Understanding viral infections using mathematical models and statistical analysis
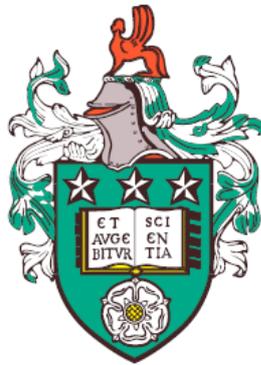
Macauley Locke

Department of Applied Mathematics

University of Leeds

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

August, 2023

Joint publications

Almost all of the work in Chapter 3.2 has been refereed and published, as follows:

- Lin, C, Wolf, J, Brice, D, Sun Y, **Locke, M**, Cherry S, Castellaw, A Wehenkel, M, Crawford, J, Zarnitsyna, V, Luque-Duque, D, Allison, K, Allen, E, Brown, S, Mandarano, A, Estepp, J, The SJTRC Study Team, Taylor, C, Molina-París, C, Schultz-Cherry, S, Tang, L, Thomas, P, McGargill, M, (2022). Pre-existing humoral immunity to human common cold coronaviruses negatively impacts the protective SARS-CoV-2 antibody response Perspective. *Cell Host & Microbe*, 8(6), p.1014.

Additionally, almost all of the work in Chapter 4 has been refereed and published, as follows:

- **Locke, M**, Lythe, G, López-García, M, Muñoz-Fontela, C, Carroll, M, and Molina-París, C, (2021). Quantification of type I interferon inhibition by viral proteins: Ebola virus as a case study. *MDPI Viruses*.

Finally, some of the work in Chapter 5.1 is in preparation for submission:

- **Locke, M***, Grebennikov, D*, Sazonov, I, Loguinova, M, Meyerhans, A, Bocharov, G, Molina-París, C. Exploring the therapeutic potential of defective interfering particles in reducing the replication of SARS-CoV-2. *PLOS computational, in preparation*, (*: First co-authorship).

Acknowledgements

Preface

This research was funded by the Engineering and Physical Science Research Council Doctoral Training Grant Studentship, provided jointly by the University of Leeds and Public Health England, now the UK Health Security Agency. The original plan for my PhD project was to work with Public Health England and other experts in Ebola virus epidemiology and immunology to develop mathematical models on Ebola virus infections. The aim was to support the statistical analysis of data sets in Chapter 3 of this Thesis and to use additional data collected from experiments to formulate and calibrate several models of Ebola infections.

However, as soon as the SARS-CoV-2 pandemic began, many of the collaborators at Public Health England and elsewhere shifted their attention to respond. As a result, not only was there a delay in receiving the completed data set for statistical analysis, but many of the experiments to obtain additional data were abandoned to focus efforts on the pandemic, especially in the case of Public Health England, a government agency.

To ensure there was enough content for this thesis and to stay within viral immunology, new collaborations were formed that gave rise to the additional work in this thesis. The first new collaboration was with St Jude Children's Research Hospital to investigate antibody responses in SARS-CoV-2 infection. This new collaboration investigated potential pre-existing immune responses to SARS-CoV-2 due to the circulation of common coronaviruses that share sequence homology with SARS-CoV-2. Since I had already begun statistical analysis for Ebola patient

data, my supervisors and I thought this work investigating another virus would fit within the scope of this Thesis.

The next collaboration was formed with Dmitry Grebennikov and Gennady Bocharov to investigate defective interfering particles as a potential therapeutic during SARS-CoV-2 infection, using deterministic and stochastic modelling. Due to the pandemic, it seemed relevant to explore SARS-CoV-2 and develop a model that would characterise this as a potential treatment option for the future. Finally, a collaboration with Eva Perez from the Pirbright Institute was formed to investigate Foot and Mouth disease. The analysis that was carried out to investigate Foot and Mouth disease formed part of a research proposal and allowed me to investigate and expand my knowledge into viruses that are predominant in other species. Additionally, this allowed me to gain a further understanding of the requirements for research funding that is necessary in a research career.

'I believe these collaborations allowed me to broaden my knowledge about virus immunology and develop additional collaborations and tools for my future career. The previous information also explains why my thesis is not solely on Ebola infection but also on SARS-CoV-2 and Foot and Mouth disease.

Abstract

There are a range of viruses that exist, that not only infect humans, but also a wide range of other species. These viruses not only present a threat to human health but also to economies, in particular, less economically developed countries. Mathematical models and statistical analysis provide ways to understand viral dynamics and immune responses better and test new hypotheses using information gathered through biological experimentation. In Chapter 3, I will analyse two cohort studies from viral outbreaks. The first study is on data from the 2014 West Africa Ebola outbreak collected by Public Health England (now known as the United Kingdom Health Security Agency) to understand the longitudinal antibody and T-cell response of survivors from the epidemic. St Jude Children's Research Hospital undertook the second study to understand possible associations between common coronaviruses and SARS-CoV-2, using data collected within the hospital.

Chapter 4 introduces three potential stochastic models to investigate type I interferon (IFN) antagonism, a tactic employed by several viruses, including SARS-CoV-2, Ebola and Crimean Congo Haemorrhagic fever to avoid immune responses. Here Ebola virus is investigated as a case study. Finally, in Chapter 5, I examine two mathematical models for viral infections. The first model explores defective interfering particles as a potential therapeutic in SARS-CoV-2 infections. The second model investigates foot and mouth disease infections *in vitro* to examine differences between three strains, including determining each strain's time to infection and basic reproduction number.

Contents

CONTENTS

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Chapter 1

Introduction

1.1 Biological introduction

During our lifetime, we are exposed to a variety of viruses. For healthy individuals, many of these pose little to no risk, and exposure often leads to increased resistance compared to before infection, or immunity from a repeat infection (Chaplin, 2010). However, there exist many viruses that not only have increased infectivity but are also highly pathogenic, leading to increased mortalities amongst the infected. The World Health Organisation (WHO) has listed several viruses that are a cause for concern, either due to lack of efficient treatments or their potential to cause an epidemic (WHO, 2015). Several of these have seen large outbreaks in recent years and include SARS-CoV-2 (Wu *et al.*, 2020), Ebola virus (EBOV) and Marburg virus (Brolin Ribacke *et al.*, 2016), Zika virus (Campos *et al.*, 2015) and Crimean Congo haemorrhagic fever virus (Appannanavar & Mishra, 2011).

For any virus to be successful and infect an organism, it must possess several key attributes. First, it must be transmissible to infect new susceptible hosts to persist amongst the population. Viruses use several routes to enter a host, including respiratory, gastrointestinal or through genital tracts (Louten, 2016b). Examples of viruses that use these routes include Influenza A and B viruses, Poliovirus, and Human Papillomaviruses (HPV), respectively. Furthermore, viruses can be spread through skin contact, in the case of herpes simplex virus-1 (HSV-1), or through penetration into the dermis or subcutaneous tissue. This can occur by injection or cuts, such as for Hepatitis A and Human Immunodeficiency virus

1. INTRODUCTION

(HIV), or through other vectors, such as ticks and mosquitoes (Piesman & Eisen, 2008; Tolle, 2009). Viruses can also be transmitted through contaminated organs or blood products, as has been observed for several viruses and has even been the cause of several infections in the 1980s (Weinberg *et al.*, 2002).

Next, any virus that enters a host must be able to access a target site for replication and replicate within a particular cell, referred to as viral tropism (Louten, 2016a). Viral tropism can be receptor-dependent, requiring the correct cell surface receptor, or receptor-independent, where intracellular molecules such as cytokines determine the tropism (McFadden *et al.*, 2009). The general steps for viral replication are attachment through a cell surface receptor, penetration into the cell, un-coating of viral genomes, replication and assembly, whether that includes integrating the host's replication machinery or additionally providing its own, maturation into an infective virion and finally, release (Ahlquist *et al.*, 2003). The viral release can be achieved either by budding, where the virions are released gradually from the cell, or through the cell bursting and releasing all the virions simultaneously. Any virus will most likely go extinct without the ability to replicate successfully within a host.

Finally, to persist in a host, any virus must be able to subvert the host's immune responses (Alcami & Koszinowski, 2000). Host immune responses are broken into two distinct arms, the innate and adaptive immune responses. Innate immune responses are non-specific and target a few highly conserved structures called pathogen-associated molecular patterns (PAMPs). These can be for example blunt or 5'-triphosphorylated ends of viral genomic ribonucleic acid (RNA) segments or long double stranded RNA (Berke *et al.*, 2013). An example of a PAMP is viral ribonucleic acids (RNA) (Medzhitov & Janeway Jr, 2000). Innate responses occur within the first few days of viral infection. Activation of innate immune responses results in the expression of various genes, which produces many different inflammatory cytokines, to initiate an infected cell state and trigger neighbouring cells' immune responses (Hoebe *et al.*, 2004). The latter part of the immune response is the adaptive arm, which takes place several days to weeks after infection and is virus-specific. Antigen-presenting cells, such as macrophages and dendritic cells (DCs), present small sections of the viral genome, referred to as epitopes, through the major histocompatibility complex (MHC). MHC class I and

II present epitopes to T-cells that allow a T-cell to recognise cells infected with a pathogen (Swain, 1983). T-cells then lead a cell-mediated response to actively kill cells infected with the virus by recognising the epitope presented by MHC on the cell surface. Meanwhile, B-cells mount an adequate antibody response to neutralise the virus within the host (Bonilla & Oettgen, 2010). While these two routes are distinct, they depend on one another to form a coordinated and effective immune response. Since viruses are obligate parasites and require the intra-cellular machinery to replicate, many viruses encode proteins that antagonise or suppress an immune response (Zinzula & Tramontano, 2013) to allow replication with minimal interference.

SARS-CoV-2 and Ebola are the most recognisable viruses in the last decade. In 2014 the West Africa EBOV outbreak led to a suspected 28,652 cases and a reported 11,326 deaths (Bell, 2016). The 2019 SARS-CoV-2 pandemic, as of 8th November 2023, has caused an over 777 million infections, and over 6.9 million deaths have been recorded worldwide (World Health Organization, November 2023). These two viruses still threaten human health and will be the main focus of the following work; an overview of each virus is provided.

1.2   SARS-CoV-2

In late 2019 an unknown agent was reported to be causing pneumonia in clusters of patients in Wuhan, China (Zhu *et al.*, 2020a). This would later be identified as a Coronavirus (family *Coronaviridae*) and subsequently named SARS-CoV-2. The WHO declared SARS-CoV-2 a "public health emergency of international concern" on 30th January 2020 (Li *et al.*, 2020a), rapidly spreading to 113 countries. By the 17th March 2020, it had caused a reported 179,112 infections and 7,426 deaths (World Health Organization *et al.*, 2020). The rapid spread of SARS-CoV-2 would see many governments implement many different lockdowns and social distancing measures to help combat the spread of infections (Tobías, 2020). However, even with these measures, new infections occurred rapidly, which meant many healthcare systems were pushed to breaking point, not only affecting the ability of them to treat infected patients effectively but also at the detriment of healthcare workers' physical and mental well-being (Sanghera *et al.*, 2020). The

1. INTRODUCTION
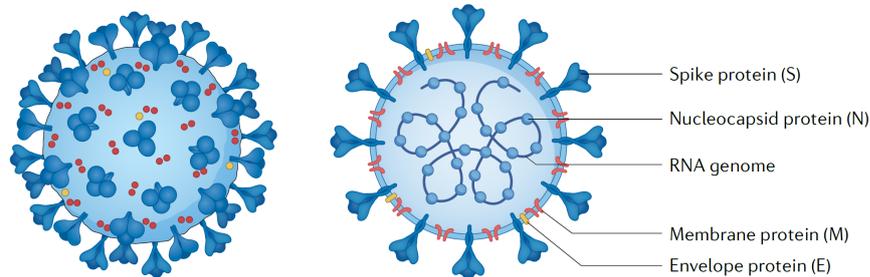
Figure 1.1: Structure of SARS-CoV-2 with spike protein (S), nucleocapsid protein (N), membrane (M) and envelope protein (E). Figure taken from Lamers & Haagmans (2022). Copyright License Number: 5596600336543.

effects of the global pandemic have also impacted the global economy, especially in those countries that rely heavily on tourism to support their economy (Pinilla *et al.*, 2021). Thankfully, vaccine development became a priority, and several effective vaccines now reduce the transmission and pathogenicity of the virus (Creech *et al.*, 2021). However, the emergence of new viral strains is still a cause for concern; therefore, research into the virus continues.

Coronaviruses are divided into four genera: α-CoVs, β-CoVs, γ-CoVs and δ-CoVS, of which only α and β are known to infect only mammals, γ infects avian species and δ infects both (Naqvi *et al.*, 2020). There are four coronaviruses endemic in human populations these being hCCoV-NL63 (α), hCCoV-229E (α), hCCoV-HKU1 (β), and hCCoV-OC43 (β), which infect the upper respiratory tract causing common-cold symptoms (Liu *et al.*, 2021). Two other coronaviruses are severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV), whose genome shares a 79% (Zhang & Holmes, 2020) and 50% (Abdelrahman *et al.*, 2020) sequence similarity with SARS-CoV-2 wild type respectively. Fatality rates of SARS and MERS are $\approx$10%, and $\approx$35% respectively (Li *et al.*, 2020b). SARS-CoV-2 however, has fatality rates that vary from 0.1%-18.1% (John Hopkins University of Medicine , March 2023) depending on the country considered. In addition, between January 2020, and December 2021, it is estimated there are 14.83 million excess deaths as a consequence of the SARS-CoV-2 pandemic (Msemburi *et al.*, 2023). This is significantly higher than the 5.42 million cases reported to the

WHO. Therefore mortality rates including excess deaths may be much higher. The seven coronaviruses are all known to be zoonotic viruses, with five coming from bat reservoirs (Singh *et al.*, 2021).

SARS-CoV-2 is a novel beta coronavirus, whose ≈30 kb single-stranded positive-sense RNA molecule genome encodes for six functional proteins: replicase (ORF1a/1b), spike protein (S), envelope (E), membrane (M) and nucleocapsid (N), with seven non-structural proteins interlaced amongst the others (Wang *et al.*, 2020). The structure of SARS-CoV-2 is illustrated in Figure 1.1 (Lamers & Haagmans, 2022). Many of the current vaccines include antigens designed to target the spike protein of the virus, while a select few include whole inactivated virus (Creech *et al.*, 2021). Mutants typically generated during the replication process are a cause of concern regarding vaccine development since most vaccines target the S protein. During the replication cycle, some of the processes that can lead to the generation of new mutants are reading frame shifts, transcription or translational errors, and viral recombination (Wang *et al.*, 2021). Mutants for SARS-CoV-2 can be formed by recombination of viral strains as shown by Jackson *et al.* (2021). Additionally, treatments can lead to selection pressures that lead to escape mutants as shown by Ragonnet-Cronin *et al.* (2023). The generation of mutations and the population's behaviour led to the rapid transmission of newly emerged strains that allowed the pandemic to persist and the need for additional booster vaccine programs (McLean *et al.*, 2022).

As with many viruses, the host's immune responses determine the outcome of the infection. This was especially the case in the early stages of the pandemic, where there were no recommended treatments or vaccines. Those individuals who displayed an immediate onset of type I and III interferon (IFN) secretion showed few or no symptoms with controlled viral replications (Vora *et al.*, 2021). However, like many other viruses, SARS-CoV-2 encodes four proteins that act as IFN antagonists. These are non-structural proteins (NSP) 13, NSP14, NSP15, and open reading frame (ORF) 6 (Yuen *et al.*, 2020), and they result in a delayed or poor IFN response, improved viral replication and increased symptom severity. Furthermore, it has been shown in individuals with a poor IFN response; an accumulation of inflammatory macrophages causes an increase in inflammatory

cytokines and poor T-cell-mediated immune response, which harms the host's outcome. (Vora *et al.*, 2021).

1.3 Ebola virus

Ebola virus is a member of the *Filoviridae* family in the order of mononegavirales. The first reported case of a virus from this family was the Marburg virus (1967) in Germany and former Yugoslavia (Feldmann & Geisbert, 16th November 2010). EBOV was first reported in 1976 with two outbreaks in South Sudan and Zaire (now known as the Democratic Republic of Congo (DRC)), where the causative agent was identified, leading to the name Ebola after a river located in the DRC. Currently, there are six reported species of *ebolavirus*, which are Ebola, Sudan, Taï Forest, Bundibugyo, Reston, and Bombali virus. Four of these species are infectious agents in humans (*Zaire, Sudan, Taï Forest* and *Bundibugyo*). Reston virus has only been seen to cause disease in non-human primates (NHPs) and pigs. Bombali virus has been recently found in bats, but whether it causes disease in other animals or humans is unknown (Centers for Disease Control and Prevention, October 2018; Geisbert *et al.*, December 2003; Goldstein *et al.*, 2018).

