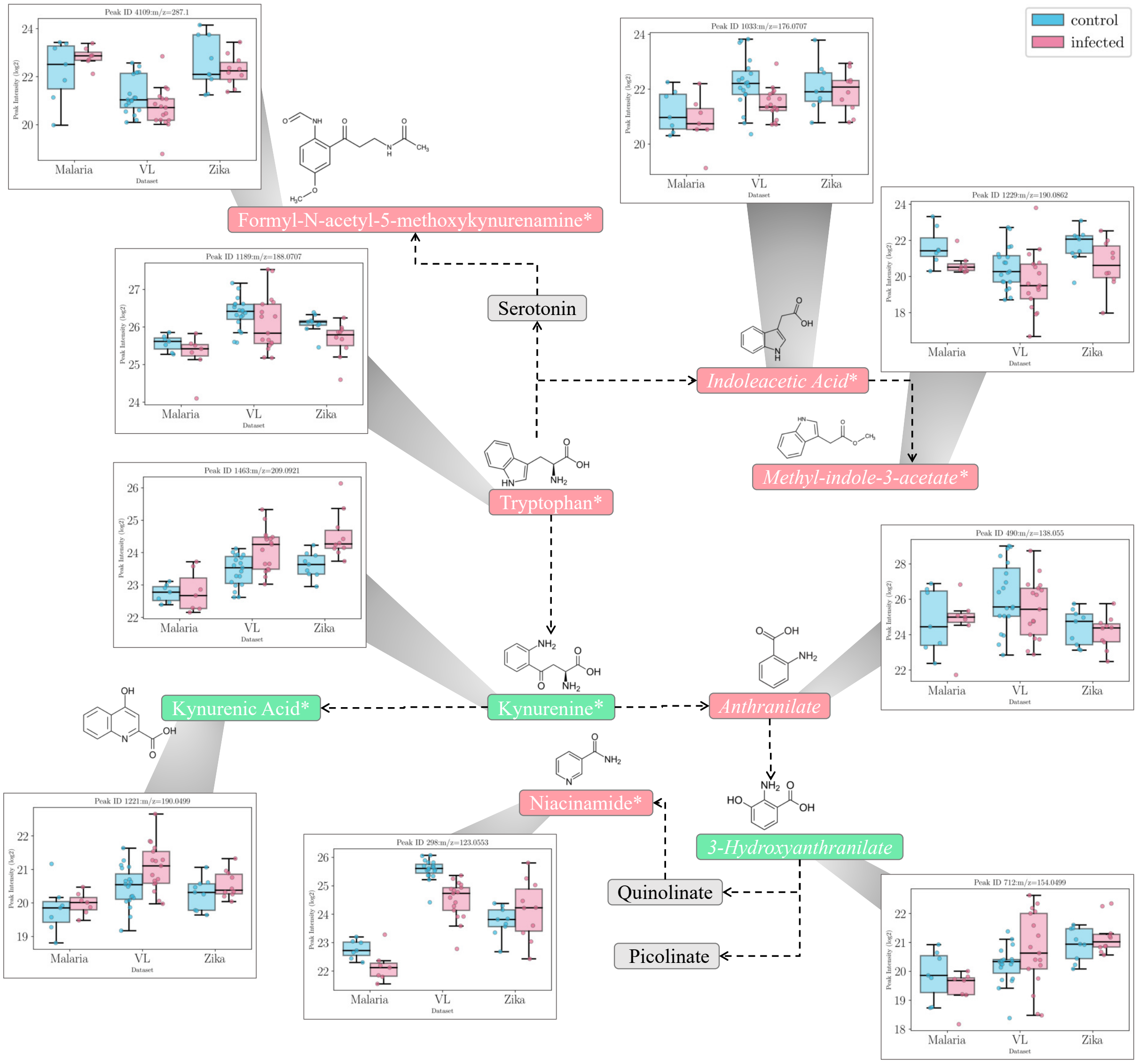
Figure 6.4: *mummichog* activity networks plotted in Cytoscape obtained following the analysis of the positive mode LC-MS data (up) and negative mode LC-MS data (down). Metabolites involved in the tryptophan metabolism and citric acid cycle are predominat in the activity networks.

6.1.2.1 Tryptophan metabolism

Focused analysis on this pathway represented in Figure 6.5 revealed significant decreases accompanied by a general downward trend in all three datasets in tryptophan and tryptophan derivatives such as indoleacetic acid and methoxyindole acetate. Methyl indole acetate and formyl-N-acetyl-5-methoxy kyunernamine were also significantly decreased in the infected group with the exception of the Malaria dataset where the logFC$_i$ value was slightly higher (Figure 6.5). In contrast, the kynurenine pathway suggests an increased activation as kynurenine and kynurenic acid present higher levels in infected patients in all three datasets. 3-Hydroxyanthranilate was in general higher as well with the exception of the Malaria dataset were logFC$_i$ value was slightly lower. Anthranilate levels were also reduced in infected patients across all datasets, although not statistically significant so. Nicotinic acid (- ESI mode) was also found in significantly lower levels in infected patients and nicotinamide was significantly lower, although Zika dataset had positive logFC$_i$ (Figure 6.3). It is important to note that unless otherwise stated, annotations are putative and based on m/z alone.

Tryptophan metabolism has previously been associated with various agents of infection [145], particularly its flow through the kynurenine pathway which produces metabolites including kynurenate and nicotinamide adenine dinucleotide (NAD+). Of particular interest is the inverse relation between kynurenine and tryptophan, as the ratio between the two is used to measure the activity of the enzyme indoleamine-2,3-dioxygenase 1 (IDO-1) [146]. IDO-1 is the rate limiting step of the tryptophan pathway and it catalyses the breakdown of tryptophan to kynurenine. IDO-1 activity is also tightly regulated by interferon gamma (IFN-γ) activity [147]. Similar to this, COX-2, an enzyme central to the fever process, can also be induced by IFN-γ [148]. The interplay between IDO-1 and COX-2 enzymes has been previously studied where inhibition of COX-2 enzyme has led to a downregulation in IDO-1 and decrease in kynurenine metabolites [149].

The effect of each separate disease on the kynurenine pathway has also been studied previously for specific infectious disease associated with fever: malaria [150], HAT [151], Zika virus [152] and VL [153]. For each disease, an increased degradation of tryptophan into kynurenine, as a consequence of an increased IDO-1 activity was observed. IDO-1 is located in immune cells such as macrophages and monocytes, and nerve cells, microglia, astrocytes and neurons [152]. In the presence of proinflammatory cytokines, particularly IFN-γ, TNF-α and less so by IFN-α and IFN-β [154], IDO-1 gets activated, which in turn leads to neuroprotective and neurotoxic metabolites being generated. In mice models with increased levels IFN-γ, higher levels of quinolinic acid in the hippocampal area were observed [154]. It could also be hypothesised in this case, that the increased level in kynurenine and decreased levels of tryptophan not only indicate an increased IDO-1 activity, but also an increased COX-2 activity.



🟥 Lower peak intensities in infected patients
🟩 Higher peak intensities in infected patients
⬜ Metabolites not found in the dataset
* p-value < 0.05

Figure 6.5: Tryptophan metabolism and the changing metabolites from each dataset. Intensities values are represented as lg2 values. The metabolites were mapped against the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway map hsa00380. Metabolites in italic font are annotated following *mummichog* analysis or HMDB matching method, while the rest are annotated using the SRM metabolites information. The boxplots representing the intensities of all the samples in each condition (red = infected, blue = control) in all three datasets.

6.1.2.2 Amino acid metabolism

The rest of the significantly affected pathways relate mostly to other amino acids metabolism such as alanine, aspartate and glutamate metabolism, methionine and cysteine metabolism and glutathione metabolism (Table 6.1). Figure 6.3 indicates a clear suppression in amino acid metabolism, especially in glutamine and related metabolites. In immune cells, glutamine is converted through glutaminolysis into glutamate, aspartate and alanine by undergoing partial oxidation [155]. This could explain the decrease in glutamine and increase in aspartate in the infected group in all three datasets. Similarly, another glutamine derived metabolite, asparagine, was found to be increased in infected patients in all three datasets. Aspartate then feeds into the urea cycle and gets converted into arginosuccinate, fumarate, arginine, urea, ornithine and citrulline which gets converted back to aspartate [156].

Glutamine also acts as a precursor for citrulline which plays an important role in arginine biosynthesis in the urea cycle. Citrulline levels were significantly lower in the infected group from the three datasets. Citrulline is also a product of arginine deamination along with nitric oxide (NO), a process which is catalysed by nitric oxide synthase. It is well known that NO takes part in the antimicrobial defence mechanism during infection and inflammation [157]. The antimicrobial effect does not stem from NO, but from the reactive nitrogen intermediates formed after its oxidation, which inactivate microbial enzymes (ribonucleotide reductase, aconitase) [158]. The lower levels of citrulline in the infected groups, which could also signify higher levels of NO, could indicate an increased production of NO to fight against the infection. For the present meta-dataset one peakset was identified as being arginine through ms2 matching (ID:760). This was indeed downregulated in all datasets, but without statistical significance.

Altered glutathione metabolism accompanied by a dysregulated methionine and cysteine metabolism could also be observed from the analysis of the metadataset. The plasma levels of the amino acid 5-oxoproline (pyroglutamic acid) were also lower in the infected patients in comparison to healthy controls. Additionally, the sulfur containing amino acids methionine and methylcysteine showed a decrease in the infected patients. A precursor of cysteine, o-acetylserine was also decreased in the infected group alongside threonine and homoserine. Taurine, another metabolite derived from cysteine metabolism, was found to have significantly lower levels in infected patients in the metadataset. Overall, the decrease in reduced thiols, may arise due to increased levels of oxidative stress in response to infection. There has been previous evidence which suggests that the alteration of the redox homeostasis and glutathione depletion affect normal body temperature [159].

Other amino acids presenting lower levels in all three datasets in the infected group but not with statistical significance include: beta-alanine, proline and betaine. In contrast, a few amino acids and their relatives presented overall higher levels in infected patients: carnitine, tyrosine and leucine.

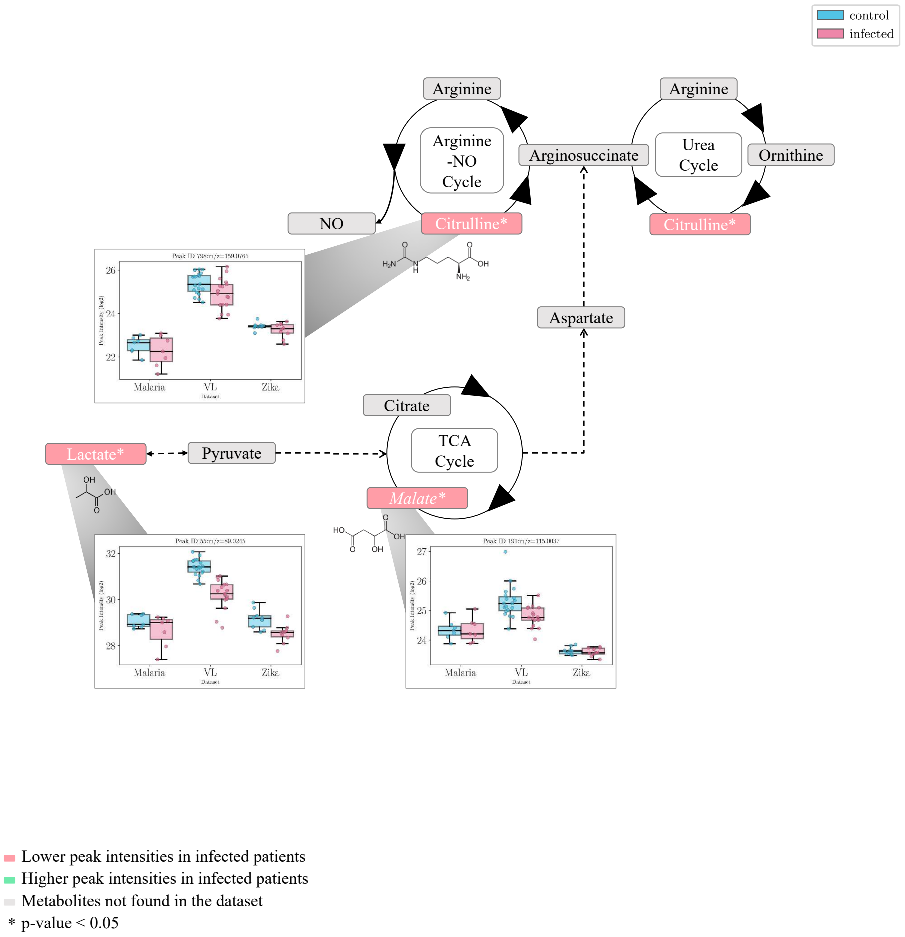


Figure 6.6: NO synthesis and the changing metabolites from each dataset. Intensities values are represented as lg2 values. The metabolites were mapped against the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway map hsa00380. Metabolites in italic font are annotated following *mummichog* analysis or HMDB matching method, while the rest are annotated using the SRM metabolites information. The boxplots representing the intensities of all the samples in each condition (red = infected, blue = control) in all three datasets.