EBOV is a zoonotic virus; it can be transferred to humans by animals acting as a reservoir, such as bats and NHPs. Transfer to humans can be from eating contaminated bush meats or contact with infected fluids from the animals (Centers for Disease Control and Prevention, October 2018). Intraspecies interactions allow rapid transmission of the virus through various ways, such as caring for an infected individual or handling the corpse from a fatality of the virus. Sexual intercourse with an infected individual can cause transmission even after symptoms have ceased presenting (Feldmann & Geisbert, 16th November 2010). The early symptoms of EBOV are a primary reason that in the initial stages of infection, there is difficulty confirming a diagnosis since they mimic many other infections. Mortality rates between virus strains vary, with the Zaire strain having the highest fatality rates of $60 - 90\%$, whilst Sudan is lower at $40 - 60\%$. Bundibugyo has only had two reported outbreaks with fatality rates of 25% and 51%. Taï Forest virus has had one reported case of which the infected individual recovered in 1994 (World Health Organization, May 2019).
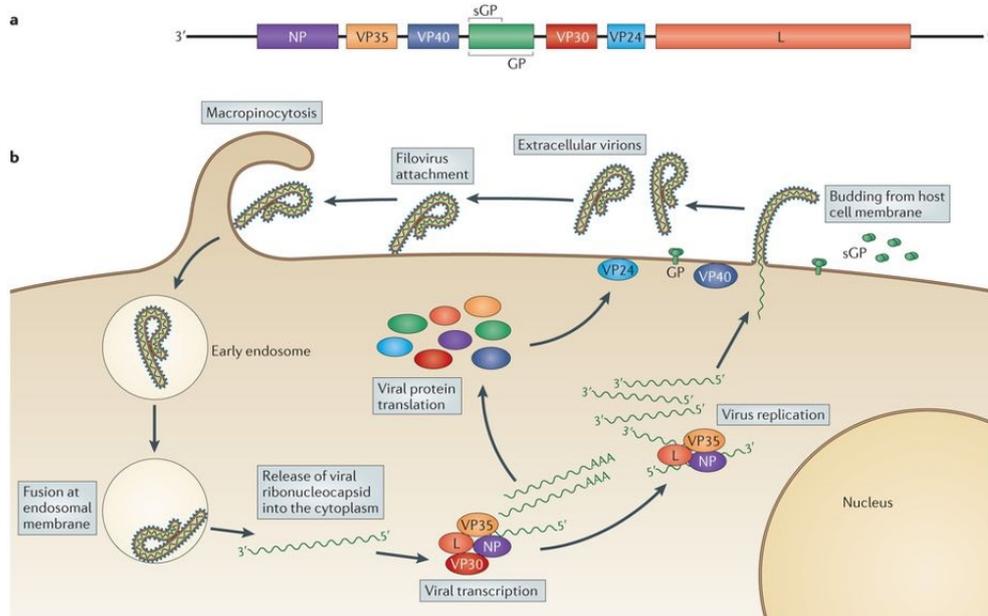
Figure 1.2: (a): Genome structure of EBOV 3' leader non-coding region followed by genes for nucleoprotein (NP), viral protein (VP) 35, VP40, glycoprotei (GP), VP30, VP24, RNA polymerase (L on diagram) finally the 5' trailer non-coding region. (b): Life cycle of EBOV from its entry into the cells after being engulfed via macropinocytosis by host cells cell membrane to its release of viral products after binding and subsequent production of further virus particles and their release from the cell by budding. Figure taken from Messaoudi *et al.* (6th October 2015). Copyright License Number: 5596600082830.

EBOV is an enveloped, non-segmented, negative-stranded RNA virus of varying lengths up to 1400nm but with a uniform diameter of 80nm (Feldmann & Geisbert, 16th November 2010). As can be seen in Figure 1.2(a), the genome consists of a 3' non-coding leader followed by nucleoprotein (NP), viral protein (VP) 35, VP40, glycoprotein (GP), VP30, VP24, RNA polymerase and finally the 5' non-coding region trailer (Messaoudi *et al.*, 6th October 2015). All the genes code for a single protein except for the GP region. The nucleoprotein associates with VP35, VP30 and RNA-dependant RNA polymerase to form the functional transcriptional-replicase complex to allow viral replication. Viral proteins play a crucial role in the subversion of the immune response: VP24 interferes with IFN signalling,

VP35 acts as an antagonist to IFN induction, and VP40 is a matrix protein that mediates particle formation. VP30 acts as an NP transcription primer and plays a role in viral replication to assemble virions (**?**).

Laboratory experimentation elucidated the pathogenesis of EBOV in humans using NHPs such as Cynomolgus Macaques. These animals provide the best analogue to humans due mainly to similarities in physiology (Geisbert *et al.*, December 2003). Upon infection through contact with contaminated fluids, the primary targets for the virus are macrophages and dendritic cells (DC) in lymphoid tissues and Kupfer cells in the liver. Progression of infection leads to further liver cells becoming infected and producing the virus. The infection also migrates to adrenal glands and further collections of tissues to increase its ability to replicate (Geisbert *et al.*, December 2003; Messaoudi *et al.*, 6th October 2015). The effects on macrophages and DCs differ, with the former being induced to produce pro-inflammatory chemokines and cytokines, whilst myeloid DCs have their functions inhibited, and Plasmacytoid dendritic cells (pDCs) have been shown to be refractory to EBOV infection (**?**)

1.4 Thesis objectives

Given the emergence of new SARS-CoV-2 strains and fresh outbreaks of the EBOV, these viruses present an ongoing threat to the global population. Therefore, it is essential to understand immune responses, viral replication dynamics, and potential therapeutics to help minimise the risk associated with infection. Mathematical modelling and statistical analysis can assist with understanding the phenomena. Before any mathematical modelling can be formulated, it is vital to understand the biological processes that drive the assumptions made. This would be impossible without results from biological experimentation, which will provide the backbone for models that we, as mathematicians, formulate to investigate additional questions. Sudden viral outbreaks, such as the 2014 West Africa outbreak and the SARS-CoV-2 pandemic, generate a lot of worry. However, they also provide a wealth of data that can help guide our understanding of disease progression and host immune responses and inform future modelling. It is, therefore, essential to analyse this data and look for relationships between different

serological markers and co-variates, such as age, race, or geographical location. These may suggest cross-reactivity among viral strains, long-term immunity in longitudinal studies and the benefits of vaccination compared to natural infection.

Chapter 3 presents an opportunity to analyse two cohort studies from viral epidemics. The first set of data comes from the 2014 West Africa Ebola outbreak (Section 3.1) and is a collection of immune response data from individuals who are known to have either been infected with EBOV, have been in close contact with an infected individual, and confirmed negative individuals. The unique perspective of this study is, subjects were asked to provide additional samples yearly to investigate longitudinal responses to EBOV infection. Such data allows further investigation into potential waning immunity and whether an individual could become susceptible to a secondary infection. Furthermore, a genetic component was also investigated. Major histocompatibility complex (MHC) alleles were typed to examine if any correlations could be found between immune responses and these alleles, as seen in other studies such as presented by Kiepiela *et al.* (2004). The second study was the St. Jude Tracking of Viral and Host Factors Associated with COVID-19 study (SJTRC) (Section 3.2) to investigate the longitudinal host responses of SARS-CoV-2 for the same set of individuals before and after infection or vaccination, something previous studies had lacked. Since SARS-CoV-2 is also a coronavirus (Lamers & Haagmans, 2022), this study provided the opportunity to examine potential cross-reactivity between endemic circulating strains of human common coronaviruses (hCCoV) and whether this provided any protection. Statistical methods introduced in Section 2.2 shall be used to investigate differences amongst clinical groups and assess the differences in longitudinal immune responses.

As highlighted previously, immune responses are broken into two distinct arms innate and adaptive immune responses. Many viral genomes encode proteins to antagonise these, with antagonism of type I IFN responses in the innate arm being a primary target for many viruses, including EBOV and SARS-CoV-2. Therefore, it seems prudent to investigate a mathematical model that can characterise this inhibition of the viral pathway. One caveat, however, is that the pathway for type I IFN is a highly complex system and data availability is limited *in vivo* due to the early response. Therefore, a complex model, which could incorporate all the

details, is impossible to parameterise. As a result, in Chapter 4, three minimal stochastic models are introduced that aim to characterise, to a first approximation, inhibition by viral proteins of type I IFN induction. When possible, stochastic descriptors will be used to study model dynamics using matrix analytic approaches; if the calculations are computationally tractable. Where these methods are not possible, Gillespie simulations will be used in their place (Section 2.3.1). Some small data sets are available *in vivo* for EBOV infection. Hence, EBOV will be used as a case study where Bayesian inference methods will allow calibration of each model, along with model selection (Section 2.4) to help determine which model best characterises the data. To enable the parameterisation of each model, mean field approximations will be determined. Additionally, this will allow for assessing parameter identifiability and sensitivity (Section 2.5), providing necessary information about our parameters, which must be carefully considered. All this information together will be used to identify which model would be best suited for limited data that comes from *in vivo* samples during viral outbreaks for future analysis of other potential viruses, since many inhibit the type I IFN pathway.

SARS-CoV-2 has many different approved vaccines as summarised by Creech *et al.* (2021). However, the continued emergence of new mutant strains and waning immunity highlights a need for additional prophylactics. A potential prophylactic is therapeutic interring particles (TIPS), sometimes called defective interring particles (DIPS or DIS). DIPs naturally occur during viral replication. These defective particles lack vital parts of the viral genome required to replicate. As a result, DIPs sequester replication machinery from functional viruses and lead to a reduction in the yield of infectious viruses (Wu *et al.*, 2022). These have been investigated within-host by Chaturvedi *et al.* (2021). However, little has been done to understand the intra-cellular cellular level of replication of SARS-CoV-2 in the presence of DIPs. Therefore, in Chapter 5, an extension of the model proposed by Grebennikov *et al.* (2021) to include the intra-cellular replication of SARS-CoV-2 in the presence of DIPS is introduced. The model examines the dynamics of co-replication and how initial dosages affect the overall release of SARS-CoV-2 virions. This model is provided in a deterministic and stochastic representation to explore the differences between the approaches. The final part of Chapter 5 provides some preliminary investigation into *in vitro* replication dynamics of Foot

and Mouth Disease virus (FMDV) with data sets provided for three viral strains. Here a deterministic approach is taken for the model's parameterisation and to determine the sensitivity and identifiability of the parameters. From there, the model is used to determine the basic reproduction number $\mathcal{R}_0$ along with the time to infection. Finally, a stochastic model is introduced to investigate the effects of initial MOI dosages on time to extinction and the probability of extinction within the experimental window.

1. INTRODUCTION

Chapter 2

Mathematical background

Biological processes are subject to stochastic effects. From a modeller's perspective, describing these systems with deterministic approaches may be less preferential. A key feature of many viruses is they require a low dose to cause an infection. In these cases, deterministic modelling approaches will provide only opposing outcomes. Either an infection persists and reaches a steady state or becomes extinct. Stochastic processes permit a modeller to capture the inherent randomness associated with small populations. This enables the study of different potential outcomes given the same initial conditions and model parameters. Furthermore, it allows for studying different descriptors, such as the mean time to viral extinction. One of the limitations of deterministic models. That being said, stochastic processes introduce extra mathematical details and complexity that limit the analytical work that can be performed. As such, this highlights the necessity for deterministic models that are easier to implement and understand. Therefore, a balance between both stochastic and deterministic approaches is vital.

This chapter will introduce probability theory and an overview of stochastic processes used to develop models in Chapter 4 and Chapter 5. Analytic methods to study stochastic processes are outlined, and where these methods cannot be implemented, stochastic simulation algorithms shall be used. Bayesian methods for parameter inference and model selection are also introduced as these provide a way for comparison between mathematical models and experimental data and models incorporating different biological hypotheses. Furthermore, a brief

overview of multi-variate analysis concepts will also be presented that relate to work undertaken in Chapter 3.

2.1 Probability theory

Here I will introduce some fundamental probability theories that will be an underlying feature in the stochastic processes that I will consider. The following definitions can be found in Allen (2010); Kijima (1997); Ross (2014).

Suppose for a random variable X that can take values from some sample space S. Then for a value $x \in S$ one can define the **c.d.f. (cumulative distribution function)** as

$$F_X(x) = \mathbb{P}(X \leq x), \ x \in S,$$

where $F_X(x) : \mathbb{R} \to [0, 1]$. If X is a discrete random variable such that S is countable then one can define the **p.m.f. (probabilty mass function)** as

$$f_X(x) = \mathbb{P}(X = x), \ x \in S.$$

Finally if X is a continuous random variable and there exists a function $f_X(x)$ that is non-negative and integrable such that the following holds

$$\mathbb{P}(a \leq x \leq b) = \int_a^b f_X(x)dx, \ a, b \in S, \ a < b,$$

then I can call the function $f_X(x)$ the **p.d.f. (probability density function)**. For a continuous random variable, I can define the expectation of X as

$$\mathbb{E}[X] = \int_S x f_X(x)dx.$$

For a discrete random variable, one has

$$\mathbb{E}[X] = \sum_{x \in S} x f_X(x).$$

These allow me to propose the following definitions.

Definition 2.1.1. The **mean** of a random variable X denoted μ_X is simply the expectation of X, $\mathbb{E}[X] = \mu_X$. The **variance** denoted by σ_X^2 or $\text{Var}(X)$ is

given by $\text{Var}(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}[X^2] - \mu_X^2$. The square root of the variance gives the **standard deviation** (SD) of the random variable X. The **skewness** denoted by $\tilde{\mu}_3$ or $\text{Skew}(X)$ is the third standardised moment and is given by $\text{Skew}(X) = \mathbb{E}[(X - \mu_X)^3]/(\mathbb{E}[(X - \mu_X)^2])^{3/2}$. **Kurtosis** of a random variable is defined as the fourth standardised moment, denoted by γ_2 or $\text{Kurt}(X)$ and is given by $\text{Kurt}(X) = \mathbb{E}[(X - \mu_X)^4]/(\mathbb{E}[(X - \mu_X)^2])^2$.

These definitions also allow me to define the expectation of a function $g(X)$ where X can be either a continuous or discrete random variable. In the continuous case, I have

$$\mathbb{E}[g(X)] = \int_S g(x) f_X(x) dx,$$

whilst in the discrete case it can be defined as

$$\mathbb{E}[g(X)] = \sum_{x \in S} g(x) f_X(x).$$

This gives the following definition of the probability generating function.

Definition 2.1.2. For a discrete random variable X defined on some set $S \subseteq \mathbb{N} \cup \{0\}$, the **p.g.f. (probability generating function)** is given by

$$\mathcal{P}_X(s) = \mathbb{E}[s^x] = \sum_{j=0}^{+\infty} s^j \mathbb{P}(X = j),$$

for $s \in \mathbb{R}$ for which the sum converges.

Next, I will introduce several probability distributions. First, I present the exponential distribution, which will be important when considering Markovian stochastic processes. Followed by several other well-known distributions that are used throughout this thesis.

Definition 2.1.3. A continuous random variable X with p.d.f. given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

for some parameter $\lambda > 0$ is referred to as **exponentially distributed** with parameter λ. The c.d.f. of an exponential distribution is given by

$$F_X(x) = 1 - e^{-\lambda x}, \ x \geq 0.$$

The mean and variance are given by

$$\mathbb{E}[X] = \frac{1}{\lambda}, \qquad Var(X) = \frac{1}{\lambda^2}.$$

Definition 2.1.4. A random variable X is said to be **normally distributed** or otherwise referred to as **normal random variable** X with parameters μ and σ^2, if its p.d.f. is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)/2\sigma^2}, \ x \in \mathbb{R},$$

with the mean and variance given by

$$\mathbb{E}[X] = \mu, \qquad Var(X) = \sigma^2.$$

Definition 2.1.5. A random variable X is said to have a gamma distribution with parameters $\alpha > 0$ and $\lambda > 0$, if its p.d.f. is given by

$$f_X(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}, & \text{if } x \geq 0, \\ 0, & \text{otherwise}, \end{cases}$$

where $\Gamma(\alpha)$, referred to as the *gamma function*, is defined as,

$$\Gamma(\alpha) = \int_0^{+\infty} e^{-y} y^{\alpha-1} dy.$$

For a gamma random variable, the mean and variance are,

$$\mathbb{E}[X] = \frac{\alpha}{\lambda}, \qquad Var(X) = \frac{\alpha}{\lambda^2}.$$

A special case of the gamma distribution with $\lambda = 1/2$ and $\alpha = n/2$, where n is a positive integer, is called the χ_n^2 ("chi-squared") distribution with n degrees of freedom. This special gamma distribution is frequently used in inferential statistics, particularly hypothesis testing.

Definition 2.1.6. A random variable X is a uniform random variable on the interval (α, β) if its p.d.f. is given by

$$f_X(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \text{if } \alpha < x < \beta, \\ 0, & \text{otherwise}. \end{cases}$$

16

The c.d.f. of a uniform distribution is given by

$$
f_X(x) = \begin{cases} 0, & \text{if } x \leq \alpha, \\ \frac{x-\alpha}{\beta-\alpha}, & \text{if } \alpha < x < \beta, \\ 1, & \text{if } x \geq \beta. \end{cases}
$$

The mean and variance of a uniform random variable X are,

$$
\mathbb{E}[X] = \frac{\beta + \alpha}{2}, \qquad Var(X) = \frac{(\beta - \alpha)^2}{12}.
$$

2.2 Multi-variate analysis

In Chapter 3, I will analyse two cohort studies of viral outbreaks. The first study is from the UK Health Security Agency (UKHSA, formally PHE) on longitudinal data of patients from the 2014 West Africa Ebola outbreak. The second study is from St Jude Children's Research Hospital investigating immune response to SARS-CoV-2. Several multi-variate methods will be used during this analysis, so several basic definitions are provided here. The definitions here come from Manly & Alberto (2016) and Chatfield (2018). The first definition is for a sample mean and sample variance. In what follows, capital letters such as $\boldsymbol{X}$ will denote random variables. In contrast, lowercase $\boldsymbol{x}$ shall denote samples within a data set or from a distribution.

Definition 2.2.1. Let $\boldsymbol{X}$ be a p-dimensional random variable such that $\boldsymbol{X}^T = [X_1, \ldots, X_p]$ where $X_1, \ldots, X_p$ are uni-variate samples from $\boldsymbol{X}$ such that the jth sample is denoted $\boldsymbol{x}_j^T = (x_{j1}, \ldots, x_{jp})$. If there are n samples then the sample mean of x_i is defined as

$$
\bar{x}_i = \sum_{j=1}^{n} x_{ji}/n,
$$

where x_{ji} represents the jth sample for the ith variable. The sample mean vector is defined as, $\bar{\mathbf{x}}^T = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$, and often to referred by $\boldsymbol{\mu}$. The **sample variance** of each p variables can also be defined,

$$
s_i^2 = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2/(n-1).
$$

2. MATHEMATICAL BACKGROUND

Also provided here are definitions for the sample covariance matrix and sample correlation matrix. The sample covariance matrix is as follows,

Definition 2.2.2. Suppose there are p variables where the ith variable x_i has sample mean $\bar{x}_i$ and variance s_i^2, whilst the kth variable x_k has sample mean $\bar{x}_k$ and variance s_k^2. Then the sample covariance between x_i and x_k is given by

$$\sigma_{ik} = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)/(n-1).$$

One can write this in a matrix known as the sample **covariance matrix** which is constructed as

$$\mathbf{S} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots\cdots\cdots & \sigma_{1p} \\ \sigma_{21} & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ \sigma_{p1} & \cdots\cdots\cdots\cdots & & \sigma_{pp} \end{pmatrix}$$

where $\sigma_{ii} = s_i^2$ is the ith variable sample variance.

Definition 2.2.3. Suppose that for the ith variable x_i with sample variance s_i^2 and the kth variable x_k with variance s_k^2 and sample covariance σ_{ik} . Then the **correlation coefficient** is given by

$$r_{ik} = \frac{\sigma_{ik}}{s_i s_k}.$$

This can be formed into a matrix known as the **sample correlation matrix** written as,

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots\cdots\cdots\cdots & r_{1p} \\ r_{21} & & & \\ & & & \\ & & & \\ & & & \\ & & & r_{p-1p} \\ r_{p1} & \cdots\cdots\cdots & r_{pp-1} & 1 \end{pmatrix}$$

where $r_{ii} = 1$ and the matrix $\mathbf{R}$ is symmetric.

If the variables have been standardised such that they have a sample mean of zero and a sample variance of one, then the sample covariance matrix is just the correlation matrix.

2.2.1 Multi-variate normality

One final definition will be for the multi-variate normal distribution.

Definition 2.2.4. A p-dimensional random variable $\boldsymbol{X}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is said to be **multi-variate normal distributed** if it has a p.d.f. of the form,

$$f_{\boldsymbol{X}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\},$$

this will be denoted as $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Before analysing a data set, it is essential to know whether the data is multi-variate normal (MVN) to apply tests such as Hotelling's T^2 (Chatfield, 2018). To do this, use the following MVN test.

Mardia Test

The Mardia test uses extensions of skewness and kurtosis to assess a data sets multi-variate normality. Suppose I have a p-dimensional random variable where n samples set have been drawn $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ then the skewness is defined as $(\tilde{\mu}_{3,p})$ and Kurtosis $(\gamma_{2,p})$ by the following measures,

$$\tilde{\mu}_{3,p} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij}^3, \ \ \gamma_{2,p} = \frac{1}{n} \sum_{i=1}^{n} m_{ii}^2,$$

where $m_{ij} = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})\boldsymbol{S}^{-1}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})$ which is the Mahalanobis distance such that $\bar{\boldsymbol{x}}$ is the sample mean vector and $\boldsymbol{S}$ is the sample covariance matrix. The test statistic for skewness is $(n/6)\tilde{\mu}_{3,p}$ which is approximately χ^2 distributed with $p(p+1)(p+2)$ degrees of freedom. Meanwhile, the test statistic for Kurtosis is $(\gamma_{2,p} - p(p+2))/(8p(p+2)^{\frac{1}{2}})$ which is approximately normally distributed ($\approx N(0,1)$). If the sample size is small ($n < 20$) then an adjustment is required for the skewness such that $(nk/6)\tilde{\mu}_{3,p} \sim \chi^2$ where $k = (p+1)(n+1)(n+3)/(n(n+1)(p+1) - 6)$

(Korkmaz *et al.*, 2014; Mardia, 1974). The calculation for degrees of freedom remains the same.

The null hypothesis asserts the data is multi-variate normally distributed. These two tests are performed, and to reject the null hypothesis, either of the tests must disagree. To accept the null hypothesis, both tests must agree.

2.2.2 Principal component analysis

Principal component analysis provides a way to examine relationships between a set of p correlated variables by transforming them into a set of p uncorrelated variables. These are referred to as principal components (PC). The principal components are linear combinations of the original variables, where each PC can potentially hold more variability than the original single variable. These new variables are derived in a decreasing order such that the first PC accounts for the most variance within the data possible. A full detailed explanation can be found in Chatfield (2018).

Lets suppose I have a p-dimensional random variable $\boldsymbol{X}^T = [X_1, \dots, X_p]$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. One can define a new set of random variables $\boldsymbol{Y}^T = [Y_1, \dots, Y_p]$, which are uncorrelated, and the variance decreases from first to last. Then each $\boldsymbol{Y}_i$ can be taken as a linear combination of $\boldsymbol{X}_i$ such that;

$$\begin{aligned} \boldsymbol{Y}_j &= \alpha_{1j}X_1 + \alpha_{2j}X_2 + \cdots + \alpha_{pj}X_p \\ &= \boldsymbol{\alpha}_j^T \boldsymbol{X}, \end{aligned}$$

where $\boldsymbol{\alpha}_{\boldsymbol{j}}^T = (\alpha_{1j}, \dots, \alpha_{pj})$ is a vector of constants. A condition is imposed $\boldsymbol{\alpha}_{\boldsymbol{j}}^T \boldsymbol{\alpha}_{\boldsymbol{j}} = 1$, so distances within the space are preserved. The vectors $\boldsymbol{\alpha}_{\boldsymbol{j}}$ can be determined by calculating eigenvectors from the covariance matrix, where eigenvalues are assumed to be ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Since this new PC system is ordered such that the later components may only account for a small part of the total variance, keeping them all is unnecessary. To choose how many components are retained, one can use Kaiser's rule and keep those eigenvalues ≥ 1 (Kaufman & Dunlap, 2000) or those that account for 90% of the total variance within the data (Manly & Alberto, 2016).

2.2.3   Multi-variate ε test

Suppose I have two groups with samples $\boldsymbol{x}$ and $\boldsymbol{y}$ drawn from two random p-dimensional distributions $\boldsymbol{X}$ and $\boldsymbol{Y}$. A question may be asked whether there are any statistical differences between these groups and hence their distributions. If samples $\boldsymbol{x}$ and $\boldsymbol{y}$ are drawn from multi-variate normal distributions, then a wide range of methods can be used. These are summarised in Manly & Alberto (2016) and Chatfield (2018). However, if the samples are not drawn from multi-variate normal distributions, one may wish to use non-parametric tests to investigate similarities between groups. One such test is the multi-variate ε test, defined as follows.

Suppose I have $k \geq 2$ independent p-dimensional random samples $\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots, \boldsymbol{x}_k \in \mathbb{R}^p$ with sample sizes $n_1, n_2, \ldots, n_k$, respectively, from p-dimensional distributions $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k$. I consider the null hypothesis

$$H_0 : \boldsymbol{X}_1 \equiv \boldsymbol{X}_2 \equiv \cdots \equiv \boldsymbol{X}_k \tag{2.1}$$

and

$$H_1 : \boldsymbol{X}_1 \neq \boldsymbol{X}_2 \neq \ldots \neq \boldsymbol{X}_k \tag{2.2}$$

Consider a pooled sample such that $\boldsymbol{x}_1 \bigcup \boldsymbol{x}_2 \bigcup \cdots \bigcup \boldsymbol{x}_k = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n\}$ so the total sample size is $\sum_{i=1}^{k} n_i = n$. Under the null hypothesis, each $\boldsymbol{w}_i$ is independent and identically distributed from a common distribution $\boldsymbol{X}$. I first define a desired significance level α then resample without replacement samples of size n from the pooled sample $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n\}$ for B replicates such that $(B + 1)\alpha \in \mathbb{Z}^+$. Let $m_j = \sum_{i=1}^{j} n_i$ with $m_0 = 0$, then for each bootstrap replicate $\boldsymbol{w}_1^{(b)}, \boldsymbol{w}_2^{(b)}, \ldots, \boldsymbol{w}_n^{(b)}$ with $b \in \{1, \ldots, B\}$, the kth sample of the bootstrap replicate b is defined as, $\mathbf{x}_i^{(b)} = \{\boldsymbol{w}_{m_{i-1}+1}^{(b)}, \ldots, \boldsymbol{w}_{m_i}^{(b)}\}$. The statistic $\varepsilon_n^{(b)}$ is defined as,

$$\varepsilon_n^{(b)} = \sum_{1 \leq i \leq j \leq k} \varepsilon_{n_i, n_j}(\mathbf{x}_i^{(b)}, \mathbf{x}_j^{(b)}) = \sum_{1 \leq i \leq j \leq k} e(\mathbf{x}_i^{(b)}, \mathbf{x}_j^{(b)}), \tag{2.3}$$

where $e(\mathbf{x}_i, \mathbf{x}_j)$ is defined for random samples $\mathbf{x}_i$ and $\mathbf{x}_j$ with sample size n_i and

n_j as,

$$
\begin{aligned}
e(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_i n_j}{n_i + n_j} & \left(\frac{2}{n_i, n_j} \sum_{l=1}^{n_i} \sum_{m=1}^{n_j} ||x_{il} - x_{jm}|| \right. \\
& \left. - \frac{1}{n_i^2} \sum_{l=1}^{n_i} \sum_{m=1}^{n_i} ||x_{il} - x_{im}|| - \frac{1}{n_j^2} \sum_{l=1}^{n_j} \sum_{m=1}^{n_j} ||x_{jl} - x_{jm}|| \right).
\end{aligned}
\tag{2.4}
$$

One can then calculate the bootstrap estimate $\mathbb{P}_n(\varepsilon_n \leq \varepsilon)$ as $\frac{\#\varepsilon_n^{(b)} \leq \varepsilon}{B}$ for some value ε, where I can reject the null hypothesis when the observed ε_n exceeds $100(1-\alpha)\%$ of the replicates (Efron, 1982; Székely *et al.*, 2004). Here, $\#\varepsilon_n^{(b)} \leq \varepsilon$ means the number of $\varepsilon_n^{(b)}$ statistics less than ε.

2.3 Stochastic processes

I shall now give definitions relevant to stochastic processes. These can be found in Allen (2010) and Padma & Vijayalakshmi (2011). The first definition is that of a stochastic process.

Definition 2.3.1. A **stochastic process** is a collection of random variables $\mathcal{X} = \{X(t) : t \in T\}$ where $X(t) \in S$, such that S is our state space that can be either finite or infinite (*e.g.* in the discrete case $\{0, 1, 2, \ldots, N\}$ or $\{0, 1, 2, \ldots\}$). Here T is defined as the index set, and since many of the applications I will discuss shall be biological processes taking place in some given time interval, I shall consider this as time $T = [0, +\infty)$.

The previous definition will allow me to define a continuous time Markov chain, a particular case of a stochastic process, such that the future state only depends on the current state and not on its past. The formal definition follows.

Definition 2.3.2. Let $\mathcal{X}$ be a collection of discrete random variables with values in a finite or infinite "state space" S (e.g., $\{0, 1, 2, \ldots, N\}$ or $\{0, 1, 2, \ldots\}$) with index set $T = [0, \infty)$. The stochastic process $\mathcal{X}$ is called a **continuous time Markov chain** (CTMC) if it satisfies

$$
\begin{aligned}
& \mathbb{P}(X(t_{n+1}) = i_{n+1} | X(t_0) = i_0, X(t_1) = i_1, \ldots X(t_n) = i_n), \\
& = \mathbb{P}(X(t_{n+1}) = i_{n+1} | X(t_n) = i_n),
\end{aligned}
$$

for any set of real numbers that satisfy $0 \le t_0 \le t_1 \le \cdots \le t_{n+1}$ and any $i_j \in S$, $j \in \{0, \dots, n+1\}$.

Suppose I have a stochastic process and want to find the probability for that process to go from state i at time s to state j at time t for two arbitrary states $i, j \in S$. Then the following definition is given,

Definition 2.3.3. Given a CTMC $\mathcal{X}$, the transition probability to get from a state i at time s to a state j at time t is defined as,

$$p_{ij}(s, t) = \mathbb{P}(X(t) = j | X(s) = i), \; s < t,$$

where $i, j \in S$. The CTMC is considered homogeneous if these probabilities only depend on the length of time interval $t - s$ and not on the particular values of s and t. Hence

$$p_{ij}(s, t) = p_{ij}(t - s) = \mathbb{P}(X(t) = j | X(s) = i) = \mathbb{P}(X(t - s) = j | X(0) = i),$$

for $s < t$. The matrix containing transition probabilities, otherwise known as the transition matrix, is defined as,

$$\mathbf{P}(t) = (p_{ij}(t))_{i,j \in S},$$

where $t \in [0, +\infty)$.

Definition 2.3.4. From the transition probabilities $p_{ij}(t)$ one can derive the transition rates q_{ij} to form the generator matrix $\mathbf{Q}$. Let me assume the transition probabilities $p_{ij}(t)$ are continuous and differentiable for $t \ge 0$ and at $t = 0$,

$$p_{ij}(0) = 0, \; i \ne j, \; p_{ii}(0) = 1, \qquad \forall i \in S.$$

Define for any $i, j \in S$

$$q_{ij} = \lim_{\Delta t \to 0^+} \frac{p_{ij}(\Delta t) - p_{ij}(0)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{p_{ij}(\Delta t)}{\Delta t}, \qquad i \ne j, \tag{2.5}$$

and similarly

$$q_{ii} = \lim_{\Delta t \to 0^+} \frac{p_{ii}(\Delta t) - p_{ii}(0)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{p_{ii}(\Delta t) - 1}{\Delta t}. \tag{2.6}$$

2. MATHEMATICAL BACKGROUND

Since $\sum_{j\in S} p_{ij}(\Delta t) = 1$ (Allen, 2010; Padma & Vijayalakshmi, 2011) the following holds,

$$1 - p_{ii}(\Delta t) = \sum_{j\in S, j\neq i} p_{ij}(\Delta t) = \sum_{j\in S, j\neq i} [q_{ij}\Delta t + \mathcal{O}(\Delta t)],$$

resulting in the following expression for q_{ii},

$$q_{ii} = \lim_{\Delta t \to 0^+} \frac{-\sum_{j\in S, j\neq i}[q_{ij}\Delta t + \mathcal{O}(\Delta t)]}{\Delta t},$$
$$= -\sum_{j\in S, j\neq i} q_{ij}, \tag{2.7}$$

where $\sum_{j\neq i} \mathcal{O}(\Delta t) = \mathcal{O}(\Delta t)$. The limits in Equations (2.5) and (2.6) can be expressed in terms of a matrix $\mathbf{Q}$. Let $\mathbf{P}(\Delta t)$ be the infinitesimal transition matrix and $\mathbf{I}$ be the identity matrix of the same size. Then the matrix $\mathbf{Q}$ is equal to

$$\mathbf{Q} = \lim_{\Delta t \to 0^+} \frac{\mathbf{P}(\Delta t) - \mathbf{I}}{\Delta t}. \tag{2.8}$$

Definition 2.3.5. Using the definitions in Equations (2.5), (2.6) and (2.8), I can define the matrix of transition rates $\mathbf{Q} = (q_{ij})_{i,j\in S}$ known as the **infinitesimal generator matrix**. For example if $S = \{0, 1, 2, \dots\}$, one gets

$$\mathbf{Q} = \begin{pmatrix} q_{00} & q_{01} & q_{02} & \cdots \\ q_{10} & q_{11} & q_{12} & \cdots \\ q_{20} & q_{21} & q_{22} & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

.