6.1.2.3 Carbon metabolism

A significant increase in glucose was also noted across datasets (Fig 6.3) indicating alterations in central carbon metabolism. Indeed, perturbations to the glycolytic process and citric acid cycle, in particular, were confirmed in - ESI mode data particularly with significant decreases in lactate and malate 6.6. The significant increase in glucose is related to stress-induced hyperglycaemia which occurs in cases such as fever and infectious diseases [160]. This occurs due to the interaction between proinflammatory cytokines (TNF-α, IL-1, IL-6), the hypothalamic-pituitary axis and the noradrenergic system [160].

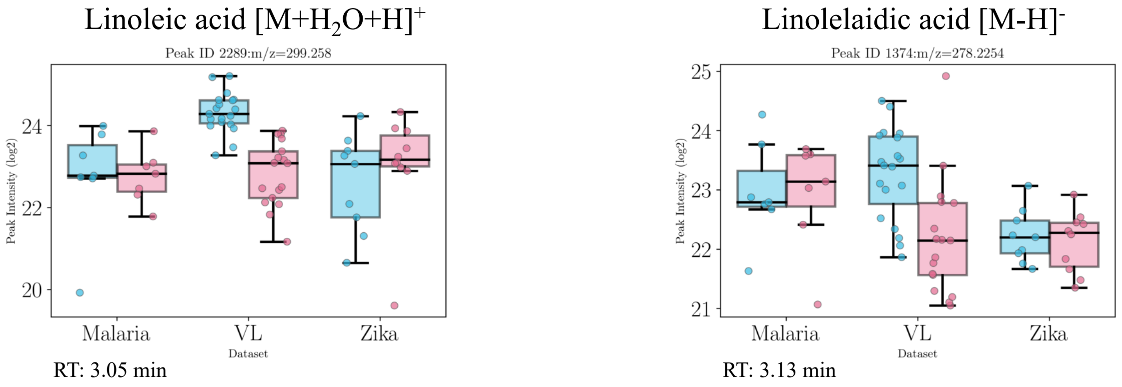
6.1.2.4 Lipid metabolism

Lipid abnormalities were also noted in the sera of infected patients, which demonstrated significant changes in the fatty acid metabolism, where there was a significant decrease mainly in particular acylcarnitines (Figure 6.2). Fatty acids, 3-hydroxybutanoate, 6-hydroxyhexanoate and octadecanoate were, on the other hand, all increased. This could be related to the previously mentioned increase in carnitine and subsequent decrease in acetylcarnitine levels which might indicate an increased activity in releasing fatty acids. Short-chain fatty acids are known to have a protective role and play a beneficial part in reducing endothelial activation, which leads to a reduction in proinflammatory cytokine production and adhesion molecule expression [161].

Sphingolipid and glycerophospholipid metabolism were also affected with sphingosine 1-phosphate being lower in all datasets. This might be indicative of liver damage, which is also reflected in the significantly modified levels of bile acids and taurine [162]. Bile acid biosynthesis metabolism also seems to be affected with taurine levels decreasing significantly in infected patients and chenodeoxycholate being significantly increased in infected patients. Choline and its derivatives were also downregulated in infected patients, which might also be related to macrophage metabolism [163].

Sphingolipids are not only important cellular membrane components, but they also have a dynamic role in cellular signalling and are involved in processes such as proliferation, endocytosis, necrosis, apoptosis and migration [164]. These are also a pathogen membrane component and have, thus, an important role in infectious diseases. Key molecules in sphingolipid signalling are: ceramide, sphingosine and sphingosine-1-phosphate (S1P). S1P is produced during inflammation or following tissue damage and it has been reported that pathogens affect S1P signalling via sphingosine kinase/S1P axis [165]. S1P exists both in the intracellular and extracellular pool, particularly in plasma. Plasma S1P has been shown to regulate various processes related to pathogenesis [166]. According to [166] reduced plasma S1P levels were associated with malaria severity in mice. There is also evidence that S1P bioavailability could have a role in attenuating endothelial damage. S1P also known to induce NO release from endothelial cells [167].

Linoelaidic acid was also putatively annotated. This is an isomer of linoleic acid, so without further confirmation the feature could also be linoleic acid, a precursor to eicosanoids. It was found to be significantly lower with a general downward trend in all datasets. Another feature was annotated as linoleic acid which was annotated both by *mummichog* and MS2 spectral validation and was also decreased in infected patients, albeit without reaching statistical significance, apart from D_Z where logFC_i was positive (0.4). Linoleic acid is a precursor for arachidonic acid from which prostaglandins and other bioactive eicosanoids are synthesised. In the context of fever, it could be hypothesized that an increased production of PGE2 due to the activation of COX-2 causes an increased turnover of arachidonic acid with a subsequent increased usage of linoleic acid, hence the decreased levels in serum. The LC-MS platform used here does not readily detect arachidonic acid or its products.



(a) **(b)**

Figure 6.7: (a) Boxplot of the intensities in each group and dataset for the compound putatively annotated as linoleic acid using *mummichog* and MS2 data. Linoleic acid is a precursor to arachidonic acid and prostaglandin E2 which plays an important role in the physiological mechanism of fever. Here the linoleic acid levels are generally lower in the infected group (red) which might be due to the increased production of arachidonic acid. (b) Peak putatively annotated as linoelaidic acid in negative mode by *mummichog*. Linoelaidic acid is an isomer of linoleic acid, so it could be either without further confirmation from tandem MS analysis. In a) the peakset represents an adduct of linoleic acid, hence the difference in m/z between the two. The downward trend is noticeable in both cases.

6.1.2.5 Nucleotide metabolism

Pyrimidine metabolism is significantly affected with cytosine being significantly higher in infected patients. Viral infections are known to cause significant metabolic changes in host cells, such as upregulation of pyrimidine nucleotide biosynthesis [168]. Uracil and its derivative, uridine, on the other hand, were found to be significantly decreased in all three datasets in the infected group. Purine metabolism is also affected, with adenine being significantly lower in infected patients.

6.1.3 Connecting the results with the molecular mechanisms of fever

The algorithm was written with the intention of comparing datasets related to infection, hence we sought differences between the two conditions from all three datasets using linear regression from limma and *mummichog* to determine the biological network of activity and significantly affected pathways in the infected samples group. This study offers a route to identify commonality in the metabolic profile in infected patients affected by pathogens that cause fever. At the centre of this meta-dataset stands the relationship between the kynurenine and tryptophan metabolites, as also identified by *mummichog* pathway analysis.

The existing proposed molecular basis for fever involves the following. The innate immune system is activated through pathogen recognition by toll-like receptors e.g. TLR-4. There are several pathogen-derived recognition molecules, with the most studied ones being the lipopolysaccharides (LPS) [159]. LPS interact with TLR-4 which further induce activation of nuclear factor κB (NF-κB). This, in turn, initiates the production of endogenous pyrogenic cytokines (IL-6, IL-1β, TNF-α) [159]. These pyrogenic cytokines then act on the organum vasculosus of the laminae terminalis in the hypothalamus. They also trigger the release of arachidonic acid from membrane phospholipids which is converted to prostaglandin H_2 (PGH_2) via the activation of cyclooxygenase-2 (COX-2), the rate limiting enzyme in the synthesis of prostaglandins [159]. The microsomal prostaglandin E_2 synthase (mPGES-1) then converts PGH_2 into prostaglandin E_2 (PGE_2) which acts on the pre-optic nucleus in the hypothalamus leading to an elevated temperature set-point. Additional negative feedback systems prevent excessive elevation of body temperature via antipyretic cytokines (IL-1Ra, IL-10, TNF-α binding protein) [1]. Other inflammatory mediators, apart from PGE_2, which could act as pyrogens are: bradykinin, corticotropin releasing hormone, nitric oxide [169], endothelin and macrophage inflammatory protein 1 (MIP-1) [159].

Based on the obtained results, a possible connection between fever and the kynurenine pathway from tryptophan metabolism could be explained by the interplay between IDO-1 and COX-2. An activation of IDO-1 could lead to a decreased inhibition of COX-2 which in turn could lead to an increased activation of PGE_2 release. The link between the two enzymes and inflammation has been previously studied. COX-2 enzyme activity suppression could also be decreased by the lower levels of niacinamide in the infected group, as niacinamide has been shown to influence its activity [170]. It is worth noting that serum metabolomics, as used here, detects only a faint echo of the changes that are occuring in specific cell types orchestrating inflammation in local anatomical sites associated with infection. Although PGE2 was not annotated in these datasets (prostaglandins being difficult to detect using the platform used), linoelaidic acid and linoleic acid were putatively annotated and found to be decreased. Linoleic acid is a precursor of arachidonic acid and bioactive eicosanoids such as PGE_2. The general decreased level in the infected group could point to an increased production of arachidonic acid and, thus, PGE_2. Setting these results into the context of the ongoing pandemic, recent metabolomics investigations on SARS-CoV2 infection in coronavirus patients also pinpointed significant alterations in the tryptophan-kynurenine pathway [171,172], with kynurenine levels increasing as disease severity did, although it would appear kynurenine increase is a general feature of febrile illness rather than a coronavirus specific response.

Many of the other metabolites found to have significant abundance level differences between the two groups have also been studied before in relation to their role in immunometabolism, due to their effect in altering the expression of either pro-inflammatory or anti-inflammatory cytokines. Glutamine, for instance, plays a major role in the immune system, as a key energy source for immune cells and it is used at a higher rate during catabolic conditions such as sepsis or other infections [155]. As suggested by [173] metabolites such as depleted glutamine and citrulline identified in this study could also be used as indicators of disease severity. Decreased oxoproline levels were also previously associated with a non-infectious fever-associated disease, Rheumatoid Arthritis [174].

It should be noted that while some of the commonly affected metabolites could indeed be related to fever or inflammation, others, such as those involved in pyrimidine metabolism, could be related strictly to the pathogenic aspect of the studied datasets, as previously noted above.

6.2 Metabolites specific to individual datasets

The sample size of the investigated datasets was small, a common limitation in metabolomics studies. D_M had only 14 samples in total, D_Z 20 samples and D_{VL} 40 samples. This can affect the statistical significance of the results when comparing the infected and control samples. Hence, integrating the datasets provided an increased sample size and statistical robustness. However, when integrating the datasets, apart from the features which behave in similar manner in infected vs. control based on the logFC_i computed for each dataset, features which behave in a specific manner to only one of the datasets can be more easily detected when using each dataset's logFC_i value to differentiate it from the other dataset with opposite logFC_i values. However, it should be noted that because of the small sample size of each dataset, the statistical significance of the fold change value between the conditions is also affected.

It should be noted that this is only putative annotation and hence the reporting of results may not be totally accurate. In this investigation, however, annotation using SRM was considered most reliable. After corroborating these results, and combining any duplicates, the metabolites specific to one dataset are listed in Table 6.2 below. The metabolites which were also detected in common for all datasets were filtered out. Additionally, the peaks with the same annotation, but opposing logFC_m values were also filtered out. For this analysis, peaksets for each individual dataset which did not align during the integration process were also taken into consideration.

D_M In the case of the malaria dataset there were 149 features with logFC values opposite from the other two datasets, out of which one was statistically significant. In the ascending order of their logFC values, the putatively annotated features are enumerated next with attached information on the peakset ID and the method of annotation: a) downregulated: cortisone (2472, mm), creatinine (175, ms2), imidazole-4-acetate (325, srm); n1-methyl-2-pyridone-5-carboxamide (716, mm), 3-hydroxyanthranilate (712, mm); b) upregulated: cystine (1839, srm), methyl indole-3-acetate (1229, mm), hippuric acid (3642, ms2). From the *mummichog* pathway analysis results methionine metabolism stood out for D_M for the upregulated features. For the - ESI mode results: a) downregulated: arachidonic acid (1458, 1453, mm), glycerol (69, mm), glycocholate (3154, mm), docosahexaenoic acid (1515, mm); b) upregulated: methyl indole-3-acetate (856, mm). Additionally, signficantly upregulated adenosine in D_M was identified in the peaksets specific to the dataset, i.e. which were not aligned with the peaksets from the other datasets.

D_Z In the case of D_Z there were 145 features with logFC values opposite from the other two datasets, out of which two were statistically significant for the dataset. In the ascending order of their logFC values the putatively annotated features are the following: a) downregulated:

epinephrine (743, mm), n(pi)-methyl-l-histidine (10779, srm). For the - ESI mode results: a) downregulated: 4-coumaryl alcohol (532, mm), choline phosphate (660, mm).