Definition 2.3.6. From Definitions 2.3.5 and 2.3.3, I can define the forward and backward Kolmogorov differential equations that describe the rate of change of the transition probabilities. If $\mathbf{P}(t)$ is the transition matrix as defined in Definition 2.3.3, then the forward Kolmogorov differential equation in matrix form is given by

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{Q}\mathbf{P}(t) \qquad \mathbf{P}(0) = \mathbf{I},$$

where $\mathbf{I}$ is the identity matrix. The backward Kolmogorov equations can be defined in a similar manner,

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{Q} \qquad \mathbf{P}(0) = \mathbf{I}.$$

These equations define a system of differential linear equations, which can be solved to find a solution of the form

$$\mathbf{P}(t) = \mathbf{P}(0)\exp(\mathbf{Q}t),$$

where the exponential matrix is defined as,

$$\exp(\mathbf{Q}t) = \sum_{k=0}^{+\infty} \frac{(\mathbf{Q}t)^k}{k!}.$$

Inter-event time

For a CTMC $\mathcal{X} = \{X(t) : t \in T\}$, the time spent in a given state before jumping to the next, can be described by the random variable W_n for the time of the nth jump (assuming $W_0 = 0$). The random variable $\{W_n\}_{n \in \mathbb{N} \cup \{0\}}$ are referred to as waiting times. From these waiting times, a random variable $T_n = W_{n+1} - W_n \geq 0$ can be defined as the time between state changes. These are known as inter-event times.

Assume at the nth jump; the CTMC is at state i such that $X(W_n) = i \in S$. Then the probability of leaving the state i is $\alpha(i)\Delta t + \mathcal{O}(\Delta t)$. Here the process moves to a new state in the time period Δt, where

$$\alpha(i) = \sum_{j \in S , j \neq i} q_{ij} = -q_{ii},$$

hence, the probability that the process remains in the same state, for the length of time Δt is

$$p_{ii}(\Delta t) = 1 - \alpha(i)\Delta t + \mathcal{O}(\Delta t).$$

Let $G_n(t)$ be the probability that $X(W_n)$ remains in state i for some time $t \geq 0$ which in terms of the inter-event time $T_n = W_{n+1} - W_n > t$ means,

$$G_n(t) = \mathbb{P}(W_{n+1} > t + W_n) = \mathbb{P}\{T_n > t\}.$$

For $t = 0$, then $G_n(0) = \mathbb{P}(T_n > 0) = 1$. For sufficiently small Δt,

$$G_n(t + \Delta t) = G_n(t)p_{ii}(\Delta t) = G_n(t)(1 - \alpha(i)\Delta t + \mathcal{O}(\Delta t)).$$

Subtracting $G_n(t)$ from both sides and dividing by Δt and then letting $\Delta t \to 0^+$ one arrives at the following ODE,

$$\frac{dG_n(t)}{dt} = -\alpha(i)G_n(t).$$

Using the initial condition $G_n(0) = 1$, the first order differential equation, can be solved by separation of variables to give

$$G_n(t) = \mathbb{P}(T_n > t) = e^{-\alpha(i)t}.$$

The above equation allows one to compute the c.d.f. of T_n, as

$$\mathbb{P}(T_n \leq t) = 1 - G_n(t) = 1 - e^{-\alpha(i)t}, \ t \geq 0.$$

Therefore the inter-event time corresponds to an exponential random variable with parameter $\alpha(i)$. Hence, T_n is an exponentially distributed random variable. This leads to the following theorem.

Theorem 2.3.1. *Let $\mathcal{X}$ be a CTMC with a transition matrix $\boldsymbol{P}(t) = (p_{ij}(t))_{i,j \in S}$ such that for sufficiently small Δt,*

$$\sum_{j \neq i} p_{ij}(\Delta t) = \alpha(i)\Delta t + \mathcal{O}(\Delta t),$$

and

$$p_{ii}(\Delta t) = 1 - \alpha(i)\Delta t + \mathcal{O}(\Delta t),$$

where $\alpha(i) = \sum_{j \in S, j \neq i} q_{ij} = -q_{ii}$. Then define the inter-event time as $T_n = W_{n+1} - W_n$ (where W_n is the time of the nth jump resulting in a change in the state space). Given $X(W_n) = T_n$, is an exponential random variable with parameter $\alpha(i)$ as defined in Definition 2.1.3. The cumulative distribution function (c.d.f) for T_n is $F_n(t) = 1 - e^{-\alpha(i)t}$ such that the mean and variance are

$$\mathbb{E}[T_n] = \frac{1}{\alpha(i)}, \qquad Var(T_n) = \frac{1}{\alpha(i)^2} \qquad \text{if } \alpha(i) > 0.$$

A well-known CTMC is obtained when I consider a space of discrete states $S = \{0, 1, 2, \dots\}$ and assume that $\mathcal{X}$ can only move one unit to the left or right in each jump. From this, one can define a general birth and death process for a CTMC with birth rates λ_i and death rates μ_i for some small change in the population $\Delta X(t) = X(t + \Delta t) - X(t)$. The definition is as follows.
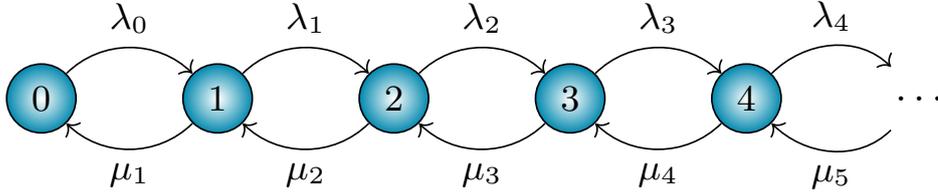
Figure 2.1: Depiction of a birth-death process with birth rate λ_i and death rate μ_i.

Definition 2.3.7. A **general birth and death process** is a continuous time Markov chain $\mathcal{X}$ which can either have a finite or infinite space state ($S = \{0, 1, \dots N\}$ or $S = \{0, 1, 2, \dots\}$). $\mathcal{X}$ can be defined in terms of the probabilities as,

$$p_{ij}(\Delta t) = \mathbb{P}(X(t + \Delta t) = j | X(t) = i)$$

$$= \begin{cases} \lambda_i \Delta t + \mathcal{O}(\Delta t), & j = i + 1, \\ \mu_i \Delta t + \mathcal{O}(\Delta t), & j = i - 1, \\ 1 - (\lambda_i + \mu_i)\Delta t + \mathcal{O}(\Delta t), & j = i, \\ \mathcal{O}(\Delta t), & \text{otherwise}, \end{cases}$$

for sufficiently small Δt where $\lambda_i \geq 0$, $\mu_i \geq 0$ for all $i \in S$ and $\mu_0 = 0$.

The birth and death rates can be related to the generator matrix $\mathbf{Q}$ as follows $\lambda_i = q_{i,i+1}$, $\mu_i = q_{i,i-1}$ and $q_{ii} = -(\lambda_i + \mu_i)$. A depiction of a simple birth-death process is shown in Figure 2.1. In this example, if $\lambda_0 = 0$ then $p_{00}(t) = 1$, and this state is referred to as an **absorbing state** and the overall Markov process is an **absorbing Markov process**.

2.3.1 Stochastic simulation algorithm

Gillespie simulations

Suppose I have a particular stochastic model that is of interest to simulate. There are several stochastic simulation algorithms; one of the most famous methods is the Gillespie simulation algorithm. Suppose for a CTMC $\mathcal{X}$ there is an initial

state $\mathbf{x}(0) = \mathbf{x}$. Let $r_1, \ldots, r_n$ be a list of possible events or reactions with each r_i, having associated transition rate $q_i(\mathbf{x})$, that is as described for the generator matrix in Definition 2.3.5. Then $q_i(\mathbf{x})\Delta t$ defines the probability that this reaction occurs in a short time interval $(t, t + \Delta t)$. One can then define $\Delta\mathbf{x}_i$ such that the entries correspond to a change in $\mathbf{x}$ for a given reaction r_i (Erban *et al.*, 2007). This method is carried out until a given condition is met. The Gillespie algorithm uses the following theorem to draw a new time step Δt from the exponential distribution using the uniform distribution (Gillespie, 2007). Algorithm 2.1 describes the steps for Gillespie simulations.

Theorem 2.3.2. *Let U be a uniform random variable defined on $[0, 1]$ and let T be a continuous random variable on $[0, +\infty)$. Then $T = F^{-1}(U)$, where F is the cumulative distribution function of the random variable T.*

Theorem 2.3.1 has previously shown that an exponential distribution for continuous-time Markov chains can describe the inter-event time. If then $U = F(T)$ where F is the c.d.f. of the exponential distribution with T being the inter-event time and U a uniform random variable, then one can rearrange to find $T = F^{-1}(U)$ as stated in Theorem 2.3.2. Calculating the inverse c.d.f. of F produces $T = -log(U)/\alpha(\mathbf{x})$, where $\mathbf{x}$ is the current state of the process. Hence the inter-event time can be constructed by transforming a uniform random variable.

Algorithm 2.1: Gillespie simulation algorithm

Let $t = 0$ and set initial population $\mathbf{x} = \mathbf{x}(0)$.

while *some condition* **do**

    Draw two uniformly distributed numbers u_1 and u_2 from the unit interval $u_1, u_2 \in [0, 1]$;

    Calculate $q_i(\mathbf{x})$ for $i = 1, 2, \ldots, n$;

    Calculate $\alpha(\mathbf{x}) = \sum_{i=1}^{n} q_i(\mathbf{x})$ otherwise known as the total propensity function;

    Choose small time increment $\Delta t = -log(u_1)/\alpha(\mathbf{x})$;

    Update time $t \leftarrow t + \Delta t$;

    Select the reaction r_i that satisfies $\frac{\sum_{j=1}^{i-1} q_j(\mathbf{x})}{\alpha(\mathbf{x})} < u_2 \leq \frac{\sum_{j=1}^{i} q_j(\mathbf{x})}{\alpha_{(\mathbf{x})}}$;

    Update population $\mathbf{x} \leftarrow \mathbf{x} + \Delta\mathbf{x}_i$;

end

2.4   Bayesian inference and model selection

This section introduces ideas of Bayesian inference that will be used to infer parameters for the models presented, given observed experimental data. Two variations of the Approximate Bayesian Computation (ABC) method will be discussed based on Bayes' theorem. Ross (2014) discusses Bayes' theorem and many other probability texts, which relate the conditional probabilities of two events A and B. The relation is given by

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)}.$$

In Bayesian inference this relation is instead formulated as

$$\pi(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{D}|\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta})\pi(\boldsymbol{D}|\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{2.9}$$

where $\boldsymbol{\theta}$ is a vector of model parameter values, and $\boldsymbol{D}$ is observed experimental data. In this formula, $\pi(\boldsymbol{\theta})$ is known as the prior distribution and represents our

prior beliefs about the parameter(s). Priors can be referred to as either informative: such as a normal distribution, or non-informative: such as a uniform distribution. Suppose an individual has strong prior knowledge about the true parameter value. In that case, an informative prior gives increased density to the region of the parameter space where the true value is thought to be. A non-informative prior allows upper and lower bounds that cover a broad parameter space. $\pi(\boldsymbol{D}|\boldsymbol{\theta})$ is the likelihood of observing the data $\boldsymbol{D}$ given the parameter vector $\boldsymbol{\theta}$, and $\pi(\boldsymbol{\theta}|\boldsymbol{D})$ is the posterior distribution of our parameter vector $\boldsymbol{\theta}$ as a result of our observed data $\boldsymbol{D}$. The integral in the denominator of (2.9) is a normalisation constant; as such, (2.9) can be written in a simpler form as a proportionality equation,

$$\pi(\boldsymbol{\theta}|\boldsymbol{D}) \propto \pi(\boldsymbol{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

In Sections 2.4.1 and 2.4.2 two methods for estimating the posterior distribution are introduced that omit the calculation of the likelihood function. This is preferred since calculating the likelihood function for mathematical models is not always possible.

2.4.1 Approximate Bayesian computation-rejection

Given a mathematical model $\mathcal{M}$, and a set of observed data $\boldsymbol{D}$, which is parameterised by the vector $\boldsymbol{\theta}$, approximate Bayesian computation (ABC) can be used to infer posterior distributions for the parameter values. ABC allows a user to define a set of prior beliefs about parameter distributions, $\pi(\boldsymbol{\theta})$, and combine this with model simulations and data to arrive at a posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{D})$. Given a sample parameter set $\boldsymbol{\theta}^* \sim \pi(\boldsymbol{\theta})$, a user can simulate data $\boldsymbol{D}^* \sim \pi(\boldsymbol{D}|\boldsymbol{\theta}^*)$ and compare this with the experimental data $\boldsymbol{D}$. If the simulated data is within a given threshold distance ε with distance measure $d(\cdot, \cdot)$ from the experimental data $\boldsymbol{D}$, then the sample parameter set $(\boldsymbol{\theta}^*, \boldsymbol{D}^*)$ is accepted. Otherwise, the parameter set is rejected, and this continues until N parameter sets are accepted. Consequently, the method is presented in Algorithm 2.2.

Algorithm 2.2: ABC rejection (Toni *et al.*, 2009)

Let $n = 0$, choose a posterior sample size N, an acceptance threshold ε, and distance measure $d(\cdot, \cdot)$.

while $n < N$ **do**

   Sample $\boldsymbol{\theta}^*$ from $\pi(\boldsymbol{\theta})$.

   Simulate a data set $\boldsymbol{D}^* \sim \pi(\boldsymbol{D}|\boldsymbol{\theta}^*)$.

   If $d(\boldsymbol{D}, \boldsymbol{D}^*) \leq \varepsilon$, accept $\boldsymbol{\theta}^*$ and set $n = n + 1$.

end

2.4.2 Approximate Bayesian computation- sequential Monte Carlo

A significant drawback of the ABC rejection algorithm is its computational inefficiency. Situations arise where the parameter space being sampled from is large, either due to numerous parameters in the model, a wide interval for prior distributions, or a non-informative prior is used. Consequently, to sample the entire parameter space, many simulations are required, and as such, the probability of acceptance is low. To overcome this, Toni *et al.* (2009) developed a new method ABC- Sequential Monte Carlo (SMC). Toni *et al.* (2009)'s approach is iterative, where the user implements ABC rejection multiple times to improve computational efficiency and speed of convergence.

Keeping the notation introduced in Section 2.4.1, I introduce a decreasing sequence of distance threshold values $\varepsilon_1 > \varepsilon_2 > \cdots > \varepsilon_T$ where T is the number of iterations of ABC to run. Each iteration will generate an accepted sample set of size N, where each of the N elements $\boldsymbol{\theta}^*$ is referred to as a particle using the terminology defined by Toni *et al.* (2009). A whole sample of particles size N is referred to as a population, where the method iterates until there are T populations of accepted particles, where iteration T comprises the final posterior distribution, which satisfies the distance measure such that $d(\boldsymbol{D}, \boldsymbol{D}^*) \leq \varepsilon_T$.

For the first iteration, a rejection ABC is implemented where parameters are sampled from prior distributions $\pi(\boldsymbol{\theta})$ and are accepted if they satisfy the

distance measure $d(\boldsymbol{D}, \boldsymbol{D}^*) \leq \varepsilon_1$. Each particle from the posterior distribution of the first iteration is assigned equal weight $w_1^{(n)} = 1/N$ for $n = 1, \ldots, N$. Subsequent iterations of the method will draw particles $\boldsymbol{\theta}^*$ by sampling the posterior distributions of the prior iteration with weights $\boldsymbol{w}_{t-1}$ where t is the index of the current iteration. Accepted parameters from the previous iteration are perturbed using a perturbation kernel $K_t(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$. These are then used to simulate the model and compare it to observed data. Once again, for each iteration, a distance measure with a distance threshold is used to determine if these perturbed parameters are accepted if $d(\boldsymbol{D}, \boldsymbol{D}^*) \leq \varepsilon_t$. Each particle in the population is assigned a weight determined by: the prior densities, the weights from the previous iteration, and the choice of perturbation kernel. The procedure continues for T iterations where N accepted parameter sets are achieved. Compared to rejection ABC, this method allows for additional reduction in the parameter space between successive iterations since the posterior distributions of the previous iteration inform sampling. The choice of perturbation kernel is user-defined, and examples include the component-wise uniform kernel or Gaussian kernel (Toni *et al.*, 2009). The algorithm for this method is shown in Algorithm 2.3.

2.4.3 Bayesian model selection

Bayesian model selection allows comparison between two or more mathematical models to determine which has the highest probability to describe the observed data $\boldsymbol{D}$. Assume that a user has observed data that describes a biological mechanism and proposed two potential mathematical models, $\mathcal{M}_1$ and $\mathcal{M}_2$ that describe that mechanism. Bayesian model selection requires determination of the *Bayes factor*, which is defined as,

$$B_{12} = \frac{\pi(\mathcal{M}_1|\boldsymbol{D})/\pi(\mathcal{M}_2|\boldsymbol{D})}{\pi(\mathcal{M}_1)/\pi(\mathcal{M}_2)}.$$

Here $\pi(\mathcal{M}_i)$ is the prior distribution of model $\mathcal{M}_i, i \in \{1, 2\}$ and $\pi(\mathcal{M}_1|\boldsymbol{D})$ is the posterior distribution for model $\mathcal{M}_i$. Assuming uniform prior distributions, the Bayes factor reduces to

$$B_{12} = \frac{\pi(\mathcal{M}_1|\boldsymbol{D})}{\pi(\mathcal{M}_2|\boldsymbol{D})},$$

which provides the evidence of selecting $\mathcal{M}_1$ over $\mathcal{M}_2$. From ABC rejection an estimate $\hat{B}_{12}$ can be obtained for B_{12}. The model selection algorithm is summarised in Algorithm 2.4.

Algorithm 2.3: ABC SMC (Toni *et al.*, 2009)

Choose a sequence of threshold values $\varepsilon_1 \geq \varepsilon_2 \cdots \geq \varepsilon_T \geq 0$, posterior sample size N, distance measure $d(\cdot, \cdot)$. Set $t = 1$ for the population indicator and particle indicator $n = 0$.

while $t < T$ **do**

$\quad$ Reset $n = 0$

$\quad$ **while** $n < N$ **do**

$\quad\quad$ If $t = 1$, then sample from the prior distribution (iteration 1). Else sample $\boldsymbol{\theta}^*$ from a previous population $\{\boldsymbol{\theta}_{t-1}^{(n)}\}$ with the associated weights $\boldsymbol{w}_{t-1}$ and perturb these to obtain $\boldsymbol{\theta}^{**}$ $K_t(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$, with $K_t(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$ being the perturbation kernel. If $\pi(\boldsymbol{\theta}^{**}) = 0$, then re-sample $\boldsymbol{\theta}^*$ until $\pi(\boldsymbol{\theta}^{**}) \neq 0$.

$\quad\quad$ Simulate the data $\boldsymbol{D}^*$ from $\pi\boldsymbol{D}|\theta^{**})$.

$\quad\quad$ If $d(\boldsymbol{D}, \boldsymbol{D}^*) \leq \varepsilon_t$, set $\boldsymbol{\theta}_t^{(n)} = \boldsymbol{\theta}^{**}$ and set $n = n + 1$. Calculate the weight for sample $\boldsymbol{\theta}_t^{(n)}$ as,

$$w_t^{(n)}\boldsymbol{\theta}_t^{(n)}) = \begin{cases} 1 & \text{if } t = 1 \\ \frac{\pi(\boldsymbol{\theta}_t^{(n)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\boldsymbol{\theta}_t^{(n)}|\boldsymbol{\theta}_{t-1}^{(j)})} & \text{if } t > 1 \end{cases}$$

$\quad$ **end**

$\quad$ Normalise the weights, and set $t = t + 1$.

end

Algorithm 2.4: ABC rejection model selection (Toni *et al.*, 2009)

Choose a posterior samples size N, an acceptance threshold ε, and distance measure $d(\cdot, \cdot)$. Set model indicator $i = 1$ and total simulations per model $r_1 = r_2 = 0$.

while $i \leq 2$ **do**

    Let $n = 0$

    while $n < N$ **do**

        Sample $\boldsymbol{\theta}_i^*$ from $\pi(\boldsymbol{\theta_i})$.

        Simulate a data set $\boldsymbol{D_i^*} \sim \pi(\boldsymbol{D}|\boldsymbol{\theta}_i^*)$.

        If $d(\boldsymbol{D}, \boldsymbol{D}_i^*) \leq \varepsilon$, accept $\boldsymbol{\theta}_i^*$ and set $r_i = r_i + 1$

        Set $n = n + 1$.

    end

    Set $i = i + 1$

end

Calculate $p_i = \frac{r_i}{N}$ for $i \in \{1, 2\}$

Calculate $\hat{B}_{12} = \frac{r_1}{r_2}$

2.5 Sensitivity analysis and identifiability

In the preceding section, methods of Bayesian inference were introduced that allow the parametrisation of mathematical models when data is available for comparison. However, since a model may introduce a large number of parameters, of which only a handful may have been quantified through biological experimentation, it is vital to understand how variation in parameter values can affect the output of a mathematical model. Careful consideration of the parameter values such that these models continue to represent the biological system under investigation is essential. Sensitivity analysis can be used for this purpose and provides the user with a numerical quantification of a given parameter's sensitivities (Sobol, 1993). Together with parameter sensitivity, it is also essential to understand if unique values to model unknown parameters can be obtained for a given model and the

data collected via experimentation. Structural identifiability provides a method to understand this.

2.5.1 Sobol global sensitivity analysis

The models developed within this thesis involve multiple parameters to represent the biological mechanisms under investigation. Since many parameters require calibration, it is important to identify which parameters have the greatest effect on the overall model output. Global sensitivity analysis allows me to evaluate the results of simultaneous changes in parameter values (Zhang *et al.*, 25th February 2015). Consider the vector of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ such that the model output is described as $Y = g(\boldsymbol{\theta})$. Sobol's approach shall be used to determine global sensitivities (Sobol, 1993). Each parameter θ_i can be considered as a random variable with an associated range. Since Y is a function of these variables, it is also a random variable with variance $\text{Var}(Y)$. I am interested in what would happen if a known value is changed, *i.e.* consider the conditional variance $\text{Var}(Y|\theta_i = \theta_i^*)$. However, since the value of θ_i^* is not known, I instead consider the average conditional variance $\mathbb{E}[\text{Var}(\theta_i = \theta_i^*)]$ where the expectation is with respect to θ_i and the variance is taken over all remaining parameters θ_j, $j \neq i$. The law of total probability gives me the following expression,

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|\theta_i)] + \text{Var}(\mathbb{E}[Y|\theta_i]),$$

from which the first-order Sobol index for parameter θ_i is defined as,

$$S_i = \frac{\text{Var}(\mathbb{E}[Y|\theta_i])}{\text{Var}(Y)}. \tag{2.10}$$

Furthermore, one can also investigate the result of multiple fixed parameter values. Let $\text{Var}(\mathbb{E}[Y|\boldsymbol{\theta}_{-\theta_j}])$ be the expected reduction in the variance by fixing all parameters except θ_j. Then the total effect of parameter θ_i can be defined as,

$$S_{Ti} = \frac{\mathbb{E}[\text{Var}(Y|\boldsymbol{\theta}_{-\theta_i})]}{\text{Var}(Y)} = 1 - \frac{\text{Var}(\mathbb{E}[Y|\boldsymbol{\theta}_{-\theta_i}])}{\text{Var}(Y)}. \tag{2.11}$$

Sobol (1993) showed that these conditional variances are obtained if a function g, integrable over $[0,1]^n$, may be decomposed into

$$g(\boldsymbol{\theta}) = g_o \sum_{i=1}^{n} g_i(\theta_i) + \sum_{1 \leq i < j \leq n} g_{ij}(\theta_i, \theta_j) + \cdots + g_{1,2,\ldots,n}(\theta_1, \theta_2, \ldots, \theta_n). \tag{2.12}$$

It is assumed that terms in (2.12) have zero mean, as such squaring both sides and integrating yields

$$\int g^2(\boldsymbol{\theta})d\boldsymbol{\theta} - g_0^2 = \text{Var}(Y) = \sum_{i=1}^{n} V_i + \sum_{1 \leq i < j \leq n} V_{i,j} + \cdots + V_{1,2,\ldots,n}, \qquad (2.13)$$

where $g_0 = \mathbb{E}[Y]$ and

$$V_i = \text{Var}(\mathbb{E}[Y|\theta_i]) \,,$$
$$V_{i,j} = \text{Var}(\mathbb{E}[Y|\theta_i, \theta_j]) - \text{Var}(\mathbb{E}[Y|\theta_i]) - \text{Var}(\mathbb{E}[Y|\theta_j]) \,,$$
$$\vdots$$
$$V_{1,\ldots,n} = \text{Var}(\mathbb{E}[Y|\theta_1, \ldots, \theta_n]) + \sum_{i=1}^{n} \text{Var}(\mathbb{E}[Y|\theta_i]) + \sum_{1 \leq i < j \leq n} \text{Var}(\mathbb{E}(Y|\theta_i, \boldsymbol{\theta}_{-\theta_j})) + \ldots.$$

The first n terms of Equation (2.13) give the first-order Sobol indices as defined in Equation (2.10). Other terms in the expansion can be used to calculate higher-order interactions. The total-order Sobol index S_{Ti} defined in Equation (2.11) for parameter θ_i is the sum of all contributions to the model variance,

$$S_{Ti} = S_1 + S_{ij} + S_{ik} + \cdots + S_{ijk} + \cdots + S_{i,\ldots,n}.$$

Generally, S_i and S_{Ti} can be computed for all parameters; this will describe the model in terms of the global sensitivity analysis properties. A larger sensitivity index indicates a greater parameter influence on the model output (Sobol, 1993; Zhang *et al.*, 25th February 2015).

2.5.2 Structural identifiability

Structural identifiability allows a user to assess if, given a set of data, uniquely defined parameter values within a given model can be obtained from biological experimentation. The following definition of structural identifiability comes from Chis *et al.* (2011).

Definition 2.5.1. Suppose there is a system of ODEs $\dot{\boldsymbol{x}}$, and a set of experimentally observed quantities $\boldsymbol{y}$. These depend on the system's state $\boldsymbol{x}$ and a vector of

unknown parameters $\boldsymbol{\theta}$. This system is denoted by $\sum(\boldsymbol{\theta})$ to highlight dependence on the parameter vector $\boldsymbol{\theta}$. Assume a biological system is described by:

$$\sum(\boldsymbol{\theta}) : \begin{cases} \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) + \sum_{j=1}^{n_u} \boldsymbol{g}_j(\boldsymbol{x}, \boldsymbol{\theta})\boldsymbol{u}_j \ , \\ \boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{\theta}) \ , \boldsymbol{x}(t_0) = \boldsymbol{x}_0(\boldsymbol{\theta}), \end{cases}$$

where $\boldsymbol{x}^T = (x_1, \ldots, x_{n_x}) \in \boldsymbol{M} \subset \mathbb{R}^{n_x}$ is the state variable, with $\boldsymbol{M}$ a subset of $\mathbb{R}^{n_x}$ containing the initial state, $\boldsymbol{u}^T = (u_1, \ldots, u_{n_u}) \in \mathbb{R}^{n_u}$ a n_u-dimensional input vector with $u_1, \ldots, u_{n_u}$ smooth functions, and $\boldsymbol{y}^T = (y_1, \ldots, y_{n_y}) \in \mathbb{R}^{n_u}$ in the n_y-dimensional output of experimentally observed quantities. The vector of unknown parameters is denoted by $\boldsymbol{\theta}^T = (\theta_1, \ldots, \theta_{n_p}) \in \boldsymbol{\Theta}$, and in general is assumed to belong to an open and connected subset of $\mathbb{R}^{n_p}$. The entries of $\boldsymbol{f}$, $\boldsymbol{g}^T = (g_1, \ldots, g_{n_u})$ and $\boldsymbol{h}$ are analytical functions of their arguments. These functions and the initial conditions may depend on the parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Assuming perfect experimental data free of noise and continuous in time:

- A parameter θ_i, $i = 1, \ldots, n_p$ is structurally globally (or uniquely) identifiable if for almost any $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$,

$$\sum(\boldsymbol{\theta}) = \sum(\boldsymbol{\theta}^*) \implies \theta_i = \theta_i^*,$$

- A parameter θ_i, $i = 1, \ldots, n_p$ is structurally locally identifiable if, for almost and $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, there exists a neighbourhood $\boldsymbol{V}(\boldsymbol{\theta}^*)$ such that

$$\boldsymbol{\theta} \in \boldsymbol{V}(\boldsymbol{\theta}^*) \qquad \text{and} \qquad \sum(\boldsymbol{\theta}) = \sum(\boldsymbol{\theta}^*) \implies \theta_i = \theta_i^*,$$

- A parameter θ_i, $i = 1, \ldots, n_p$ is structurally non-identifiable if for almost and $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, there exists **no** neighbourhood $\boldsymbol{V}(\boldsymbol{\theta}^*)$ such that

$$\boldsymbol{\theta} \in \boldsymbol{V}(\boldsymbol{\theta}^*) \qquad \text{and} \qquad \sum(\boldsymbol{\theta}) = \sum(\boldsymbol{\theta}^*) \implies \theta_i = \theta_i^*.$$

Structural identifiabilty can be assessed using the SIAN toolbox (Hong *et al.*, 2020) a Maple based program or using Castro & de Boer (2020) method of testing identifiabilty using scaling.

2. MATHEMATICAL BACKGROUND

Chapter 3

Statistical analysis of patient data from viral infections

Throughout this thesis, multiple biological hypotheses will guide the formulation of several mathematical models. In studies investigating biological phenomena, scientists often gather large amounts of data to either support or disprove their theories and provide conclusions to their investigations. These findings subsequently inform model formulation by mathematicians to investigate additional biological questions. In this chapter, two studies of viral outbreaks will be investigated. The first study is data provided by the UK Health Security Agency (UKHSA), formally known as Public Health England (PHE), on patient data from survivors of the 2014 West Africa Ebola outbreak. The second data set is from adult immune response data to the SARS-CoV-2 virus during the early stages of the 2020 SARS-CoV-2 pandemic, obtained from employees at St Jude Children's Research Hospital.

3.1 Ebola virus patient data

In early 2014 the World Health Organisation (WHO) reported an outbreak of EBOV in Guinea, where the index patient would later be identified as an 18-month-old boy who lived in rural Guinea (Brolin Ribacke *et al.*, 2016; Carias *et al.*, 2019). This initial case soon spread to other village members and communities, where subsequent infections would be identified in neighbouring Sierra Leone and

3. STATISTICAL ANALYSIS OF PATIENT DATA FROM VIRAL INFECTIONS

Liberia. Cases continued to rise until international concern about the further spread of EBOV led to additional intervention. This led to a gradual decline in EBOV in late 2015. After the epidemic had ended, there were 28,652 suspected infections, and 11,326 reported deaths (Bell, 2016). The renewed global concern, along with the scale of the epidemic, led multiple research bodies to investigate the kinetics of EBOV infections, along with long-term immunity in survivors of EBOV infection (Thom *et al.*, 2021; Timothy *et al.*, 2019).

Public Health England began a collaborative study with multiple research institutions to examine the longitudinal immune responses to EBOV of known survivors and contacts. A study of this type would allow the scientific community to understand better the antibody and T-cell responses to further exposure over time. It would also allow investigation of any potential waning immunity, which for many viruses such as SARS-CoV-2 has been shown to enable subsequent reinfection (Stich *et al.*, 2022). This study had two key responses to be examined. The first was regarding antibody responses and how these could potentially vary with time. MacNeil *et al.* (2011) showed within a 120-day period that immunoglobulin (Ig) G levels remained relatively high, while IgM levels declined with time, while Khurana *et al.* (2020) demonstrated a robust response over a 360-day period. The question remains whether, over the years, these responses remain robust for both neutralising antibodies and IgG to the whole virus. The second question is understanding IFN-γ responses and how these change over time. IFN-γ has a half life of 30 minutes and 4.5 hours after intravenous and intramuscular injection, respectively (Foon *et al.*, 1985). IFN-γ is an essential cytokine in both the innate and adaptive immune responses, where it can amplify antigen presentation, maintain CD4 and CD8 T-cell activity and help mediate anti-microbial functions (Kak *et al.*, 2018). Rhein *et al.* (2015) has shown IFN-γ to reduce the replicative ability of EBOV. However, IFN-γ in the presence of several other cytokines such as interleukin (IL)-2, IL-10 and tumour necrosis factor (TNF)-α has found to be a marker of a fatal outcome (Villinger *et al.*, 1999). Since IFN-γ is produced to maintain CD4 and CD8 activity, quantifying its production can be used as a marker for T-cell activity.

In the following work, I investigate the longitudinal responses of a cohort of individuals from the 2014 West Africa Ebola outbreak. Both antibody and IFN-γ

responses will be analysed considering the individual's infection status: whether they were infected, a close contact, or negative control. Furthermore, human leukocyte antigens will be investigated by adding a genetic factor and how these correlate with IFN-γ responses and T-cell activity. This may indicate a particular HLA gene that provides greater protection than others.