D$_{VL}$ In the case of D$_{VL}$ there were 151 features with logFC values opposite from the other two datasets, out of which 57 were statistically significant for the dataset. The higher number of statistically significant logFC values in the case of this dataset stem mainly from the higher number of samples (N=40) compared to the other two datasets. Most of the highly downregulated features were not annotated by mm/ms2/srm method. Manual annotation was performed in this case, most of them being assigned to various phophatidyl cholines: PC 38:4 (3022), PC p-38:5 (2958), PC 40:7 (3000). In the ascending order of their logFC values the putatively annotated features are the following: a) downregulated: porphobilinogen (177, mm); b) upregulated: phenylacetylglycine (897, srm), ornithine (198, mm), 2-phenylacetamide (236, mm), creatinine (183, srm, ms2,mm); $|\text{logFC}|<1$: pyridoxine (1191, srm). For the - ESI mode results: a) downregulated: 2-oxobutanoate (120, srm), b) upregulated: d-threose (271, srm), d-erythrose (447, srm), l-iditol (797, mm). Additionally, lysine and dihydrouracil were significantly upregulated in the individual D$_{VL}$.

Metabolites	**D_M**	**D_Z**	**D_{VL}**	**Taxonomy**
2-phenylacetamide			↑↑	amide derivatives
creatinine			↑↑	amino acids and derivatives
cystine	↑↑			amino acids and derivatives
lysine			↑↑	amino acids and derivatives
n(pi)-methylhistidine		↓↓		amino acids and derivatives
ornithine			↑↑	amino acids and derivatives
phenylacetylglycine			↑↑	amino acids and derivatives
porphobilinogen			↓↓	amino acids and derivatives
erythrose			↑↑	carbohydrates and derivatives
glycerol	↓↓			carbohydrates and derivatives
iditol			↑↑	carbohydrates and derivatives
threose			↑↑	carbohydrates and derivatives
2-oxobutanoate			↓↓	carboxylic acids and derivatives
hippuric acid	↑↑			carboxylic acids and derivatives
3-hydroxyanthranilate	↓			carboxylic acids and derivatives
epinephrine		↓↓		catecholamines
4-coumaryl alcohol		↓↓		cynnamyl alcohols
imidazole-4-acetate	↓↓			imidazoles and derivatives
methyl indole-3-acetate	↑↑			indoles and derivatives
arachidonic acid	↓↓			lipids and derivatives
docosahexanoic acid	↓↓			lipids and derivatives
n1-methyl-2-pyridone-5-carboxamide	↓			nicotinamides
adenosine	↑↑			purines and derivatives
dihydrouracil			↑↑	pyrimidines and derivatives
PC 38:4			↓↓	phophatidylcholines
PC 40:7			↓↓	phophatidylcholines
PC p- 38:5			↓↓	phophatidylcholines
choline phosphate		↓		phosphocholines
pyridoxine			↑	pyridoxines
glycocholate	↓↓			steroids and derivatives
cortisone	↓↓			steroids and derivatives

Table 6.2: Putatively annotated features specific to individual datasets. The arrows symbolise whether the metabolite was found to be upregulated or downregulated. Double arrows signify a $| logFC |> 1$. Metabolites which were also found to be in common between the datasets were not highlighted and not considered for further evaluation. The metabolites highlighted in light blue were found to be significant in the individual datasets which where not integrated.

In case of malaria, [173] identified that amino acid metabolism was predominantly affected, especially arginine because of lower glutamine and proline, low tryptophan, elevated kynurenine and a reduction in certain lipids (sphingomyelins). In the current analysis, these were all found to be commonly modified between the datasets. This provides a good example of why learning about the common changes between infectious diseases with similar underlying symptoms is helpful to prevent against assuming that these might be malaria specific. Metabolites which were downregulated in D_M, but upregulated in the other two datasets, were: imidazole-4-acetate, glycerol, 3-hydroxyanthranilate, arachidonic acid, docosahexanoic acid, n1-methyl-2-pyridone-5-carboxamide, glycocholate and cortisone. Imidazole-4-acetate is a derivative of histidine, whose levels increases as a response malaria, and of aminoisoquinolines used to treat malaria [175]. Probably the decreased levels could suggest an increased uptake of imidazole-4-acetate from the blood stream. The decreased levels of arachidonic acid could be related to the fact that for D_M, malaria was induced in a group of subjects in a controlled environment and fever symptoms were less exacerbated compared to the other two datasets. Thus, arachidonic acid which is a fever mediator, was not elevated in this case. Metabolites which were upregulated in D_M, but downregulated in the other two, were: cystine, methyl indole-3-acetate, hippuric acid and adenosine. In an early study on malaria infected erythrocites, an increase in adenosine deaminase activity was observed [176]. Additionally, [177] observed increased plasma adenosine levels in malaria infected monkeys. The increased adenosine levels obtained in D_M could reflect this increase in adenosine deaminase caused by malaria infection, as adenosine is its substrate. It is to be noted that adenine was decreased for all datasets. In the case of D_M the increase of adenosine could also reflect an increased breakdown of adenine.

In a study by [178] elevated levels of gut microbial acids, among which hippuric acid, were linked to acidosis caused during severe malaria. Their levels were also correlated with disease severity. Moreover, it has been suggested that the profile of the bacterial flora could also have an impact on the severity degree of the infection [179]. This disturbance of the normal gut barrier function could be related to the sequestration of infected red blood cells, which is central to the patophysiology of falciparum malaria. This gut barrier dysfunction may cause the translocation of microbial acids into the circulation. Biomarkers of gut integrity such as citrulline and arginine also presented a significant reduction [178]. Reduced levels of citrulline were also observed as a common disruption in the presently integrated datasets, which might indicate gut integrity was affected in all diseases.

In the case of D_{VL} phosphatidyl cholines were decreased in infected samples. This might be related to liver damage caused during the infection with *L.infantum*. Other metabolites which were downregulated in D_{VL}, but upregulated in the other two, were: porphobilinogen and 2-oxobutanoate (precursor in isoleucine biosynthesis). Metabolites which were upregulated in D_{VL}, but downregulated in the other two, were: 2-phenylacetamide, creatinine (consistent with findings in [180]), lysine, ornithine, phenylacetylglycine, dihydrouracil, pyridoxine, iditol and the sugars threose and erythrose. Some of the enumerated biomarkers are also used for determining certain physiological dysfunctions. For example, elevated creatinine could point to a renal dysfunction.

In the case of D_Z four putatively annotated features were specifically downregulated (n(pi)-methylhistidine, epinephrine, 4-coumaryl alcohol, choline phosphate) and one upregulated (hydroxykynurenine). Hidroxykynurenine is a metabolite involved in the kynurenine pathway which was disrupted in all datasets. It is to be noted that some of the metabolites identified, e.g. epinephrine, could also be derived from the medicine used to treat the disease.

In conclusion, several disease-specific biomarkers were identified for each dataset which did not coincide with the biomarkers found in common between the three datasets. Discussing each biomarker was beyond the scope of this thesis. However, several of the specific biomarkers identified were successfully matched against existing literature in confirming specific disease physiology.

6.2.1 Conclusion

By integrating LC-MS datasets in this manner both biomarkers underlying the commonality between different diseases and biomarkers specific to each disease could potentially be identified. The features detected in this investigation and outlined in this chapter could be used on the PD-CMOS platform, as it has already been demonstrated that it can be used for the detection of metabolites [181]. Additionally, this sort of approach also outlines the fact even though some metabolic biomarkers appear to be specific for certain disease, they could, in fact, represent a commonality between multiple disparate diseases.

Chapter 7

Conclusion and future research directions

Two principal methods for the detection of fever-associated diseases were investigated for this thesis: detection using a biosensor-based immunoassay and detection of fever-associated biomarkers to improve the diagnosis of relevant diseases. This thesis makes a series of contributions, summarised according to each chapter containing results in the list below. Additionally, several aspects were also identified which could be further addressed in future research work. These are detailed in the following paragraphs.

1. In Chapter 3, three computational methods were developed for the quantitative detection of a reaction spot after an immunoassay was run on the PD-CMOS biosensor platform. Two of the methods were developed using already existing packages while the third method was a new approach based on generative modelling. The three methods were then compared to determine which was the optimal one for signal processing. In the end, the method based on Sequential Monte Carlo algorithm proved to be the one which correctly estimated the spot intensity, especially at higher noise values.

   Improving the computational methods for the reaction spot detection In this chapter, only synthetically generated images mimicking those which would be obtained from the biosensor were used for developing and testing the three methods mentioned above. Images obtained following multiple immunoassays ran in the laboratory on the biosensor platform would be the next step necessary in further testing the developed algorithms. In this way, detailed information regarding the actual signal obtained from the biosensor, as well as the type and magnitude of noise, would be obtained and incorporated into the stochastic algorithms based on GP regression and SMC generative modelling. Moreover, the algorithm was developed only for the detection of one reaction spot. This could be further extended for the simultaneous detection of multiple reaction spots. Further research could thus be directed towards the evaluation of the three signal processing methods.

2. In Chapter 4, an approach to developing a PD-CMOS based immunosensor for the detection of fever-associated HAT disease specific antibodies using the recombinant antigens -rL1.3, rL1.5, rISG65, rISG75- was proposed.

   (a) The recombinant antigens were engineered using molecular cloning techniques in *E.coli*.

   (b) The immunoassay surface was chemically functionalised using APTES and GA in order to enable covalent attachment of the antigens onto the surface.

   (c) The PD-CMOS biosensor platform was tested for the detection of an antibody anti-ISG65, aimed at a specific trypanosome surface antigen, and exhibited good results based on its LOD and LOQ values.

   Suggestions for improving the immunosensor platform Due to the Covid-19 pandemic, laboratory work was suspended and future research would be required for improving the immunoassay, especially in the area of antigen printing on the surface and delivery of fluids. In this case, a microfluidics approach was suggested [86]. This would optimise the consistency in delivering the fluids during the immunoassay providing, thus, a more robust immunoassay format. For the antigen deposition, physical separators such as wells or channels could be used. Alternatively, microstamping methods or inkjet printing methods could be tested out. Moreover, other antigens and antibodies should be tested for coating the surface.

3. In Chapter 5, a computational approach to integrating multiple disparate metabolomics LC-MS datasets was proposed. This was performed through the alignment between the peaksets from each dataset after a non-linear feature-based warping method was developed and applied to correct the RT drift between the datasets. The warping method was based on Gaussian Process regression modelling, a non-parametric regression approach which works well with small training data. The new predicted RTs for each dataset were utilised for aligning the three infectious disease metabolomics datasets using a direct-matching method.

   RT drift correction algorithm The research and analysis conducted in this chapter has provided with a solid method in determining common perturbations in plasma molecules from patients infected with different pathogens. Future research directions which could be undertaken in this area are summarised next. Firstly, in terms of the alignment algorithm accuracy, further evaluation could be performed. Similarly to [142], this could be done by using a test sample which is run twice through the LC-MS platform using different run lengths. By applying the RT drift correction algorithm to one of the samples, the ability of all peaks to be found after aligning the two samples could be evaluated. Another method which could be used both for algorithm evaluation and improved molecular networking would be based on MS2 data analysis. In order to do this, acquiring specific MS2 data would be needed. For the data analysed in this PhD, the MS2 data was obtained from the ions with the highest intensity. However, if the MS2 data for all three datasets were to be obtained from all ions with statistical significance, then both evaluation and annotation would be improved. In the first place, this would allow for a more comprehensive comparison based on cosine similarity scores between the MS2 spectra from all datasets for each peakset, enabling thus a better evaluation of the alignment. Secondly, molecular networking could be performed by using the new MS2 data [182].

In terms of the accessibility of the developed method, a basic web application was developed for the moment, but further improvements and features could be added in order to develop the RT drift correction software. Other sets of metabolomics datasets could be used to determine further metabolites related to fever. For example, datasets from other infectious diseases or non-infectious diseases associated with fever could be integrated and commonality in terms of metabolite perturbation could be determined.

4. In Chapter 6, molecular mechanisms related to fever were confirmed following the analysis performed in Chapter 5. The biomarkers and molecular mechanisms specific to each of the fever-associated disease were also addressed. Several of the metabolites identified which of importance to fever mechanism include those related to the kynurenine pathway of the tryptophan metabolism. The molecules identified as potential fever biomarkers, or the enzymes involved in their synthesis could potentially be used on the biosensor platform used in the first part of the thesis in similar manner to [181].

In conclusion, through computational and laboratory-based methods, areas such as signal processing of an immunosensor platform, biomarker detection through a novel RT correction method were researched for improving the detection of fever associated diseases.

Appendix A

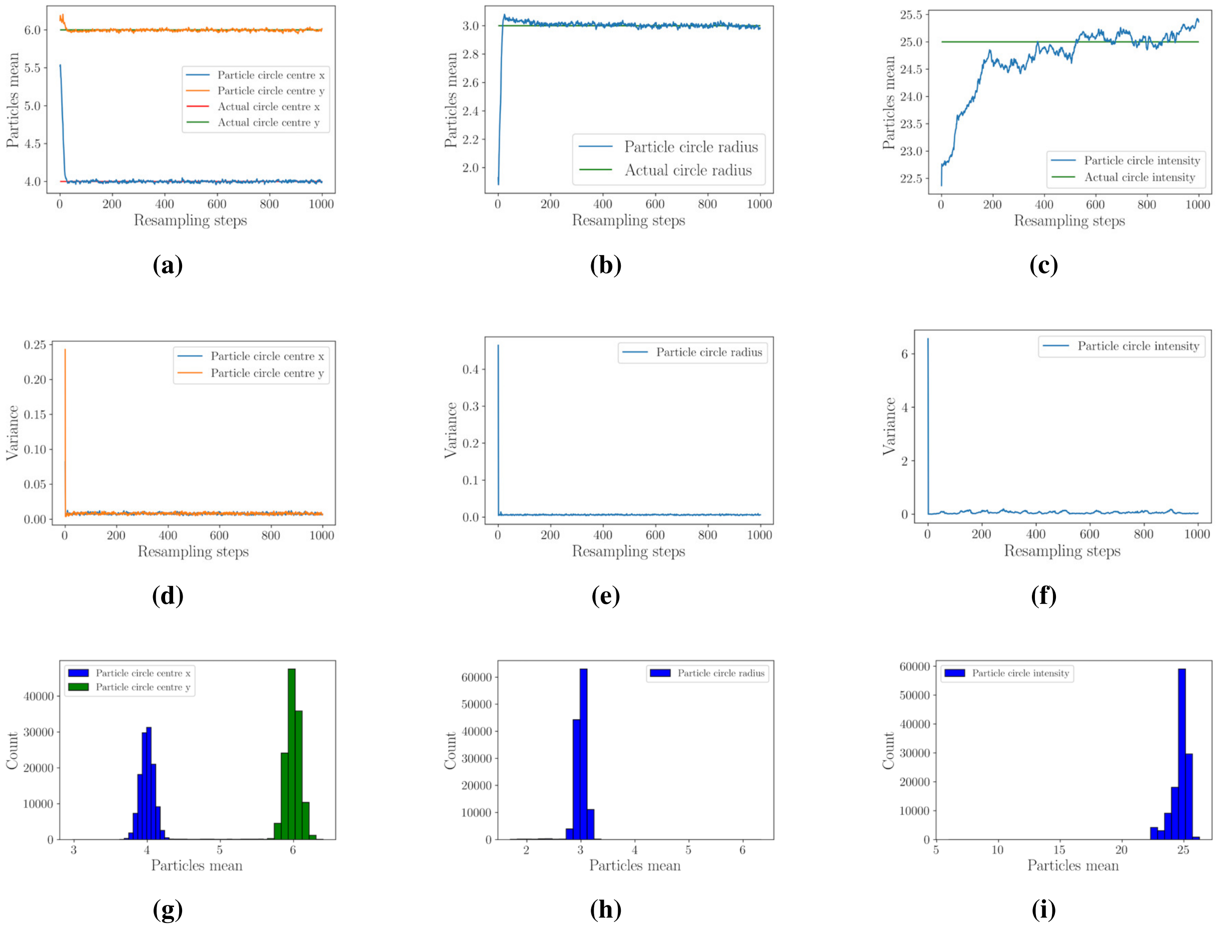
Background literature: Adducts

Ion	Reverse Calculation
$[M+H]^{+}$	mz - PROTON
$[M+2H]^{2+}$	(mz - PROTON)*2
$[M+3H]^{3+}$	(mz - PROTON)*3
$[M(^{13}C)+H]^{+}$	mz - 1.0034 - PROTON
$[M(^{13}C)+2H]^{2+}$	(mz - 0.5017 - PROTON)*2
$[M(^{13}C)+3H]^{3+}$	(mz - 0.3344 - PROTON)*3
$[M(^{34}S)+H]^{+}$	mz -1.9958 - PROTON
$[M(^{37}Cl)+H]^{+}$	mz -1.9972 - PROTON
$[M+Na]^{+}$	mz - 21.9820 - PROTON
$[M+H+Na]^{2+}$	(mz - 10.991 - PROTON)*2
$[M+K]^{+}$	mz - 37.9555 - PROTON
$[M+H_2O+H]^{+}$	mz - 18.0106 - PROTON
$[M-H_2O+H]^{+}$	mz + 18.0106 - PROTON
$[M-H_4O_2+H]^{+}$	mz + 36.0212 - PROTON
$[M-NH_3+H]^{+}$	mz + 17.0265 - PROTON
$[M-CO+H]^{+}$	mz + 27.9950 - PROTON
$[M-CO_2+H]^{+}$	mz + 43.9898 - PROTON
$[M-HCOOH+H]^{+}$	mz + 46.0054 - PROTON
$[M+HCOONa]^{+}$	mz - 67.9874 - PROTON
$[M-HCOONa+H]^{+}$	mz + 67.9874 - PROTON
$[M+NaCl]^{+}$	mz - 57.9586 - PROTON
$[M-C_3H_4O_2+H]^{+}$	mz + 72.0211 - PROTON
$[M+HCOOK]^{+}$	mz - 83.9613 - PROTON
$[M-HCOOK+H]^{+}$	mz + 83.9613 - PROTON
$[M-H]^{-}$	mz + PROTON
$[M-2H]^{2-}$	(mz + PROTON)*2
$[M-3H]^{3-}$	(mz + PROTON)*3
$[M-H_2O-H]^{-}$	mz + PROTON + 18.0111135
$[M+Na-2H]^{-}$	mz - 20.974666
$[M+Cl]^{-}$	mz - 34.969402
$[M+K-2H]^{-}$	mz - 36.948606
$[M+FA-H]^{-}$	mz - 44.998201
$[M+Hac-H]^{-}$	mz - 59.013851
$[M+Br]^{-}$	mz - 78.918885
$[M+TFA-H]^{-}$	mz - 112.985586
$[2M-H]^{-}$	(mz + PROTON)/2
$[2M+FA-H]^{-}$	(mz - 44.998201)/2
$[2M+Hac-H]^{-}$	(mz - 59.013851)/2
$[3M-H]^{-}$	(mz + PROTON)/3

Table A.1: Formula for reverse calculating the m/z of adducts for positive and negative ESI mode data, where PROTON = 1.00727646677. Obtained from [183]. TFA = trifluoroacetate, FA = formic acid, Hac = acetic acid.

Appendix B

Evaluating the SMC based algorithm



(a)

(b)

(c)

(d)

(e)

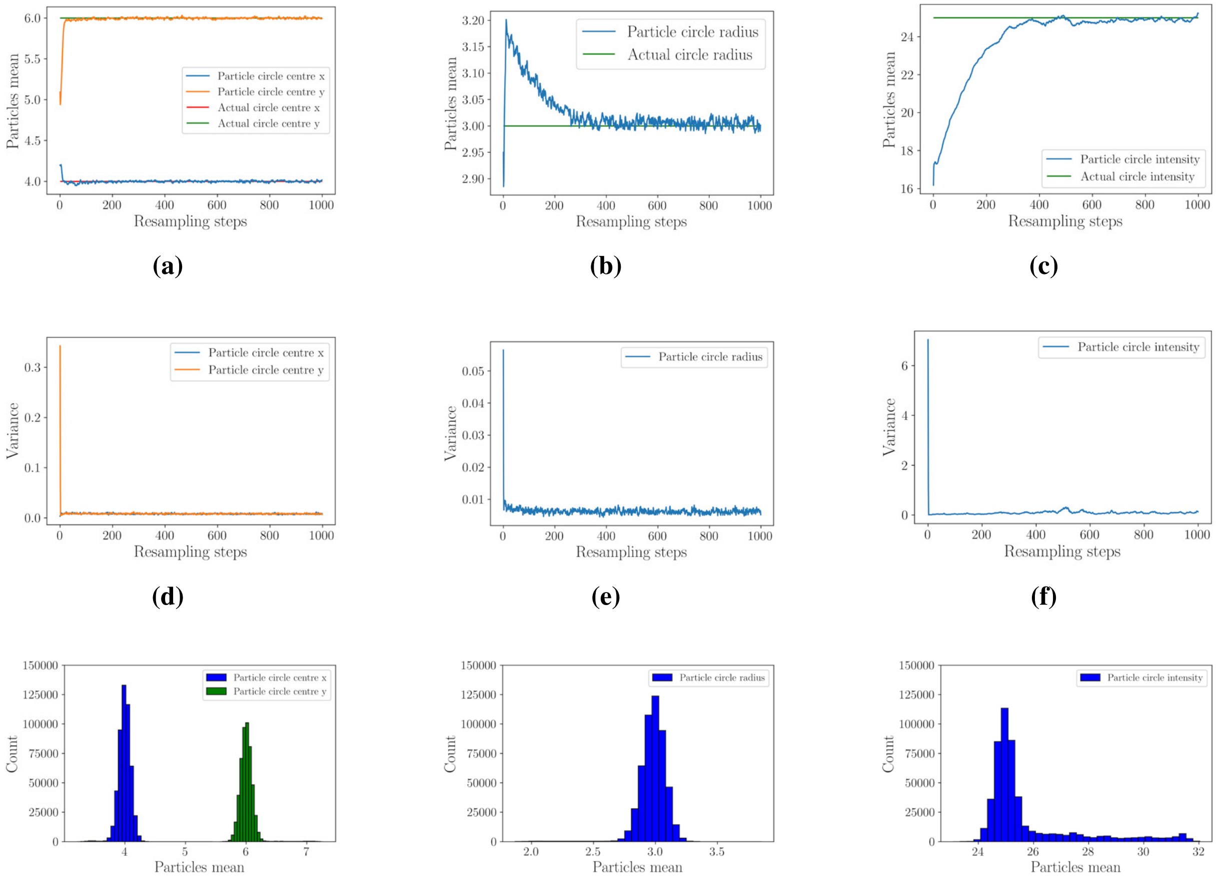
(f)

(g)

(h)

(i)

Figure B.1: Results for the particle filtering algorithm for N = 125 particle, X = 1000 resampling steps and $\sigma_{weight} = 10$. The top figures (a-c) represent the resampling steps it takes to reach convergence to the artificial image. The mean of the particles at each resampling step were computed for the x,y circle centre coordinates (pixel location), radius and intensity. Figures (d-f) represent the variance of the particles' characteristics at each resampling step. Figures (g-i) represent the distribution of the particles' characteristics at each resampling step.