3.1.1 Data summary

Participants for this study were recruited from various treatment centres in Guinea, mainly from Gueckedou, the epicentre of the outbreak and Coyah, a village located roughly 385 miles away (Thom *et al.*, 2021; Timothy *et al.*, 2019). Blood samples were collected from survivors 3-14 months post-infection, where survivor status was confirmed either from treatment centre records or through the presentation of a valid survivor's certificate. Any additional pertinent data was gathered from patients, and further data was collected from contact cases, including blood samples where possible. Contact cases are individuals who lived with or cared for infected individuals. Where possible longitudinal samples were taken one and two years post-first sample collection, a subset of these individuals have year three and four samples.

For each participant, blood samples were collected and used for the following set of experiments: (i) Marburg ELISA was undertaken to quantify IgG antibody responses using whole inactivated EBOV measured in arbitrary units, (ii) neutralisation assays to quantify the levels of neutralising antibodies which is an indicator of the amount of EBOV that is being stopped from entering the host's cells, measured in geometric mean neutralisation titres (GMT), (iii) ELISpot assays to examine the amount of IFN-γ produced by an individual in response to the glycoprotein in its entirety or individual sections, measured in spot forming units (SFU), to be used as a marker of T-cell activity. In addition to these experiments for 2015, an IgG ELISA was used to quantify the antigen expression to the glycoprotein explicitly, while for 2016-2018, the experiments were for IgG ELISA using World Health Organisation standards. Where possible, repeat measurements were taken from the same individuals yearly; however, many individuals missed one or more years worth of samples due to unknown circumstances. The Marburg

3. STATISTICAL ANALYSIS OF PATIENT DATA FROM VIRAL INFECTIONS
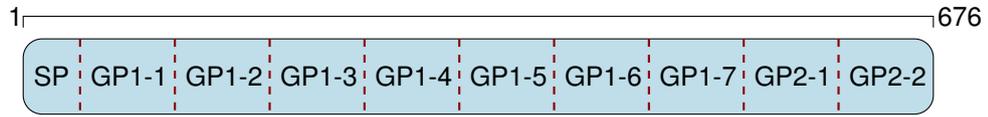
Figure 3.1: Schematic representation of the EBOV glycoprotein highlighting notable regions used in ELISpot analysis. SP stands for signal protein.

ELISA here gets its name from Marburg, Germany, where the experiment was performed. Not an ELISA in response to the Marburg virus. The GP was broken down into 10 different sections as illustrated in Figure 3.1, with an overlap of 11 amino acids where SP is 46 amino acids long and the others have an average length of 70 amino acids. Further details of the sub-units can be found in Tipton *et al.* (2021).

The PHE longitudinal EBOV response cohort consists of 299 individuals recruited from two regions in Guinea: Coyah and Gueckedou, which are 385 miles apart. 170 (56.9%) individuals were recruited to participate in the study from Coyah, while 123 (41.1%) originated from Gueckedou, with the remaining 6 (2%) participants having an unspecified region. One of these individuals was from a Western country that visited Guinea during the outbreak. As such, this individual will be excluded from all future analyses, leaving 298 participants to be examined. Within the cohort 146 (48.9%) are female, 138 (46.3%) are male, while the remaining 14 (4.8%) participants did not provide their gender. Each individual is also given an infection status: a survivor (confirmed infected), negative control, and contact, as described in the introduction to this section. In total there are 117 (59.4%) contacts, 25 (8.4%) negative controls and the remaining 152 (51%) are confirmed survivors with 4 individuals having an undefined status. The average age of participants at the start of the study was 34.58 years (SD 11.85); however, 164 participants did not provide their age for part of the study. Where possible, information regarding dates of symptom onset, treatment centre admission, first positive result, and discharge was collected with other medical

information regarding their HIV status and hepatitis B/C status. However, for many individuals, pieces of this information were missing.

3.1.2 Analysis of EBOV response data

Many studies have investigated differences in immune responses to EBOV. Baize et al. (1999) examined the differences in humoral responses for fatalities and survivors, where survivors expressed higher levels of antibody isotypes M and G than fatalities for circulating EBOV. Fatalities were also shown to express higher levels of IFN-γ early in the infection, while survivors have delayed IFN-γ production. These results were further supported by Colavita et al. (2019). In the study presented here, there are no fatalities included and no data on disease severity from ay source. Still, the differences in immune responses between survivors, close contacts and negative controls were examined for differences in humoral responses for IgG and neutralising antibodies. Analysing the response data for close contacts of survivors may indicate potential cases of asymptomatic individuals. Furthermore, one can explore the levels of IFN-γ produced that will act as an indicator of the host's T-cell responses (Thom et al., 2021). During the 2014 West Africa outbreak, EBOV spread from Guinea to Liberia and Sierra Leone. This spread amongst countries leads to two distinct lineages as noted by Carroll et al. (2015) and further supported by Simon-Loriere et al. (2015), and the possibility that individuals in different regions may be exposed to distinct circulating strains and hence exhibit different immune responses. Since our cohort includes two regions within Guinea, it would be interesting to examine whether there are any differences in immune responses. Simon-Loriere et al. (2015) noted that the original infectious strain originated in Guinea in Gueckedou, which spread around the country and into Sierra Leone and Liberia. However, Carroll et al. (2015) showed that different lineages of the virus spread from Liberia back into Guinea, so it is possible that different regions were exposed to separate virus lineages. Therefore, it is worth examining if the regional difference may imply different circulating strains. In the following work, infection status and regional differences will be investigated.

3. STATISTICAL ANALYSIS OF PATIENT DATA FROM VIRAL INFECTIONS

Infection Status	Year	Mardia Skew (p-value)	Mardia Kurtosis (p-value)	MVN
Contact	2015	2267.32 (≈ 0)	15.39 (≈ 0)	No
	2016	2929.99 (≈ 0)	23.18 (≈ 0)	No
	2017	1383.76 (≈ 0)	2.71 (6.7×10^{-3})	No
Survivor	2015	-1.5×10^{-6} (1)	936.90 (≈ 0)	No
	2016	4805.81 (≈ 0)	41.52 (≈ 0)	No
	2017	6333.81 (≈ 0)	61.49 (≈ 0)	No
	2018	916.58 (7×10^{-3})	2.73 (6×10^{-3})	No

Table 3.1: Results from the Mardia multi-variate normal test for each year and group based on infection status. Test performed at the 5% significance level.

Since the analysis here will examine differences on a multi-variate level, assessing whether the data provided follows a multi-variate normal distribution is essential. If it does, methods such as Hotelling's T-statistic can be used along with several additional techniques summarised in Chatfield (2018). However, if the data is not multi-variate normal, non-parametric techniques such as the ε-test introduced in Section 2.2.3 will have to be used. To do this, the multi-variate Mardia test described in Section 2.2.1 is applied to each year of data for the infection status groups "survivor" and "contact". Many of the "negative controls" were missing one or more experimental samples, so they could not be included in the normality test. The test only applies to IAG ELISA, Marburg ELISA, IgG ELISA and ELISpot assay results. These are all continuous quantitative measurements. The results from this test can be found in Table 3.1, which shows that none of the data for any group by infection status and year is multi-variate normally distributed and as such testing procedures that require an assumption of multi-variate normality cannot be used with this data. For "contacts" in 2018, there is insufficient data to perform the Mardia test using the R function, which requires $n > 6$ observations. Meanwhile, the results of the Mardia test applied to the cohorts are summarised in Table 3.2. Looking at the region also shows no MVN; for 2018, there were no samples collected from Coyah, so it was impossible to compare any results in 2018.

Region	Year	Mardia Skew (p-value)	Mardia Kurtosis (p-value)	MVN
Coyah	2015	5438.32 (≈ 0)	55.23 (≈ 0)	No
	2016	3055.54 (≈ 0)	21.58 (≈ 0)	No
	2017	1434.16 (≈ 0)	8.67 (6.7×10^{-3})	No
Gueckedou	2015	1415 (≈ 0)	4.86 (1.15×10^{-6})	No
	2016	3324.75 (≈ 0)	30.96 (≈ 0)	No
	2017	2407.56 (≈ 0)	17.24 (≈ 0)	No
	2018	916.58 (7×10^{-3})	2.73 (6×10^{-3})	No

Table 3.2: Results from the Mardia multi-variate normal test for each year and group based on region. Test performed at the 5% significance level.

Examining regional differences in immune responses

The samples within this cohort were collected from two regions Gueckedou and Coyah. Since Gueckedou was the origin of the epidemic, there may be differences in responses in Coyah due to mutations within the virus before it arrived in the city. When examining the antibody responses for neutralising antibodies and isotype G, there is a time dependency on immune responses. Figure 3.2 shows the responses for the three experiments for each year, which shows that, for 2015, individuals from Gueckedou exhibit a higher antibody response with less variation, compared to those individuals from Coyah. However, in 2016 this difference shifted with those from Coyah exhibiting higher levels of antibodies, and by 2017 there was no significant difference in antibody responses. One explanation for this difference in responses could be due to the time each individual was infected. Since most individuals in this study lack a symptom onset date, it is impossible to check whether this difference is due to time. One could hypothesise that the differences in 2015 are explained by individuals in Gueckedou being infected earlier and therefore having more time to produce effective IgG antibodies, while individuals in Coyah may not have had as long to mount an effective IgG antibody response. By 2016 individuals in Gueckedou have reached a steady state in antibody levels after infection, while individuals in Coyah still have slightly higher levels of antibodies due to a more recent infection. As a result, by 2017, both regions have entered a basal steady-state level of antibodies circulating in the blood, and hence the

3. STATISTICAL ANALYSIS OF PATIENT DATA FROM VIRAL INFECTIONS

differences become statistically insignificant. This hypothesis would be consistent with Rimoin *et al.* (2018), who found detectable levels of neutralising antibodies 40 years after infection with EBOV.

As mentioned by Carroll *et al.* (2015), responses may vary depending on the region due to the different circulating lineages of the virus. Since there are several years of data, it is possible to examine if the IFN-γ expression determined by ELISpot is higher in one region than another for several years. Since this experiment uses the glycoprotein (GP), which may express different genome sequences due to mutation, T-cell responses may differ. Figure 3.3 illustrates the results of principal component analysis for 2015 applied to ELISpot data for IFN-γ where the GP is split into its various subunits to test whether a particular region may elicit a different response to another. As seen from Figure 3.3, responses from individuals in Gueckedou are more heterogeneous than those from Coyah. Figure 3.4 illustrates the correlation circle for the results from PCA for the GP subunits under consideration. The angle between variables represents the correlation; the smaller an angle, the stronger the correlation, while negatively correlated variables will have an angle greater than 90 degrees between them. Furthermore, this figure also illustrates the quality of the representation in the first two principal components. The distance of the variable from the origin shows the quality of representation. From this information, it can be seen that except for GP1-4, the first two principal components represents all GP subunits equally. Furthermore, it also shows that subunits SP, GP1-1, GP1-2 and GP1-3 are highly correlated with one another and share a minimal correlation with the remaining subunits, which all share a high correlation. Altogether, this information suggests that responses between these groups in 2015 differ and using the ε-test introduced in Section 2.2.3 supports this with a p-value $\approx 5 \times 10^{-3}$. Figure 3.5 presents a uni-variate view of IFN-γ responses to each GP subunit for 2015. This figure shows that the responses for individuals from Gueckedou are significantly higher than those from Coyah. However, as hypothesised with the antibody responses in Figure 3.2, this could be a result of time from infection. Taking the information presented in Figures 3.2 and 3.5 suggests that at the time of sampling, individuals from Coyah had more recent infections and had not yet had time to develop an effective immune response and T-cell response to EBOV infections. This is

46
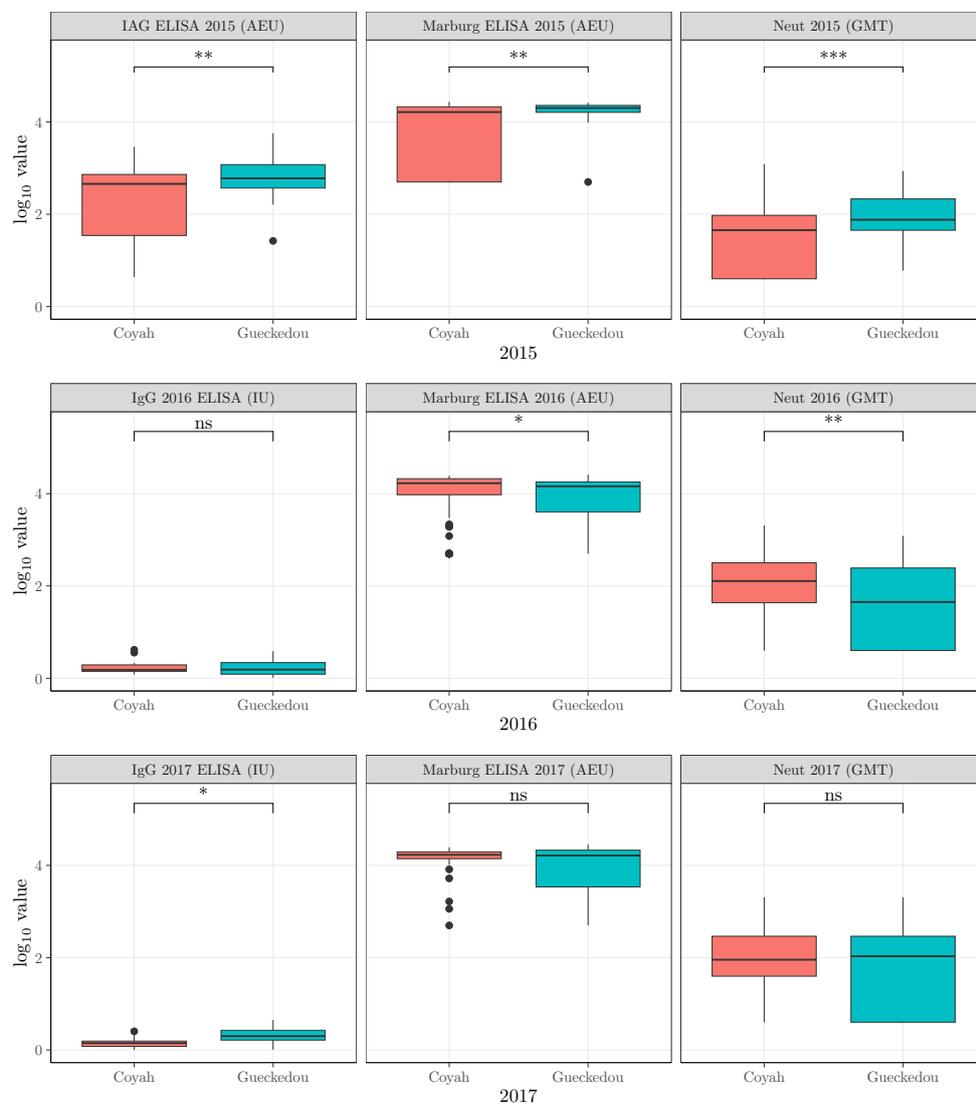
Figure 3.2: Antibody responses assessed by ELISA, Marburg ELISA, and neutralisation assays for each year where both regions in the cohort were available for study. Statistical significance was determined by the Wilcoxon–Mann–Whitney test with Bonferroni adjustment (ns, not significant; $*p < 5 \times 10^{-2}$ $**p < 10^{-2}$; $***p < 10^{-3}$; $****p < 10^{-4}$). Arbitrary ELISA units (AEU), geometric mean neutralisation titre (GMT), international units (IU).
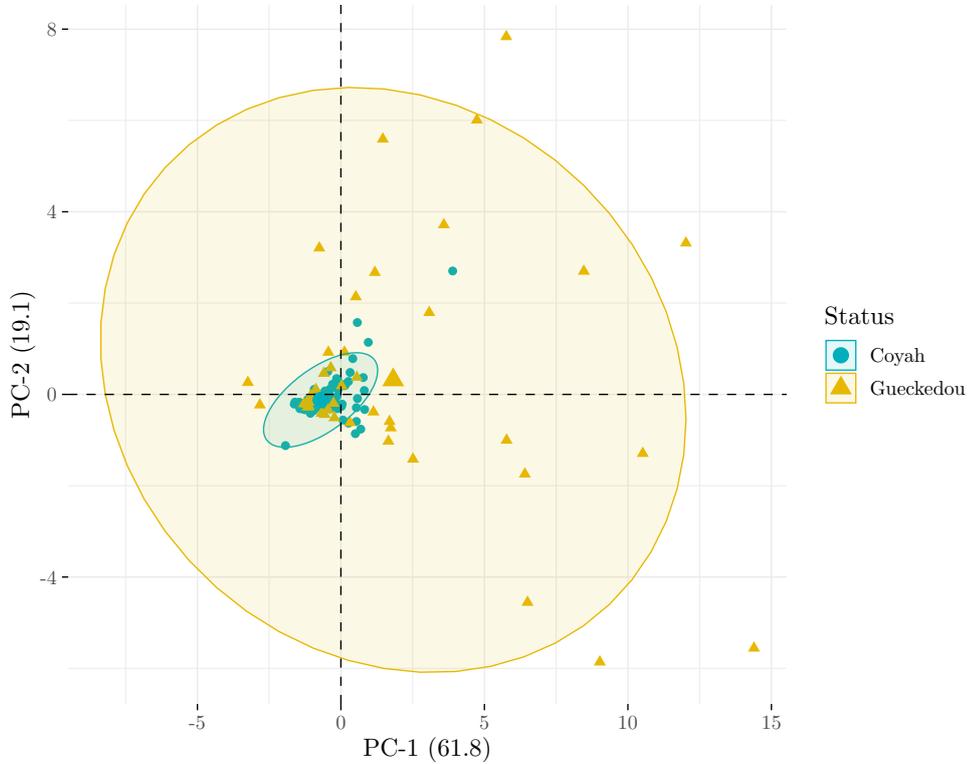
Figure 3.3: PCA for Coyah against Gueckedou, considering IFN-γ ELISpot result in response to EBOV GP broken down by GP subunit. The first PC accounts for 61.8% of the total variance, while the second accounts for 19.1%. Ellipses represents a 95% confidence region.