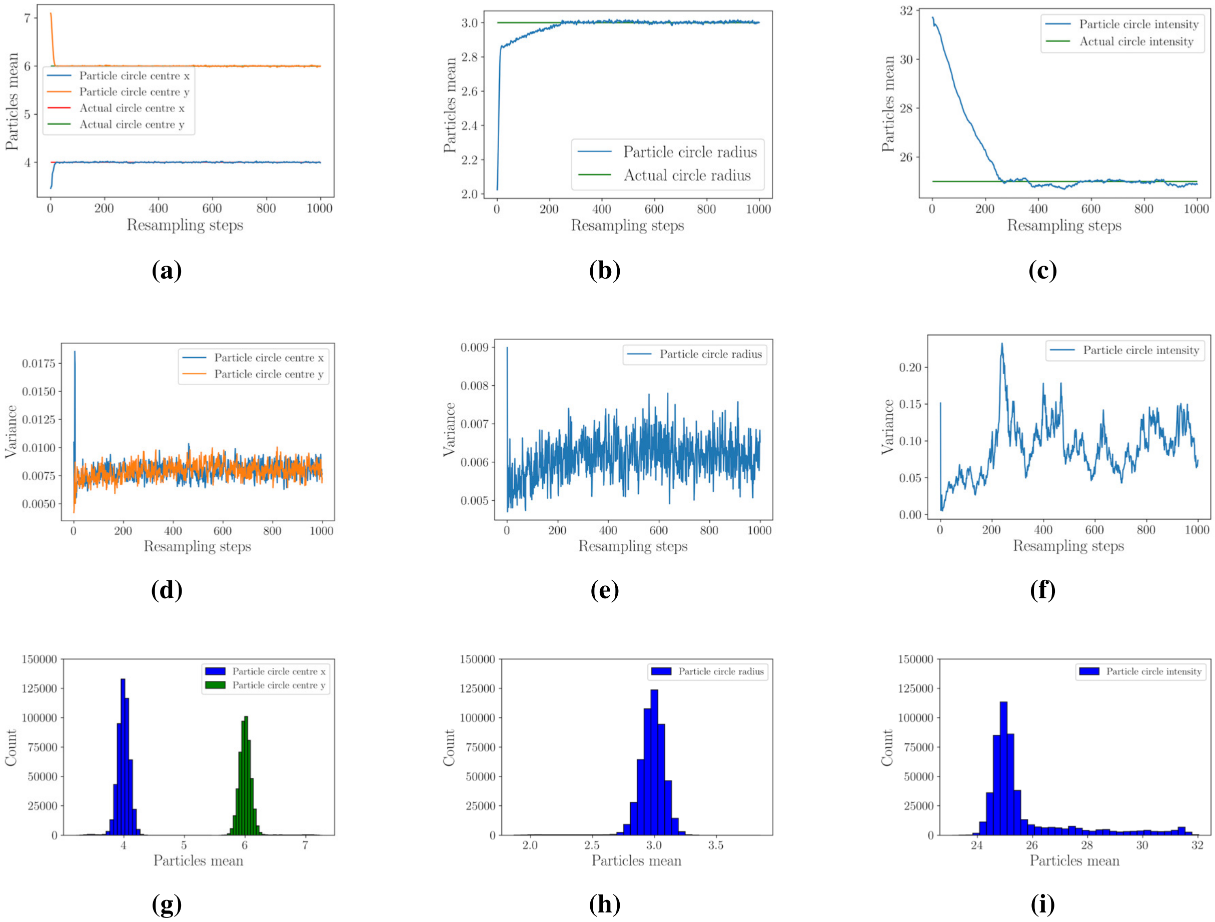


(a) **(b)** **(c)**

(d) **(e)** **(f)**

(g) **(h)** **(i)**

Figure B.2: Results for the particle filtering algorithm for N = 250 particle, X = 1000 resampling steps and $\sigma_{weight} = 10$. The top figures (a-c) represent the resampling steps it takes to reach convergence to the artificial image. The mean of the particles at each resampling step were computed for the x,y circle centre coordinates (pixel location), radius and intensity. Figures (d-f) represent the variance of the particles' characteristics at each resampling step. Figures (g-i) represent the distribution of the particles' characteristics at each resampling step.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

Figure B.3: Results for the particle filtering algorithm for N = 500 particle, X = 1000 resampling steps and $\sigma_{weight} = 10$. The top figures (a-c) represent the resampling steps it takes to reach convergence to the artificial image. The mean of the particles at each resampling step were computed for the x,y circle centre coordinates (pixel location), radius and intensity. Figures (d-f) represent the variance of the particles' characteristics at each resampling step. Figures (g-i) represent the distribution of the particles' characteristics at each resampling step.

Bibliography

[1] Walter EJ, Hanna-Jumma S, Carraretto M, Forni L. The Pathophysiological Basis and Consequences of Fever. Critical Care. 2016;20. doi:10.1186/s13054-016-1375-5.

[2] Wangdi K, Kasturiaratchi K, Nery SV, Lau CL, Gray DJ, Clements ACA. Diversity of Infectious Aetiologies of Acute Undifferentiated Febrile Illnesses in South and Southeast Asia: A Systematic Review. BMC Infectious Diseases. 2019;19(1):577. doi:10.1186/s12879-019-4185-y.

[3] Escadafal C, Nsanzabana C, Archer J, Chihota V, Rodriguez W, Dittrich S. New Biomarkers and Diagnostic Tools for the Management of Fever in Low- and Middle-Income Countries: An Overview of the Challenges. Diagnostics. 2017;7(3). doi:10.3390/diagnostics7030044.

[4] Patil SB, Dheeman DS, Al-Rawhani MA, Velugotla S, Nagy B, Cheah BC, et al. An Integrated Portable System for Single Chip Simultaneous Measurement of Multiple Disease Associated Metabolites. Biosensors and Bioelectronics. 2018;122:88–94. doi:10.1016/j.bios.2018.09.013.

[5] Nagy B, Al-Rawhani MA, Cheah BC, Barrett MP, Cumming DRS. Immunoassay Multiplexing on a Complementary Metal Oxide Semiconductor Photodiode Array. ACS Sensors. 2018;3(5):953–959. doi:10.1021/acssensors.7b00972.

[6] Năstase AM, Barrett MP, Cárdenas WB, Cordeiro FB, Zambrano M, Andrade J, et al. Alignment of Multiple Metabolomics LC-MS Datasets from Disparate Diseases to Reveal Fever-Associated Metabolites. bioRxiv. 2021; p. 2021.09.02.458540. doi:10.1101/2021.09.02.458540.

[7] Nayak S, Blumenfeld NR, Laksanasopin T, Sia SK. Point-of-Care Diagnostics: Recent Developments in a Connected Age. Analytical Chemistry. 2017;89(1):102–123. doi:10.1021/acs.analchem.6b04630.

[8] Liu G, Lin Y. Electrochemical Sensor for Organophosphate Pesticides and Nerve Agents Using Zirconia Nanoparticles as Selective Sorbents. Analytical Chemistry. 2005;77(18):5894–5901. doi:10.1021/ac050791t.

[9] Clark LC, Lyons C. Electrode Systems for Continuous Monitoring in Cardiovascular Surgery. Annals of the New York Academy of Sciences. 1962;102(1):29–45. doi:10.1111/j.1749-6632.1962.tb13623.x.

[10] Bhalla N, Jolly P, Formisano N, Estrela P. Introduction to Biosensors. Essays in Biochemistry. 2016;60(1):1–8. doi:10.1042/EBC20150001.

[11] Hall EAH. Biosensors. Open University Press; 1990.

[12] Rodovalho VR, Alves LM, Castro ACH, Brito-Madurro AG, Santos AR. Biosensors Applied to Diagnosis of Infectious Diseases – An Update. Austin Journal of Biosensors & Bioelectronics. 2015;1(3):12.

[13] Mehrotra P. Biosensors and Their Applications - A Review. Journal of Oral Biology and Craniofacial Research. 2016 May-Aug;6(2):153–159. doi:10.1016/j.jobcr.2015.12.002.

[14] Sin ML, Mach KE, Wong PK, Liao JC. Advances and Challenges in Biosensor-Based Diagnosis of Infectious Diseases. Expert Review of Molecular Diagnostics. 2014;14(2):225–244. doi:10.1586/14737159.2014.888313.

[15] Nambiar S, Yeow JTW. Conductive Polymer-Based Sensors for Biomedical Applications. Biosensors and Bioelectronics. 2011;26(5):1825–1832. doi:10.1016/j.bios.2010.09.046.

[16] Viguier C, Viguier C, Crean C, O'Kennedy R. Trends and Perspectives in Immunosensors. In: Antibodies Applications and New Developments. Bentham Books; 2012. p. 184–208.

[17] de la Escosura-Muñiz A, Parolo C, Merkoçi A. Immunosensing Using Nanoparticles. Materials Today. 2010;13(7):24–34. doi:10.1016/S1369-7021(10)70125-5.

[18] Sharma MK, Rao VK, Agarwal GS, Rai GP, Gopalan N, Prakash S, et al. Highly Sensitive Amperometric Immunosensor for Detection of Plasmodium Falciparum Histidine-Rich Protein 2 in Serum of Humans with Malaria: Comparison with a Commercial Kit. Journal of Clinical Microbiology. 2008;46(11):3759–3765. doi:10.1128/JCM.01022-08.

[19] Sharma MK, Agarwal GS, Rao VK, Upadhyay S, Merwyn S, Gopalan N, et al. Amperometric Immunosensor Based on Gold Nanoparticles/Alumina Sol–Gel Modified Screen-Printed Electrodes for Antibodies to Plasmodium Falciparum Histidine Rich Protein-2. Analyst. 2010;135(3):608–614. doi:10.1039/B918880K.

[20] Sharma MK, Rao VK, Merwyn S, Agarwal GS, Upadhyay S, Vijayaraghavan R. A Novel Piezoelectric Immunosensor for the Detection of Malarial Plasmodium Falciparum Histidine Rich Protein-2 Antigen. Talanta. 2011;85(4):1812–1817. doi:10.1016/j.talanta.2011.07.008.

[21] Sikarwar B, Sharma PK, Srivastava A, Agarwal GS, Boopathi M, Singh B, et al. Surface Plasmon Resonance Characterization of Monoclonal and Polyclonal Antibodies of Malaria for Biosensor Applications. Biosensors and Bioelectronics. 2014;60:201–209. doi:10.1016/j.bios.2014.04.025.

[22] Hemben A, Ashley J, Tothill IE. Development of an Immunosensor for PfHRP 2 as a Biomarker for Malaria Detection. Biosensors. 2017;7(3). doi:10.3390/bios7030028.

[23] Souto DEP, Silva JV, Martins HR, Reis AB, Luz RCS, Kubota LT, et al. Development of a Label-Free Immunosensor Based on Surface Plasmon Resonance Technique for the Detection of Anti-Leishmania Infantum Antibodies in Canine Serum. Biosensors and Bioelectronics. 2013;46:22–29. doi:10.1016/j.bios.2013.01.067.

[24] Cabral-Miranda G, de Jesus JR, Oliveira PRS, Britto GSG, Pontes-de-Carvalho LC, Dutra RF, et al. Detection of Parasite Antigens in Leishmania Infantum–Infected Spleen Tissue by Monoclonal Antibody-, Piezoelectric-Based Immunosensors. Journal of Parasitology. 2014;100(1):73–78. doi:10.1645/GE-3052.1.

[25] Kaushik A, Yndart A, Kumar S, Jayant RD, Vashist A, Brown AN, et al. A Sensitive Electrochemical Immunosensor for Label-Free Detection of Zika-virus Protein. Scientific Reports. 2018;8(1):9700. doi:10.1038/s41598-018-28035-3.

[26] Zelada-Guillén GA, Tweed-Kent A, Niemann M, Göringer HU, Riu J, Rius FX. Ultrasensitive and Real-Time Detection of Proteins in Blood Using a Potentiometric Carbon-Nanotube Aptasensor. Biosensors and Bioelectronics. 2013;41:366–371. doi:10.1016/j.bios.2012.08.055.

[27] Foxman B. Chapter 5 - A Primer of Molecular Biology. In: Foxman B, editor. Molecular Tools and Infectious Disease Epidemiology. San Diego: Academic Press; 2012. p. 53–78.

[28] Bintimdsallah SS, Amin MS, Mohamed H, Mamun M. Cmos Downsizing: Present, Past And Future. Journal of Applied Sciences Research. 2012;8(8).

[29] Al-Rawhani MA, Cheah BC, Macdonald AI, Martin C, Hu C, Beeley J, et al. A Colorimetric CMOS-Based Platform for Rapid Total Serum Cholesterol Quantification. IEEE Sensors Journal. 2017;17(2):240–247. doi:10.1109/JSEN.2016.2629018.

[30] Marimuthu M, Kandasamy K, Ahn CG, Sung GY, Kim MG, Kim S. CMOS Image Sensor for Detection of Interferon Gamma Protein Interaction as a Point-of-Care Approach. Analytical and Bioanalytical Chemistry. 2011;401(5):1641. doi:10.1007/s00216-011-5231-9.

[31] Devadhasan JP, Kim S. CMOS Image Sensor Based HIV Diagnosis: A Smart System for Point-of-Care Approach. BioChip Journal. 2013;7(3):258–266. doi:10.1007/s13206-013-7309-2.

[32] Baader J, Klapproth H, Bednar S, Brandstetter T, Rühe J, Lehmann M, et al. Polysaccharide Microarrays with a CMOS Based Signal Detection Unit. Biosensors and Bioelectronics. 2011;26(5):1839–1846. doi:10.1016/j.bios.2010.01.021.

[33] Klapproth H, Bednar S, Baader J, Lehmann M, Freund I, Brandstetter T, et al. Development of a Multi-Analyte CMOS Sensor for Point-of-Care Testing. Sensing and Bio-Sensing Research. 2015;5:117–122. doi:10.1016/j.sbsr.2015.08.004.

[34] Bishop C. Pattern Recognition and Machine Learning. Information Science and Statistics. New York: Springer-Verlag; 2006.

[35] World Population Prospects - Population Division - United Nations;. https://population.un.org/wpp/.

[36] Rogers S, Girolami M. A First Course in Machine Learning. Chapman and Hall/CRC; 2016.

[37] Rasmussen CE. Gaussian Processes in Machine Learning. In: Bousquet O, von Luxburg U, Rätsch G, editors. Advanced Lectures on Machine Learning. vol. 3176. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004. p. 63–71.

[38] Micchelli CA, Xu Y, Zhang H. Universal Kernels. Journal of Machine Learning Research. 2006;7:17.

[39] Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning. Cambridge, Mass: MIT Press; 2006.

[40] GPy. GPy: A Gaussian Process Framework in Python; 2014. http://github.com/SheffieldML/GPy.

[41] Kalaitzis A, Honkela A, Gao P, Lawrence ND. Gptk: Gaussian Processes Tool-Kit; 2014.

[42] Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC. Reproducibility Standards for Machine Learning in the Life Sciences. Nature Methods. 2021;18(10):1132–1135. doi:10.1038/s41592-021-01256-7.

[43] Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine Learning in Bioinformatics. Briefings in Bioinformatics. 2006;7(1):86–112. doi:10.1093/bib/bbk007.

[44] Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: Beyond Biomarkers and towards Mechanisms. Nature Reviews Molecular Cell Biology. 2016;17(7):451–459. doi:10.1038/nrm.2016.25.

[45] Zamboni N, Saghatelian A, Patti GJ. Defining the Metabolome: Size, Flux, and Regulation. Molecular cell. 2015;58(4):699–706. doi:10.1016/j.molcel.2015.04.021.

[46] Piras V, Tomita M, Selvarajoo K. Is Central Dogma a Global Property of Cellular Information Flow? Frontiers in Physiology. 2012;3:439. doi:10.3389/fphys.2012.00439.

[47] Yang Q, Zhang Ah, Miao Jh, Sun H, Han Y, Yan Gl, et al. Metabolomics Biotechnology, Applications, and Future Trends: A Systematic Review. RSC Advances. 2019;9(64):37245–37257. doi:10.1039/C9RA06697G.

[48] Vossman. A Hairpin Loop from a Pre-mRNA; Accessed on 4 October 2021. https://commons.wikimedia.org/wiki/File:Pre-mRNA-1ysv-tubes.png.

[49] Bcndoye. Crystal Structure of C-lactate Dehydrogenase; Accessed on 4 October 2021. https://commons.wikimedia.org/wiki/File:Lactate_Dehydrogenase_C.png.

[50] Salek R. Metabolomics - Molecules of Life; Accessed on 4 October 2021. https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/Metabolomics%20-%20molecules%20of%20life.pdf.

[51] Li S. Computational Methods and Data Analysis for Metabolomics. Humana Press; 2020.

[52] González-Domínguez R, González-Domínguez Á, Segundo C, Schwarz M, Sayago A, Mateos RM, et al. High-Throughput Metabolomics Based on Direct Mass Spectrometry Analysis in Biomedical Research. In: D'Alessandro A, editor. High-Throughput Metabolomics: Methods and Protocols. Methods in Molecular Biology. New York, NY: Springer; 2019. p. 27–38.

[53] Pitt JJ. Principles and Applications of Liquid Chromatography-Mass Spectrometry in Clinical Biochemistry. The Clinical Biochemist Reviews. 2009;30(1):19–34.

[54] Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. Science. 1989;246(4926):64–71. doi:10.1126/science.2675315.

[55] Olsen JV, Macek B, Lange O, Makarov A, Horning S, Mann M. Higher-Energy C-trap Dissociation for Peptide Modification Analysis. Nature Methods. 2007;4(9):709–712. doi:10.1038/nmeth1060.

[56] Xiao JF, Zhou B, Ressom HW. Metabolite Identification and Quantitation in LC-MS/MS-based Metabolomics. Trends in analytical chemistry : TRAC. 2012;32:1–14. doi:10.1016/j.trac.2011.08.009.

[57] Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. HMDB 4.0: The Human Metabolome Database for 2018. Nucleic Acids Research. 2018;46(D1):D608–D617. doi:10.1093/nar/gkx1089.

[58] Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, et al. LMSD: LIPID MAPS Structure Database. Nucleic Acids Research. 2007;35(Database issue):D527–532. doi:10.1093/nar/gkl838.

[59] Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: A Public Repository for Sharing Mass Spectral Data for Life Sciences. Journal of Mass Spectrometry. 2010;45(7):703–714. doi:10.1002/jms.1777.

[60] Pence HE, Williams A. ChemSpider: An Online Chemical Information Resource. Journal of Chemical Education. 2010;87(11):1123–1124. doi:10.1021/ed100697w.

[61] Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Analytical Chemistry. 2006;78(3):779–787. doi:10.1021/ac051437y.

[62] Pluskal T, Castillo S, Villar-Briones A, Orešič M. MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data. BMC Bioinformatics. 2010;11(1):395. doi:10.1186/1471-2105-11-395.

[63] Du X, Smirnov A, Pluskal T, Jia W, Sumner S. Metabolomics Data Preprocessing Using ADAP and MZmine 2. In: Li S, editor. Computational Methods and Data Analysis for Metabolomics. Methods in Molecular Biology. New York, NY: Springer US; 2020. p. 25–48.

[64] Zhang W, Zhao PX. Quality Evaluation of Extracted Ion Chromatograms and Chromatographic Peaks in Liquid Chromatography/Mass Spectrometry-Based Metabolomics Data. BMC Bioinformatics. 2014;15(11):S5. doi:10.1186/1471-2105-15-S11-S5.

[65] Myers OD, Sumner SJ, Li S, Barnes S, Du X. Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. Analytical Chemistry. 2017;89(17):8689–8695. doi:10.1021/acs.analchem.7b01069.

[66] Feng X, Zhang W, Kuipers F, Kema I, Barcaru A, Horvatovich P. Dynamic Binning Peak Detection and Assessment of Various Lipidomics Liquid Chromatography-Mass

Spectrometry Pre-Processing Platforms. Analytica Chimica Acta. 2021;1173:338674. doi:10.1016/j.aca.2021.338674.

[67] Yassine M. Simple Functional Diagram of an LCMS System with Increased Clarity; Accessed on 20 July 2021. https://commons.wikimedia.org/wiki/File:Liquid_Chromatography_Mass_Spectrometer.png.

[68] Norena-Caro D. Liquid Chromatography MS Spectrum 3D Analysis; Accessed on 5 August 2021. https://en.wikipedia.org/w/index.php?title=File:Liquid_chromatography_MS_spectrum_3D_analysis.png&oldid=770313763.

[69] Smith R, Ventura D, Prince JT. LC-MS Alignment in Theory and Practice: A Comprehensive Algorithmic Review. Briefings in Bioinformatics. 2015;16(1):104–117. doi:10.1093/bib/bbt080.

[70] Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, et al. OpenMS – An Open-Source Software Framework for Mass Spectrometry. BMC Bioinformatics. 2008;9(1):163. doi:10.1186/1471-2105-9-163.

[71] Hoffmann N, Keck M, Neuweger H, Wilhelm M, Högy P, Niehaus K, et al. Combining Peak- and Chromatogram-Based Retention Time Alignment Algorithms for Multiple Chromatography-Mass Spectrometry Datasets. BMC Bioinformatics. 2012;13(1):214. doi:10.1186/1471-2105-13-214.

[72] Ballardini R, Benevento M, Arrigoni G, Pattini L, Roda A. MassUntangler: A Novel Alignment Tool for Label-Free Liquid Chromatography–Mass Spectrometry Proteomic Data. Journal of Chromatography A. 2011;1218(49):8859–8868. doi:10.1016/j.chroma.2011.06.062.

[73] Voss B, Hanselmann M, Renard BY, Lindner MS, Köthe U, Kirchner M, et al. SIMA: Simultaneous Multiple Alignment of LC/MS Peak Lists. Bioinformatics. 2011;27(7):987–993. doi:10.1093/bioinformatics/btr051.

[74] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000;28(1):27–30. doi:10.1093/nar/28.1.27.

[75] Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed Minimum Reporting Standards for Chemical Analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics: Official Journal of the Metabolomic Society. 2007;3(3):211–221. doi:10.1007/s11306-007-0082-2.

[76] Crisan D, Míguez J, Ríos-Muñoz G. On the Performance of Parallelisation Schemes for Particle Filtering. EURASIP Journal on Advances in Signal Processing. 2018;2018(1):31. doi:10.1186/s13634-018-0552-x.

[77] Wigren A, Murray L, Lindsten F. Improving the Particle Filter in High Dimensions Using Conjugate Artificial Process Noise. IFAC-PapersOnLine. 2018;51(15):670–675. doi:10.1016/j.ifacol.2018.09.207.

[78] van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. Scikit-Image: Image Processing in Python. PeerJ. 2014;2:e453. doi:10.7717/peerj.453.

[79] Pearson RD, Liu X, Sanguinetti G, Milo M, Lawrence ND, Rattray M. Puma: A Bioconductor Package for Propagating Uncertainty in Microarray Analysis. BMC Bioinformatics. 2009;10(1):211. doi:10.1186/1471-2105-10-211.

[80] Cox KL, Devanarayan V, Kriauciunas A, Manetta J, Montrose C, Sittampalam S. Immunoassay Methods. In: Markossian S, Grossman A, Brimacombe K, Arkin M, Auld D, Austin CP, et al., editors. Assay Guidance Manual. Bethesda (MD): Eli Lilly & Company and the National Center for Advancing Translational Sciences; 2004.

[81] Crowther JR. Stages in ELISA. In: Crowther JR, editor. The Elisa Guidebook. Methods in Molecular Biology™. Totowa, NJ: Humana Press; 2000. p. 45–82.

[82] Singh A, Upadhyay V, Upadhyay AK, Singh SM, Panda AK. Protein Recovery from Inclusion Bodies of Escherichia Coli Using Mild Solubilization Process. Microbial Cell Factories. 2015;14(1):41. doi:10.1186/s12934-015-0222-8.

[83] Lackie PM. Immunogold Silver Staining for Light Microscopy. Histochemistry and Cell Biology. 1996;106(1):9–17. doi:10.1007/BF02473198.

[84] Lin AV. Indirect ELISA. In: Hnasko R, editor. ELISA: Methods and Protocols. Methods in Molecular Biology. New York, NY: Springer; 2015. p. 51–59.

[85] Darain F, Wahab MA, Tjin SC. Surface Activation of Poly(Methyl Methacrylate) by Plasma Treatment: Stable Antibody Immobilization for Microfluidic Enzyme-Linked Immunosorbent Assay. Analytical Letters. 2012;45(17):2569–2579. doi:10.1080/00032719.2012.698673.

[86] Kulsharova G, Dimov N, Marques MPC, Szita N, Baganz F. Simplified Immobilisation Method for Histidine-Tagged Enzymes in Poly(Methyl Methacrylate) Microfluidic Devices. New Biotechnology. 2018;47:31–38. doi:10.1016/j.nbt.2017.12.004.

[87] Trinh KTL, Zhang H, Kang DJ, Kahng SH, Tall BD, Lee NY. Fabrication of Polymerase Chain Reaction Plastic Lab-on-a-Chip Device for Rapid Molecular Diagnoses. International Neurourology Journal. 2016;20(Suppl 1):S38–48. doi:10.5213/inj.1632602.301.

[88] Gibbs J, Vessels M, Rothenberg M. Immobilization Principles – Selecting the Surface for ELISA Assays. Corning Inc; 2001.

[89] Khnouf R, Karasneh D, Albiss BA. Protein Immobilization on the Surface of Polydimethylsiloxane and Polymethyl Methacrylate Microfluidic Devices. Electrophoresis. 2016;37(3):529–535. doi:10.1002/elps.201500333.

[90] Crowther JR. Titration of Reagents. In: Crowther JR, editor. The Elisa Guidebook. Methods in Molecular Biology™. Totowa, NJ: Humana Press; 2000. p. 83–113.

[91] Jackson AP, Sanders M, Berry A, McQuillan J, Aslett MA, Quail MA, et al. The Genome Sequence of Trypanosoma Brucei Gambiense, Causative Agent of Chronic Human African Trypanosomiasis. PLOS Neglected Tropical Diseases. 2010;4(4):e658. doi:10.1371/journal.pntd.0000658.

[92] Lindner AK, Lejon V, Chappuis F, Seixas J, Kazumba L, Barrett MP, et al. New WHO Guidelines for Treatment of Gambiense Human African Trypanosomiasis Including Fexinidazole: Substantial Changes for Clinical Practice. The Lancet Infectious Diseases. 2020;20(2):e38–e46. doi:10.1016/S1473-3099(19)30612-7.

[93] Tran T, Büscher P, Vandenbussche G, Wyns L, Messens J, De Greve H. Heterologous Expression, Purification and Characterisation of the Extracellular Domain of Trypanosome Invariant Surface Glycoprotein ISG75. Journal of Biotechnology. 2008;135(3):247–254. doi:10.1016/j.jbiotec.2008.04.012.

[94] Rogé S, Nieuwenhove LV, Meul M, Heykers A, de Koning AB, Bebronne N, et al. Recombinant Antigens Expressed in Pichia Pastoris for the Diagnosis of Sleeping Sickness Caused by Trypanosoma Brucei Gambiense. PLOS Neglected Tropical Diseases. 2014;8(7):e3006. doi:10.1371/journal.pntd.0003006.

[95] Chappuis F, Loutan L, Simarro P, Lejon V, Büscher P. Options for Field Diagnosis of Human African Trypanosomiasis. Clinical Microbiology Reviews. 2005;18(1):133–146. doi:10.1128/CMR.18.1.133-146.2005.