further supported by the information reported by Simon-Loriere *et al.* (2015) that the incident cases of EBOV were first reported in Gueckedou. Performing the ε-test on the data for 2016 and 2017 for the two regions returns p-values $< 10^{-2}$ for both years, indicating that there is still a difference in responses, even though by this time responses should be more similar as seen in the antibody data. When examining the data more closely for 2016 and 2017, one GP-subunit remains significant for both years and follows the same trend. GP1-6 in individuals from Coyah resulted in higher levels of IFN-γ in both 2016 and 2017, suggesting that the strain circulating in Coyah at the time of infection shared a similar sequence to the Kitwit variant used for all experiments performed in this study,
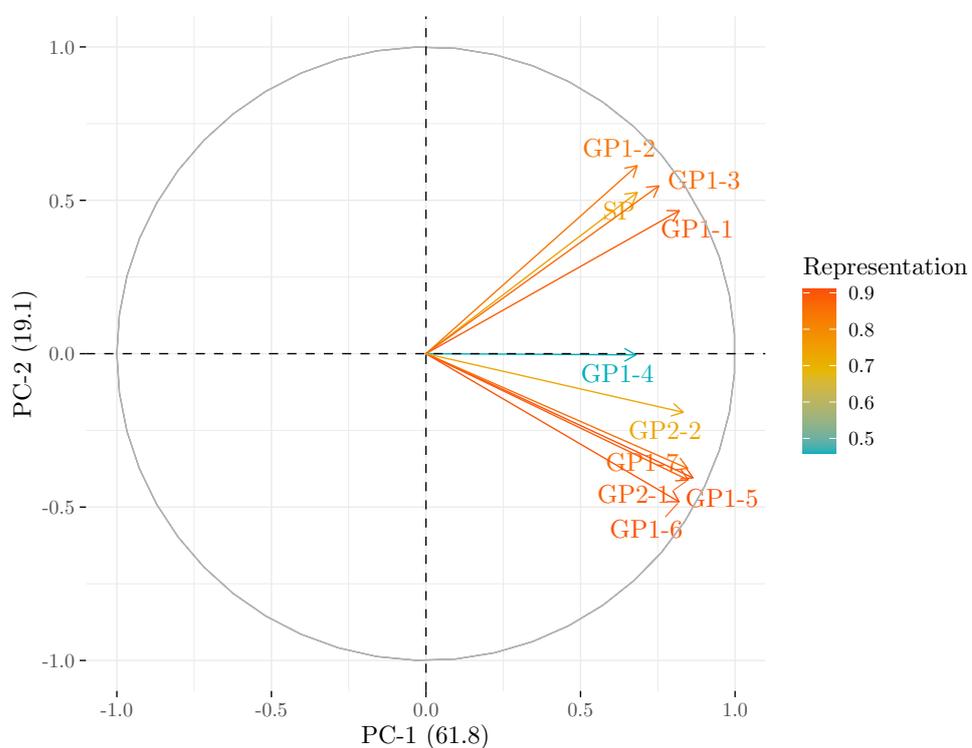
Figure 3.4: Correlation circle of each GP subunit ELISpot for IFN-γ. The angle between variables represents the correlation of variables with the length of the line from the origin, representing the quality of representation. The PCs in this figure match those presented in Figure 3.3.

49
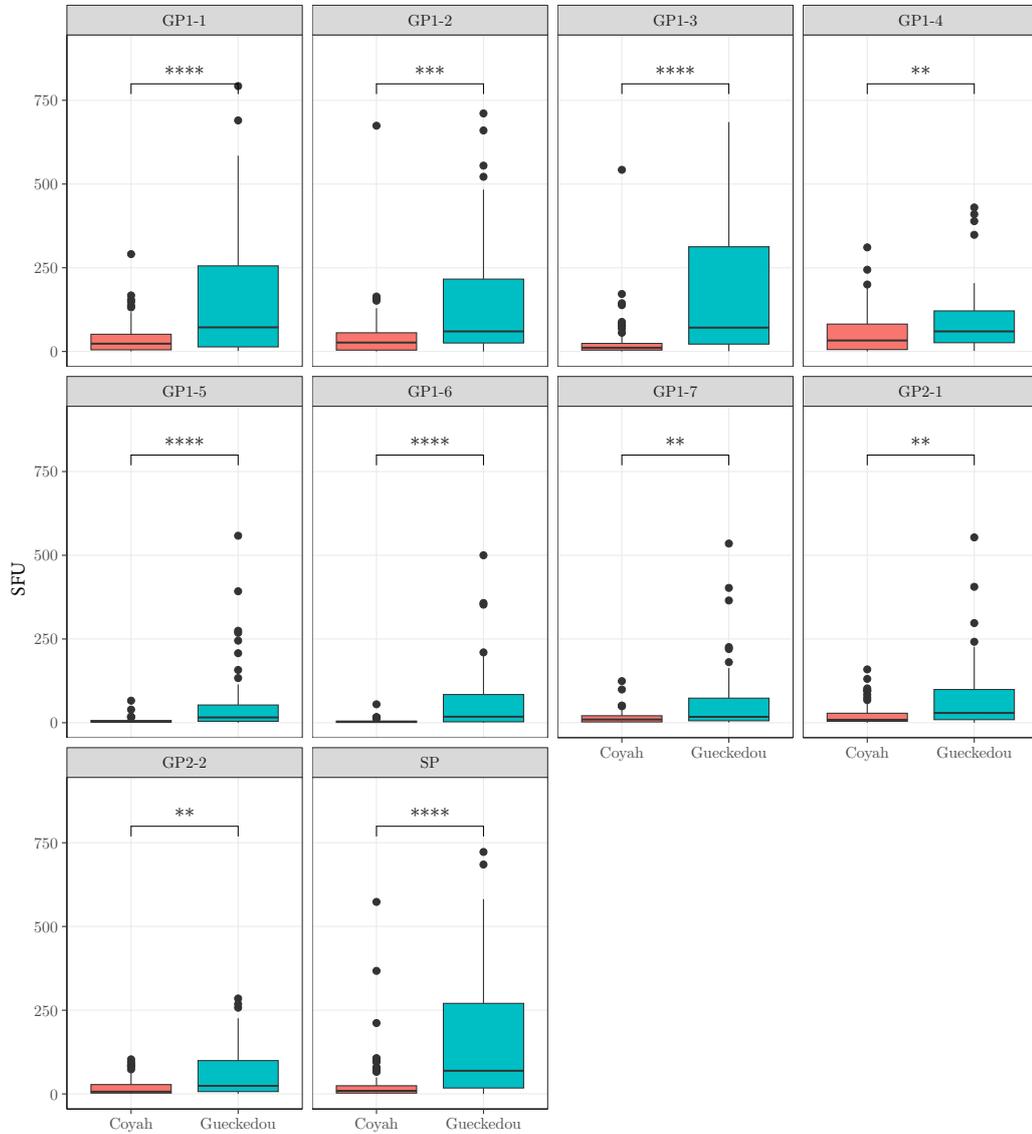
Figure 3.5: Box plots for each GP subunit for regions Coyah ($n = 116$) and Gueckedou ($n = 45$) from ELISpot experiments to quantify IFN-γ production. Statistical significance was determined by the Wilcoxon–Mann–Whitney test with Bonferroni adjustment (ns, not significant; *$p < 5 \times 10^{-2}$ **$p < 10^{-2}$; ***$p < 10^{-3}$; ****$p < 10^{-4}$). Spot forming units (SFU)
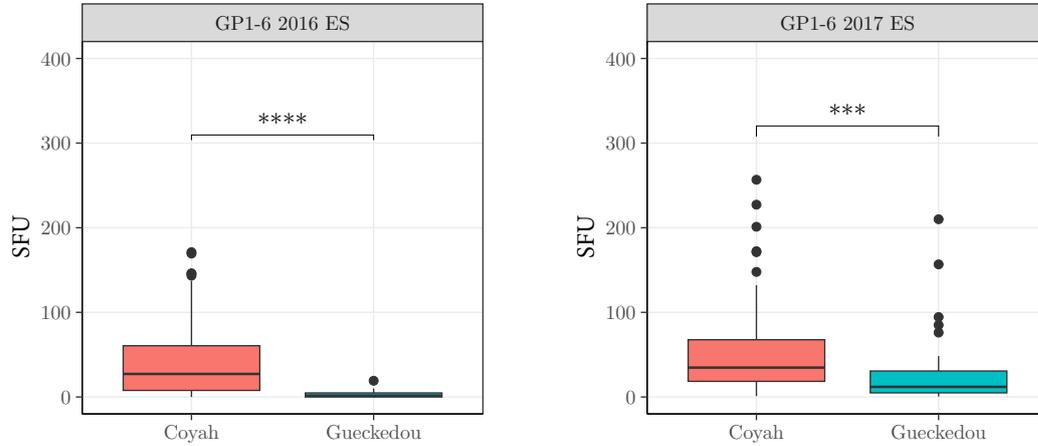
Figure 3.6: GP1-6 ELISpot for IFN-γ concerning years 2016 and 2017 for each cohort. In 2016 there were $n = 84$ and $n = 78$ individuals from Coyah and Gueckedou, respectively. Meanwhile, in 2017 there were $n = 62$ individuals for both cohorts. Statistical significance was determined by the Wilcoxon–Mann–Whitney test with Bonferroni adjustment (ns, not significant; *$p < 5 \times 10^{-2}$ **$p < 10^{-2}$; ***$p < 10^{-3}$; ****$p < 10^{-4}$). Spot forming units (SFU)

particularly in the GP1-6 region. However, this hypothesis can not be tested with the data provided during this study since no sequencing data is available for the lineage of EBOV that each patient was infected with. If this data was available, sequence alignments with the known lineages reported by Simon-Loriere *et al.* (2015) and Carroll *et al.* (2015) could be used to identify whether there was a regional difference in circulating lineages or whether there was a mixture of lineages circulating. This would be particularly interesting for Coyah since the epidemic spread from Gueckedou to this area.

Evidence for asymptomatic EBOV responses

Amongst this PHE cohort are different types of infection status: contact, negative and survivor. It is expected that there will be differences in immune responses, but Bellan *et al.* (2014) and Leroy *et al.* (2000) present evidence that there are asymptomatic responses to EBOV. Hence, since the cohort includes contacts with infected individuals, it is worth investigating whether these people did get

3. STATISTICAL ANALYSIS OF PATIENT DATA FROM VIRAL INFECTIONS

infected or some were simply asymptomatic. Glynn *et al.* (2017) has previously investigated the percentage of asymptomatic individuals and found around 2.6% of contacts were asymptomatic. To study the differences in immune responses, it is worth using PCA to provide a visualisation of the data in a reduced dimensional form and see if our groups defined by infection status differ and whether we can potentially identify any individuals who may be asymptomatic.

Figure 3.7 illustrates the first two principal components for the experimental results from 2015 for IgG ELISA, Marburg ELISA, neutralisation assay and GP IFN-γ ELISpot. The first two principal components account for 81.2% of the total variance within the data. Using Kaiser's rule would keep only the first principal component. As seen from Figure 3.7, two distinct groups are being formed within the data. The first group is individuals infected with the virus, and the second is uninfected individuals, which is made up primarily of contacts and negatives. The significant difference in these two groups comes from a shift in the negative direction in the first principal component, and a slight shift in the negative in the second principal component. The first two principal components account for approximately 90% of the total variance for GP IFN-γ ELISpot and IgG ELISA. In comparison, they only account for around 70% of the variance generated by the Marburg ELISA and neutralisation assay. The correlation circle in Figure 3.8 also shows a strong correlation between the neutralisation assay and both the Marburg ELISA and IgG ELISA and a weaker correlation between the GP ELISpot and neutralisation assay and Marburg ELISA. This makes sense since the Marburg ELISA and neutralisation assay measure humoral responses. There is, however, no correlation between the IgG ELISA and the GP ELISpot experiments.

Upon closer examination of Figure 3.7, it can be seen that there are four individuals labelled as contacts (blue circles) who lie outside the 95% confidence region. Two more individuals lie close to the boundary. The raw data shows these individuals have patient IDs *C017, C086, C098, C110, G012, G034* who are all labelled as contacts of confirmed infected individuals. It is possible that these individuals were actually asymptomatic and therefore did not seek treatment for EBOV disease. Since the first two principal components represent all four experimental outputs, these contacts will likely have immune responses to EBOV if these are examined. They, therefore, will have circulating antibodies and T-cell
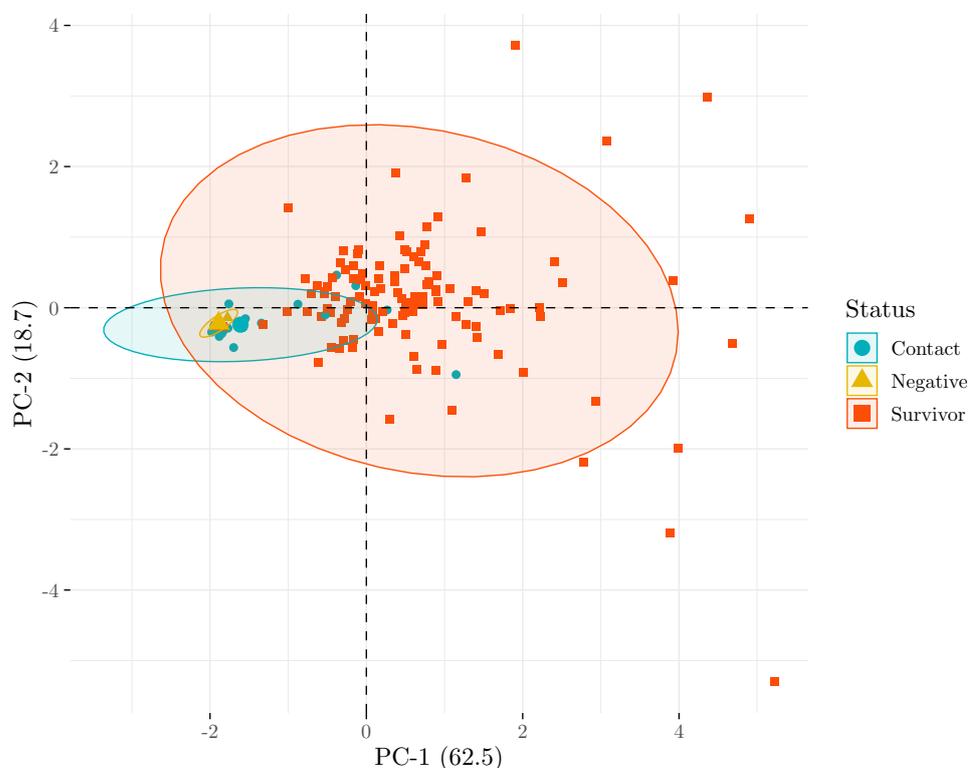
Figure 3.7: PCA defined by infectious status, considering IFN-γ ELISpot result in response to EBOV GP, neutralisation assay, IgG ELISA and Marburg ELISA for 2015 samples. The first PC accounts for 62.5% of the total variance, while the second accounts for 18.1%. Ellipses represent a 95% confidence region.

responses against EBOV. To investigate this, it is worth looking at each experiment on a uni-variate level and seeing if these individuals have immune responses since if they are, in fact, not infected, they should not produce an immune response. The uni-variate box plots are shown in Figure 3.9.