[96] Houghton RL, Stevens YY, Hjerrild K, Guderian J, Okamoto M, Kabir M, et al. Lateral Flow Immunoassay for Diagnosis of Trypanosoma Cruzi Infection with High Correlation to the Radioimmunoprecipitation Assay. Clinical and Vaccine Immunology. 2009;16(4):515–520. doi:10.1128/CVI.00383-08.

[97] Sullivan L, Fleming J, Sastry L, Mehlert A, Wall SJ, Ferguson MAJ. Identification of sVSG117 as an Immunodiagnostic Antigen and Evaluation of a Dual-Antigen Lateral Flow Test for the Diagnosis of Human African Trypanosomiasis. PLOS Neglected Tropical Diseases. 2014;8(7):e2976. doi:10.1371/journal.pntd.0002976.

[98] Van Nieuwenhove L, Rogé S, Lejon V, Guisez Y, Büscher P. Characterization of Trypanosoma Brucei Gambiense Variant Surface Glycoprotein LiTat 1.5. Genetics and molecular research: GMR. 2012;11(2):1260–1265. doi:10.4238/2012.May.9.5.

[99] Rooney B, Piening T, Büscher P, Rogé S, Smales CM. Expression of Trypanosoma Brucei Gambiense Antigens in Leishmania Tarentolae. Potential for Use in Rapid Serodiagnostic Tests (RDTs). PLOS Neglected Tropical Diseases. 2015;9(12):e0004271. doi:10.1371/journal.pntd.0004271.

[100] Sullivan L, Wall SJ, Carrington M, Ferguson MAJ. Proteomic Selection of Immunodiagnostic Antigens for Human African Trypanosomiasis and Generation of a Prototype Lateral Flow Immunodiagnostic Device. PLOS Neglected Tropical Diseases. 2013;7(2):e2087. doi:10.1371/journal.pntd.0002087.

[101] Ziegelbauer K, Multhaup G, Overath P. Molecular Characterization of Two Invariant Surface Glycoproteins Specific for the Bloodstream Stage of Trypanosoma Brucei. The Journal of Biological Chemistry. 1992;267(15):10797–10803.

[102] Ziegelbauer K, Rudenko G, Kieft R, Overath P. Genomic Organization of an Invariant Surface Glycoprotein Gene Family of Trypanosoma Brucei. Molecular and Biochemical Parasitology. 1995;69(1):53–63. doi:10.1016/0166-6851(94)00194-r.

[103] Vera A, González-Montalbán N, Arís A, Villaverde A. The Conformational Quality of Insoluble Recombinant Proteins Is Enhanced at Low Growth Temperatures. Biotechnology and Bioengineering. 2007;96(6):1101–1106. doi:10.1002/bit.21218.

[104] Schein CH, Noteborn MHM. Formation of Soluble Recombinant Proteins in Escherichia Coli Is Favored by Lower Growth Temperature. Bio/Technology. 1988;6(3):291–294. doi:10.1038/nbt0388-291.

[105] Gräslund S, Nordlund P, et al. Protein Production and Purification. Nature Methods. 2008;5(2):135–146. doi:10.1038/nmeth.f.202.

[106] Franco JR, Simarro PP, Diarra A, Ruiz-Postigo JA, Jannin JG. The Human African Trypanosomiasis Specimen Biobank: A Necessary Tool to Support Research of New Diagnostics. PLOS Neglected Tropical Diseases. 2012;6(6):e1571. doi:10.1371/journal.pntd.0001571.

[107] Lejon V, Rebeski DE, Ndao M, Baelmans R, Winger EM, Faye D, et al. Performance of Enzyme-Linked Immunosorbent Assays for Detection of Antibodies against T. Congolense and T. Vivax in Goats. Veterinary Parasitology. 2003;116(2):87–95. doi:10.1016/S0304-4017(03)00257-7.

[108] Miranda A, Martínez L, De Beule PAA. Facile Synthesis of an Aminopropylsilane Layer on Si/SiO2 Substrates Using Ethanol as APTES Solvent. MethodsX. 2020;7:100931. doi:10.1016/j.mex.2020.100931.

[109] Jenko KL, Zhang Y, Kostenko Y, Fan Y, Garcia-Rodriguez C, Lou J, et al. Development of an ELISA Microarray Assay for the Sensitive and Simultaneous Detection of Ten Biodefense Toxins. Analyst. 2014;139(20):5093–5102. doi:10.1039/C4AN01270D.

[110] Joh DY, Hucknall AM, Wei Q, Mason KA, Lund ML, Fontes CM, et al. Inkjet-Printed Point-of-Care Immunoassay on a Nanoscale Polymer Brush Enables Subpicomolar Detection of Analytes in Blood. Proceedings of the National Academy of Sciences. 2017;114(34):E7054–E7062. doi:10.1073/pnas.1703200114.

[111] Gonzalez-Macia L, Morrin A, Smyth MR, Killard AJ. Advanced Printing and Deposition Methodologies for the Fabrication of Biosensors and Biodevices. Analyst. 2010;135(5):845–867. doi:10.1039/B916888E.

[112] Singh MI, Jain V. Tagging the Expressed Protein with 6 Histidines: Rapid Cloning of an Amplicon with Three Options. PLoS ONE. 2013;8(5). doi:10.1371/journal.pone.0063922.

[113] Erickson HP. Size and Shape of Protein Molecules at the Nanometer Level Determined by Sedimentation, Gel Filtration, and Electron Microscopy. Biological Procedures Online. 2009;11:32–51. doi:10.1007/s12575-009-9008-x.

[114] Fischer H, Polikarpov I, Craievich AF. Average Protein Density Is a Molecular-Weight-Dependent Function. Protein Science : A Publication of the Protein Society. 2004;13(10):2825–2828. doi:10.1110/ps.04688204.

[115] Hosseini S, Ibrahim F, Djordjevic I, Koole LH. Recent Advances in Surface Functionalization Techniques on Polymethacrylate Materials for Optical Biosensor Applications. The Analyst. 2014;139(12):2933–2943. doi:10.1039/c3an01789c.

[116] Costantini F, Sberna C, Petrucci G, Manetti C, de Cesare G, Nascetti A, et al. Lab-on-Chip System Combining a Microfluidic-ELISA with an Array of Amorphous Silicon Photosensors for the Detection of Celiac Disease Epitopes. Sensing and Bio-Sensing Research. 2015;6:51–58. doi:10.1016/j.sbsr.2015.11.003.

[117] Ohtake S, Kita Y, Arakawa T. Interactions of Formulation Excipients with Proteins in Solution and in the Dried State. Advanced Drug Delivery Reviews. 2011;63(13):1053–1073. doi:10.1016/j.addr.2011.06.011.

[118] Rebeski DE, Winger EM, Robinson MM, Gabler CM, Dwinger RH, Crowther JR. Evaluation of Antigen-Coating Procedures of Enzyme-Linked Immunosorbent Assay Method for Detection of Trypanosomal Antibodies. Veterinary Parasitology. 2000;90(1-2):1–13. doi:10.1016/s0304-4017(00)00231-4.

[119] Morang'a C, Ayieko C, Awinda G, Achilla R, Moseti C, Ogutu B, et al. Stabilization of RDT Target Antigens Present in Dried Plasmodium Falciparum-Infected Samples for Validating Malaria Rapid Diagnostic Tests at the Point of Care. Malaria Journal. 2018;17(1):10. doi:10.1186/s12936-017-2155-7.

[120] Nielsen K. Use of Dried Smooth Lipopolysaccharide Antigen Coated Polystyrene Plates for Diagnosis of Bovine Brucellosis by Enzyme Immunoassay. Journal of Immunoassay. 1998;19(1):39–48. doi:10.1080/01971529808005470.

[121] Filipponi L, Livingston P, Kašpar O, Tokárová V, Nicolau DV. Protein Patterning by Microcontact Printing Using Pyramidal PDMS Stamps. Biomedical Microdevices. 2016;18:9. doi:10.1007/s10544-016-0036-4.

[122] Cui L, Pang J, Lee YH, Ooi EE, Ong CN, Leo YS, et al. Serum Metabolome Changes in Adult Patients with Severe Dengue in the Critical and Recovery Phases of Dengue Infection. PLOS Neglected Tropical Diseases. 2018;12(1):e0006217. doi:10.1371/journal.pntd.0006217.

[123] Gale TV, Schieffelin JS, Branco LM, Garry RF, Grant DS. Elevated L-Threonine Is a Biomarker for Lassa Fever and Ebola. Virology Journal. 2020;17(1):188. doi:10.1186/s12985-020-01459-y.

[124] Na J, Khan A, Kim JK, Wadood A, Choe YL, Walker DI, et al. Discovery of Metabolic Alterations in the Serum of Patients Infected with Plasmodium Spp. by High-Resolution Metabolomics. Metabolomics. 2019;16(1):9. doi:10.1007/s11306-019-1630-2.

[125] Pegalajar-Jurado A, Fitzgerald BL, Islam MN, Belisle JT, Wormser GP, Waller KS, et al. Identification of Urine Metabolites as Biomarkers of Early Lyme Disease. Scientific Reports. 2018;8(1):12204. doi:10.1038/s41598-018-29713-y.

[126] Vincent IM, Daly R, Courtioux B, Cattanach AM, Biéler S, Ndung'u JM, et al. Metabolomics Identifies Multiple Candidate Biomarkers to Diagnose and Stage Human African Trypanosomiasis. PLOS Neglected Tropical Diseases. 2016;10(12):e0005140. doi:10.1371/journal.pntd.0005140.

[127] Watrous JD, Henglin M, Claggett B, Lehmann KA, Larson MG, Cheng S, et al. Visualization, Quantification, and Alignment of Spectral Drift in Population Scale Untargeted Metabolomics Data. Analytical Chemistry. 2017;89(3):1399–1404. doi:10.1021/acs.analchem.6b04337.

[128] Attygalle AB, Pavlov J. Nominal Mass? Journal of The American Society for Mass Spectrometry. 2017;28(8):1737–1738. doi:10.1007/s13361-017-1647-6.

[129] Dudzik D, Barbas-Bernardos C, García A, Barbas C. Quality Assurance Procedures for Mass Spectrometry Untargeted Metabolomics. a Review. Journal of Pharmaceutical and Biomedical Analysis. 2018;147:149–173. doi:10.1016/j.jpba.2017.07.044.

[130] Broadhurst D, Goodacre R, Reinke SN, Kuligowski J, Wilson ID, Lewis MR, et al. Guidelines and Considerations for the Use of System Suitability and Quality Control Samples in Mass Spectrometry Assays Applied in Untargeted Clinical Metabolomic Studies. Metabolomics. 2018;14(6):72. doi:10.1007/s11306-018-1367-3.

[131] Jarrin EP, Cordeiro FB, Medranda WC, Barrett M, Zambrano M, Regato M. A Machine Learning-Based Algorithm for the Assessment of Clinical Metabolomic Fingerprints in Zika Virus Disease. In: 2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI); 2019. p. 1–6.

[132] Botana L, Matía B, San Martin JV, Romero-Maté A, Castro A, Molina L, et al. Cellular Markers of Active Disease and Cure in Different Forms of Leishmania Infantum-Induced Disease. Frontiers in Cellular and Infection Microbiology. 2018;8. doi:10.3389/fcimb.2018.00381.

[133] Milne K, Ivens A, Reid AJ, Lotkowska ME, O'Toole A, Sankaranarayanan G, et al. Mapping Immune Variation and Var Gene Switching in Naive Hosts Infected with Plasmodium Falciparum. eLife. 2021;10:e62800. doi:10.7554/eLife.62800.

[134] Lee JY, Styczynski MP. NS-kNN: A Modified k-Nearest Neighbors Approach for Imputing Metabolomics Data. Metabolomics : Official journal of the Metabolomic Society. 2018;14(12):153. doi:10.1007/s11306-018-1451-8.

[135] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies. Nucleic Acids Research. 2015;43(7):e47–e47. doi:10.1093/nar/gkv007.

[136] Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, et al. Predicting Network Activity from High Throughput Metabolomics. PLOS Computational Biology. 2013;9(7):e1003123. doi:10.1371/journal.pcbi.1003123.

[137] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2003;13(11):2498–2504. doi:10.1101/gr.1239303.

[138] Rogers S. Mass-Spec-Utils: Some Useful MS Code; 2020.

[139] Bittremieux W, Chen C, Dorrestein PC, Schymanski EL, Schulze T, Neumann S, et al. Universal MS/MS Visualization and Retrieval with the Metabolomics Spectrum Resolver Web Service. bioRxiv. 2020; p. 2020.05.09.086066. doi:10.1101/2020.05.09.086066.

[140] Wu CT, Wang Y, Wang Y, Ebbels T, Karaman I, Graça G, et al. Targeted Realignment of LC-MS Profiles by Neighbor-Wise Compound-Specific Graphical Time Warping with Misalignment Detection. Bioinformatics. 2020;36(9):2862–2871. doi:10.1093/bioinformatics/btaa037.

[141] Li L, Ren W, Kong H, Zhao C, Zhao X, Lin X, et al. An Alignment Algorithm for LC-MS-based Metabolomics Dataset Assisted by MS/MS Information. Analytica Chimica Acta. 2017;990:96–102. doi:10.1016/j.aca.2017.07.058.

[142] Habra H, Kachman M, Bullock K, Clish C, Evans CR, Karnovsky A. metabCombiner: Paired Untargeted LC-HRMS Metabolomics Feature Matching and Concatenation of Disparately Acquired Data Sets. Analytical Chemistry. 2021;doi:10.1021/acs.analchem.0c03693.

[143] Domingo-Almenara X, Montenegro-Burke JR, Benton HP, Siuzdak G. Annotation: A Computational Solution for Streamlining Metabolomics Analysis. Analytical Chemistry. 2018;90(1):480–489. doi:10.1021/acs.analchem.7b03929.

[144] Viant MR, Kurland IJ, Jones MR, Dunn WB. How Close Are We to Complete Annotation of Metabolomes? Current Opinion in Chemical Biology. 2017;36:64–69. doi:10.1016/j.cbpa.2017.01.001.

[145] Mehraj V, Routy JP. Tryptophan Catabolism in Chronic Viral Infections: Handling Uninvited Guests. International Journal of Tryptophan Research : IJTR. 2015;8:41–48. doi:10.4137/IJTR.S26862.

[146] Badawy AAB, Guillemin G. The Plasma [Kynurenine]/[Tryptophan] Ratio and Indoleamine 2,3-Dioxygenase: Time for Appraisal. International Journal of Tryptophan Research : IJTR. 2019;12:1178646919868978. doi:10.1177/1178646919868978.

[147] Takikawa O, Kuroiwa T, Yamazaki F, Kido R. Mechanism of Interferon-Gamma Action. Characterization of Indoleamine 2,3-Dioxygenase in Cultured Human Cells Induced by Interferon-Gamma and Evaluation of the Enzyme-Mediated Tryptophan Degradation

in Its Anticellular Activity. Journal of Biological Chemistry. 1988;263(4):2041–2048. doi:10.1016/S0021-9258(19)77982-4.

[148] Cesario A, Rocca B, Rutella S. The Interplay between Indoleamine 2,3-Dioxygenase 1 (IDO1) and Cyclooxygenase (COX)-2 In Chronic Inflammation and Cancer. Current Medicinal Chemistry. 2011;18(15):2263–2271. doi:10.2174/092986711795656063.

[149] Mukherjee P, Basu GD, Tinder TL, Subramani DB, Bradley JM, Arefayene M, et al. Progression of Pancreatic Adenocarcinoma Is Significantly Impeded with a Combination of Vaccine and COX-2 Inhibition. The Journal of Immunology. 2009;182(1):216–224. doi:10.4049/jimmunol.182.1.216.

[150] Clark CJ, Mackay GM, Smythe GA, Bustamante S, Stone TW, Phillips RS. Prolonged Survival of a Murine Model of Cerebral Malaria by Kynurenine Pathway Inhibition. Infection and Immunity. 2005;73(8):5249–5251. doi:10.1128/IAI.73.8.5249-5251.2005.

[151] Rodgers J, Stone TW, Barrett MP, Bradley B, Kennedy PGE. Kynurenine Pathway Inhibition Reduces Central Nervous System Inflammation in a Model of Human African Trypanosomiasis. Brain: A Journal of Neurology. 2009;132(Pt 5):1259–1267. doi:10.1093/brain/awp074.

[152] Marim FM, Teixeira DC, Queiroz-Junior CM, Valiate BVS, Alves-Filho JC, Cunha TM, et al. Inhibition of Tryptophan Catabolism Is Associated With Neuroprotection During Zika Virus Infection. Frontiers in Immunology. 2021;12:2843. doi:10.3389/fimmu.2021.702048.

[153] Gangneux JP, Poinsignon Y, Donaghy L, Amiot L, Tarte K, Mary C, et al. Indoleamine 2,3-Dioxygenase Activity as a Potential Biomarker of Immune Suppression during Visceral Leishmaniasis. Innate Immunity. 2013;19(6):564–568. doi:10.1177/1753425912473170.

[154] Campbell B, Charych E, Lee A, Möller T. Kynurenines in CNS Disease: Regulation by Inflammatory Cytokines. Frontiers in Neuroscience. 2014;8:12. doi:10.3389/fnins.2014.00012.

[155] Cruzat V, Macedo Rogero M, Noel Keane K, Curi R, Newsholme P. Glutamine: Metabolism and Immune Function, Supplementation and Clinical Translation. Nutrients. 2018;10(11). doi:10.3390/nu10111564.

[156] Wijnands KAP, Castermans TMR, Hommen MPJ, Meesters DM, Poeze M. Arginine and Citrulline and the Immune Response in Sepsis. Nutrients. 2015;7(3):1426–1463. doi:10.3390/nu7031426.

[157] Jamaati H, Mortaz E, Pajouhi Z, Folkerts G, Movassaghi M, Moloudizargari M, et al. Nitric Oxide in the Pathogenesis and Treatment of Tuberculosis. Frontiers in Microbiology. 2017;8:2008. doi:10.3389/fmicb.2017.02008.

[158] Burgner D, Rockett K, Kwiatkowski D. Nitric Oxide and Infectious Diseases. Archives of Disease in Childhood. 1999;81(2):185–188. doi:10.1136/adc.81.2.185.

[159] Wrotek S, Sobocińska J, Kozłowski HM, Pawlikowska M, Jędrzejewski T, Dzialuk A. New Insights into the Role of Glutathione in the Mechanism of Fever. International Journal of Molecular Sciences. 2020;21(4):1393. doi:10.3390/ijms21041393.

[160] Collier B, Dossett LA, May AK, Diaz JJ. Glucose Control and the Inflammatory Response. Nutrition in Clinical Practice. 2008;23(1):3–15. doi:10.1177/011542650802300103.

[161] Li M, van Esch BCAM, Henricks PAJ, Folkerts G, Garssen J. The Anti-inflammatory Effects of Short Chain Fatty Acids on Lipopolysaccharide- or Tumor Necrosis Factor α-Stimulated Endothelial Cells via Activation of GPR41/43 and Inhibition of HDACs. Frontiers in Pharmacology. 2018;0. doi:10.3389/fphar.2018.00533.

[162] Hannun YA, Obeid LM. Sphingolipids and Their Metabolism in Physiology and Disease. Nature Reviews Molecular Cell Biology. 2018;19(3):175–191. doi:10.1038/nrm.2017.107.

[163] Sanchez-Lopez E, Zhong Z, Stubelius A, Sweeney SR, Booshehri LM, Antonucci L, et al. Choline Uptake and Metabolism Modulate Macrophage IL-1β and IL-18 Production. Cell Metabolism. 2019;29(6):1350–1362.e7. doi:10.1016/j.cmet.2019.03.011.

[164] Wu Y, Liu Y, Gulbins E, Grassmé H. The Anti-Infectious Role of Sphingosine in Microbial Diseases. Cells. 2021;10(5):1105. doi:10.3390/cells10051105.

[165] Arish M, Husein A, Kashif M, Saleem M, Akhter Y, Rub A. Sphingosine-1-Phosphate Signaling: Unraveling Its Role as a Drug Target against Infectious Diseases. Drug Discovery Today. 2016;21(1):133–142. doi:10.1016/j.drudis.2015.09.013.

[166] Dhangadamajhi G, Singh S. Sphingosine 1-Phosphate in Malaria Pathogenesis and Its Implication in Therapeutic Opportunities. Frontiers in Cellular and Infection Microbiology. 2020;0. doi:10.3389/fcimb.2020.00353.

[167] Igarashi J, Bernier SG, Michel T. Sphingosine 1-Phosphate and Activation of Endothelial Nitric-oxide Synthase: Differential Regulation of Akt and MAP Kinase Pathways by EDG and Bradykinin Receptors in Vascular Endothelial Cells. Journal of Biological Chemistry. 2001;276(15):12420–12426. doi:10.1074/jbc.M008375200.

[168] Wang Y, Wang W, Xu L, Zhou X, Shokrollahi E, Felczak K, et al. Cross Talk between Nucleotide Synthesis Pathways with Cellular Immunity in Constraining Hepatitis E Virus Replication. Antimicrobial Agents and Chemotherapy. 2016;60(5):2834–2848. doi:10.1128/AAC.02700-15.

[169] Kozak W, Kozak A. Selected Contribution: Differential Role of Nitric Oxide Synthase Isoforms in Fever of Different Etiologies: Studies Using Nosgene-deficient Mice. Journal of Applied Physiology. 2003;94(6):2534–2544. doi:10.1152/japplphysiol.01042.2002.

[170] Yanez M, Jhanji M, Murphy K, Gower RM, Sajish M, Jabbarzadeh E. Nicotinamide Augments the Anti-Inflammatory Properties of Resveratrol through PARP1 Activation. Scientific Reports. 2019;9(1):10219. doi:10.1038/s41598-019-46678-8.

[171] Shen B, Yi X, Sun Y, Bi X, Du J, Zhang C, et al. Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. Cell. 2020;doi:10.1016/j.cell.2020.05.032.

[172] Thomas T, Stefanoni D, Reisz JA, Nemkov T, Bertolone L, Francis RO, et al. COVID-19 Infection Alters Kynurenine and Fatty Acid Metabolism, Correlating with IL-6 Levels and Renal Status. JCI Insight. 2020;5(14). doi:10.1172/jci.insight.140327.

[173] Colvin HN, Cordy RJ. Insights into Malaria Pathogenesis Gained from Host Metabolomics. PLOS Pathogens. 2020;16(11):e1008930. doi:10.1371/journal.ppat.1008930.

[174] Lipton JM, Ticknor CB. Central Effect of Taurine and Its Analogues on Fever Caused by Intravenous Leukocytic Pyrogen in the Rabbit. The Journal of Physiology. 1979;287:535–543.

[175] Enwonwu CO, Afolabi BM, Salako LO, Idigbe EO, Bashirelahi N. Increased Plasma Levels of Histidine and Histamine in Falciparum Malaria: Relevance to Severity of Infection. Journal of Neural Transmission. 2000;107(11):1273–1287. doi:10.1007/s007020070017.

[176] Daddona PE, Wiesmann WP, Lambros C, Kelley WN, Webster HK. Human Malaria Parasite Adenosine Deaminase. Characterization in Host Enzyme-Deficient Erythrocyte Culture. Journal of Biological Chemistry. 1984;259(3):1472–1475. doi:10.1016/S0021-9258(17)43431-4.

[177] Onabanjo AO, Maegraith BG. The Probable Pathogenic Role of Adenosine in Malaria. British Journal of Experimental Pathology. 1970;51(6):581–586.

[178] Leopold SJ, Ghose A, Allman EL, Kingston HWF, Hossain A, Dutta AK, et al. Identifying the Components of Acidosis in Patients With Severe Plasmodium Falciparum Malaria Using Metabolomics. The Journal of Infectious Diseases. 2019;219(11):1766–1776. doi:10.1093/infdis/jiy727.

[179] Mukherjee D, Chora ÂF, Mota MM. Microbiota, a Third Player in the Host–Plasmodium Affair. Trends in Parasitology. 2020;36(1):11–18. doi:10.1016/j.pt.2019.11.001.

[180] Oliveira MJC, Silva Júnior GB, Abreu KLS, Rocha NA, Garcia AVV, Franco LFLG, et al. Risk Factors for Acute Kidney Injury in Visceral Leishmaniasis (Kala-Azar). The American Journal of Tropical Medicine and Hygiene. 2010;82(3):449–453. doi:10.4269/ajtmh.2010.09-0571.

[181] Annese VF, Patil SB, Hu C, Giagkoulovits C, Al-Rawhani MA, Grant J, et al. A Monolithic Single-Chip Point-of-Care Platform for Metabolomic Prostate Cancer Detection. Microsystems & Nanoengineering. 2021;7(1):1–15. doi:10.1038/s41378-021-00243-4.

[182] Schmid R, Petras D, Nothias LF, Wang M, Aron AT, Jagels A, et al. Ion Identity Molecular Networking for Mass Spectrometry-Based Metabolomics in the GNPS Environment. Nature Communications. 2021;12(1):3832. doi:10.1038/s41467-021-23953-9.

[183] Huang N, Siegel MM, Kruppa GH, Laukien FH. Automation of a Fourier Transform Ion Cyclotron Resonance Mass Spectrometer for Acquisition, Analysis, and e-Mailing of High-Resolution Exact-Mass Electrospray Ionization Mass Spectral Data. Journal of the American Society for Mass Spectrometry. 1999;10(11):1166–1173. doi:10.1016/S1044-0305(99)00089-6.