From Figure 3.9, it can first be seen that for contacts, GP ELISpot and IgG ELISA that there are no individuals marked as outliers, and hence do not fall outside the overall pattern of the distribution. However, when the neutralising antibody response and Marburg ELISA are examined, there are seven people and six people, respectively, that lie outside the pattern of the distributions. When the patients' IDs are retrieved from the raw data, the six individuals with the highest
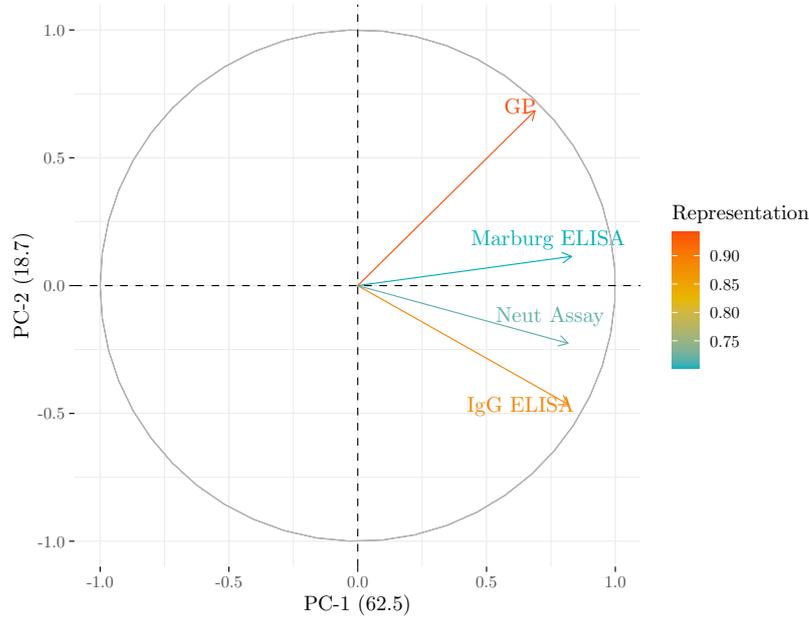
Figure 3.8: Correlation circle for the four experimental outputs shown in Figure 3.7. The angle between variables represents the correlation of variables with the line length from the origin representing the quality of representation. The PCs in this figure match those presented in Figure 3.3.

neutralising antibody levels and the six individuals in the Marburg experiment are the same six individuals found during the PC analysis. All this evidence suggests that these six individuals are not negative contacts but instead asymptomatic individuals. Within this cohort, this would mean that 3.7% of individuals with recorded values in 2015 are asymptomatic, which is consistent with the results found by Glynn *et al.* (2017).

3.1.3 Longitudinal responses

The main reason for this study was to investigate the longitudinal responses of the same group of individuals over several years. Data is available for 2015-2019; however, to best understand the longitudinal effects, it is sensible to use the same set of individuals and examine the trends in their immune response. To maximise
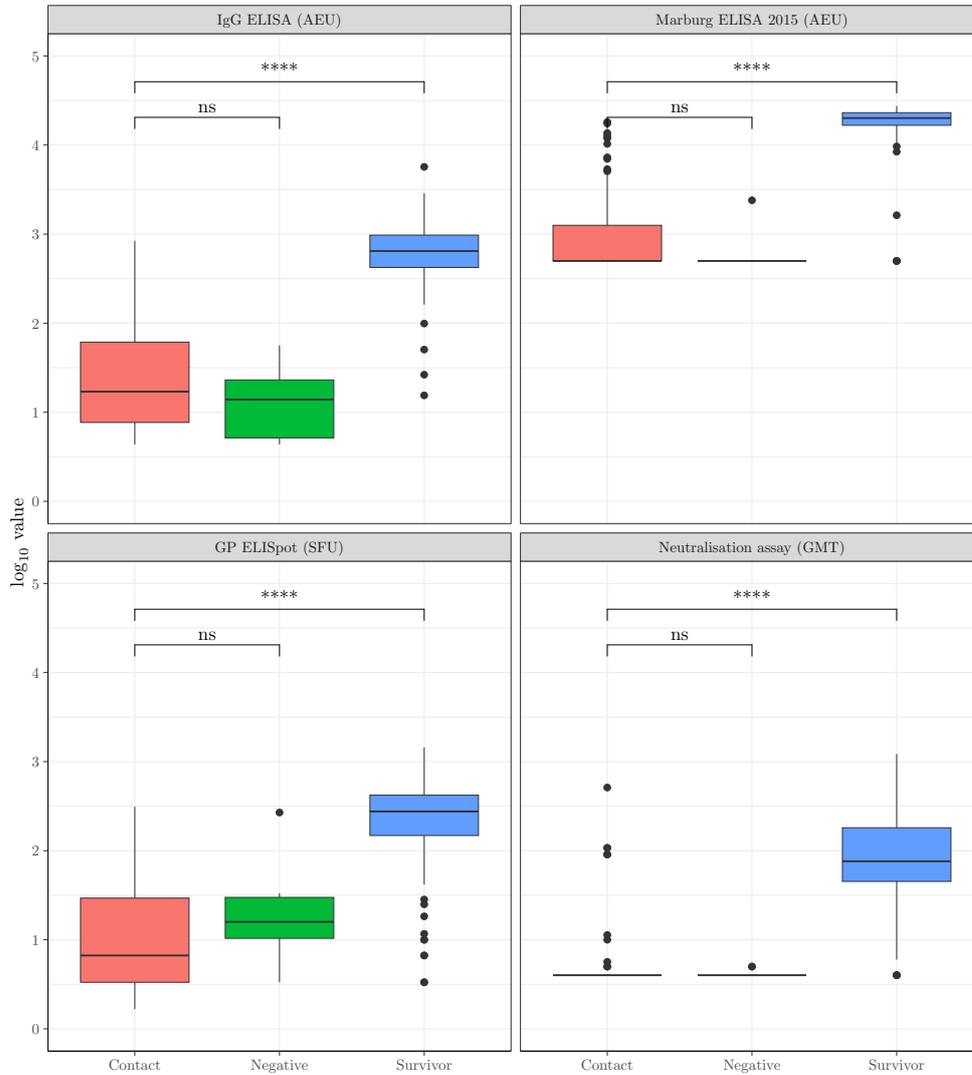
Figure 3.9: Uni-variate plots for each experiment performed in 2015 considered in Figures 3.7 and 3.8, where groups are defined by contact status. Statistical significance was determined by the Wilcoxon–Mann–Whitney test with Bonferroni adjustment (ns, not significant; $*p < 5 \times 10^{-2}$ $**p < 10^{-2}$; $***p < 10^{-3}$; $****p < 10^{-4}$).

the number of individuals that can be used in the study, 2018 results are excluded due to the lack of samples supplied this year. Including 2018 and only using those individuals who provided a sample for all years would result in only 30 people to study. Meanwhile, if 2018 is excluded, 84 individuals are available to explore the trend in response. This is only true for the neutralisation assay and Marburg ELISA results, while for GP ELISpot data, there are only 57 individuals with complete experimental data between 2015-2017.

Figure 3.10a shows the longitudinal immune responses for the 57 individuals who provided a complete set of samples between 2015 and 2017. As can be seen from this figure, there is a lot of heterogeneity in responses, but for the most part, most people are exhibiting an increase in IFN-γ production over time. Since IFN-γ is an indicator of T-cell responses, this implies that the T-cell response is increasing over time. To verify this, a linear mixed effects model was fitted to the data, the results of which are shown in Figure 3.10b. The random effect was placed on the individual's response, and the link between time and IFN-γ production and hence T-cell response was assessed. The linear mixed effects model results returned a slope gradient with value 0.11 (to two decimal places) with p-value 1.98×10^{-3}. Since the gradient of the slope is very shallow, it suggests that infected survivors will see an increase in T-cells that can respond over time, and the ELISpot assay should increase at 1.28 SFU per year. Most of the increase occurred between 2016 and 2017, with T-cell responses strengthening. This could be due to further T-cell expansion or further development of the memory T-cells (Jameson & Masopust, 2009).

Adaken *et al.* (2021) reported a decline in neutralising antibody titres, while Davis *et al.* (2019) reported a decline in IgG antibody responses to EBOV with time. This is due to antibody decay, and in general, decay has been noted for several viruses, including SARS-CoV-2 (Xia *et al.*, 2021). Within this study, both neutralising antibodies (neutralisation assay) were investigated along with IgG antibodies (Marburg ELISA), and it is worth investigating whether, for this cohort, the current data agrees with the findings by Adaken *et al.* (2021) and Davis *et al.* (2019). Figure 3.11a illustrates the longitudinal responses for the 87 individuals who provided samples for 2015 through 2019 with 2018 excluded. As can be seen from this figure, most individuals are experiencing a decrease in their

(a) Individual responses

(b) Linear mixed effects model fit
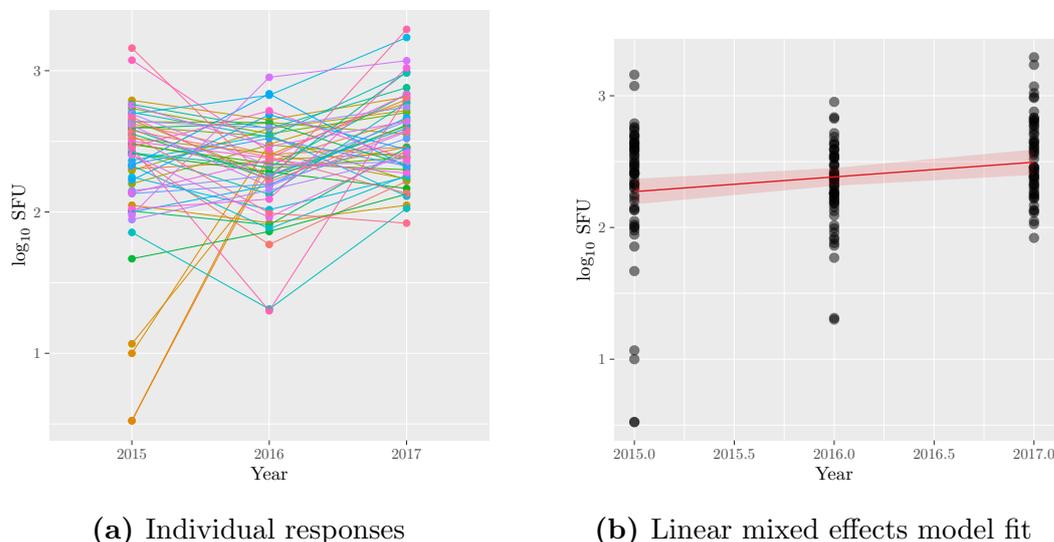
Figure 3.10: GP ELISpot results in $\log_{10}$ SFU. (a) Individual immune responses for patients with complete data for 2015 through 2017, where joined points is data from one person. (b) Linear mixed effects model fit for complete cases. The mixed effect was placed on the individual; the year was a fixed effect. The shaded area represents a 95% credible region for the fit.

antibody levels, and this decay rate depends on the individual. A linear mixed effects model was fitted to the Marburg ELISA data to determine the population decay rate while considering this individual variation. Figure 3.11b shows the fit from this analysis. The linear mixed effects modelling results, where the random effect was placed on the individual, revealed that the population level gradient is -0.023 with p-value 2.73×10^{-9}. This suggests that, on average, an infected survivor's IgG antibody response will decrease at 0.94 AEU per year. Hence, as also found by Davis *et al.* (2019), the IgG antibodies are decaying; however, this is slow, according to the data in this study.

Figure 3.12 illustrates the results from neutralisation assays to EBOV for 2015 to 2019, whereas before 2018 is excluded to maximise the available data. The box plots in Figure 3.12 appear to be suggesting, as with the IgG antibody response, that they are decaying. However, the responses for 2015 were lower than in 2016, 2017 and 2019. As mentioned in the previous sections, this could be because a large number of individuals in the cohort for 2015 had not had enough time to

(a) Individual responses

(b) Linear mixed effects model fit
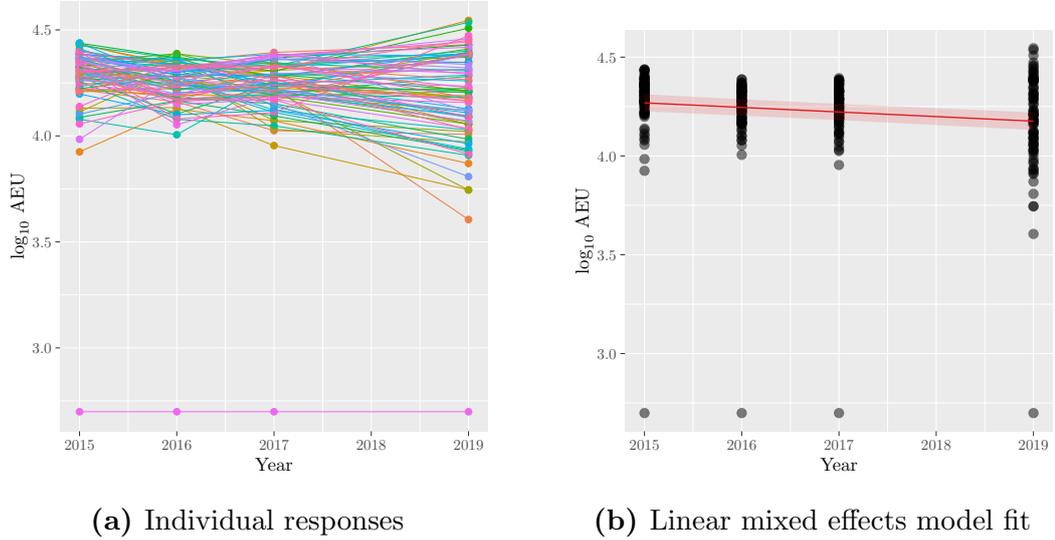
Figure 3.11: Marburg ELISA results in $\log_{10}$ AEUs. (a) Individual immune responses for patients with complete data for 2015 through 2019, where joined points is data from one person. 2018 is excluded to maximise available data (b). Linear mixed effects model fit for complete cases. The mixed effect was placed on the individual; the year was a fixed effect. The shaded area represents a 95% credible region for the fit.

develop an effective neutralising antibody response. For this reason, to assess whether neutralising antibodies are decaying, only 2016, 2017 and 2019 will be used to investigate the population's average decay. Figure 3.13 illustrates both the longitudinal responses for each individual (3.13a) along with the linear mixed effects model fit (3.13b). As seen from Figure 3.13a, there is a lot of heterogeneity in neutralising antibody responses, but most individuals see a reduction in their neutralising antibody levels. This is supported by Figure 3.13b from which the model fit provides a population average decay of 0.79 GMT per year with a p-value of 2.43×10^{-7}. Hence, for our cohort, both IgG and neutralising antibodies are decaying in agreement with Adaken *et al.* (2021) and Davis *et al.* (2019). However, it is worth noting this rate of decay is very slow. Since it has been previously found that survivors after 40 years still have detectable antibody titres (Rimoin *et al.*, 2018), likely, this decay is not linear and perhaps asymptotic. With additional data it would be possible to investigate the true trend of the decay.
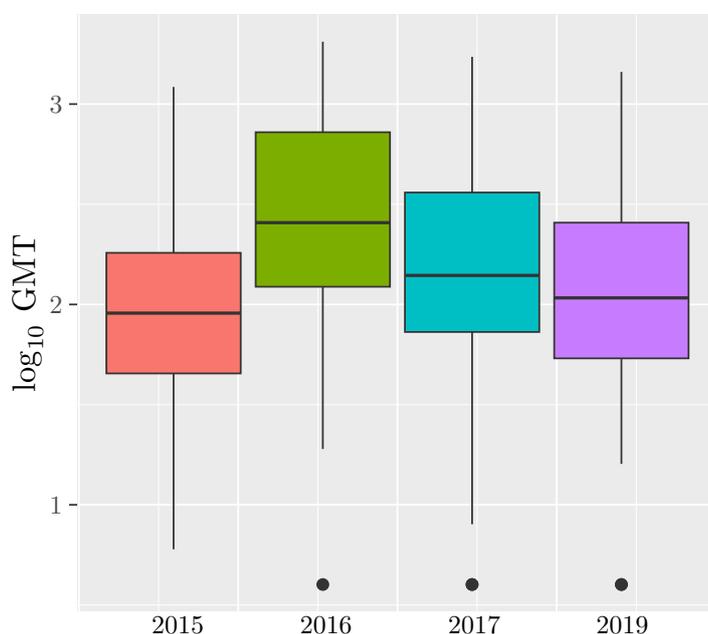
Figure 3.12: Neutralisation assay boxplot in $\log_{10}$ GMT from 2015 through until 2019.

3.1.4 Human leukocyte antigen analysis

As mentioned in Chapter 1, one of the unique parts of this study was the genetic component, particularly investigating associations with human leukocyte antigens and different immune responses. There have been several studies that have investigated the role of HLA on disease outcomes. Kiepiela *et al.* (2004) studied the effects of HLA-A and HLA-B on mediating HIV responses, while Boon *et al.* (2002) found that HLAs mediate the magnitude and specificity of CD8 T-cell responses to influenza virus. In this section, two questions are to be investigated. The first question is whether HLA type affects the production of IFN-γ and hence the T-cell response; secondly, whether there are any regional differences in HLA expression between Coyah and Gueckedou.

HLA type and the association with IFN-γ

IFN-γ is produced by CD8 and CD4 T-cells in response to infection as found by both Gazzinelli *et al.* (1991) and Thom *et al.* (2021). However, in the case

(a) Individual responses

(b) Linear mixed effects model fit
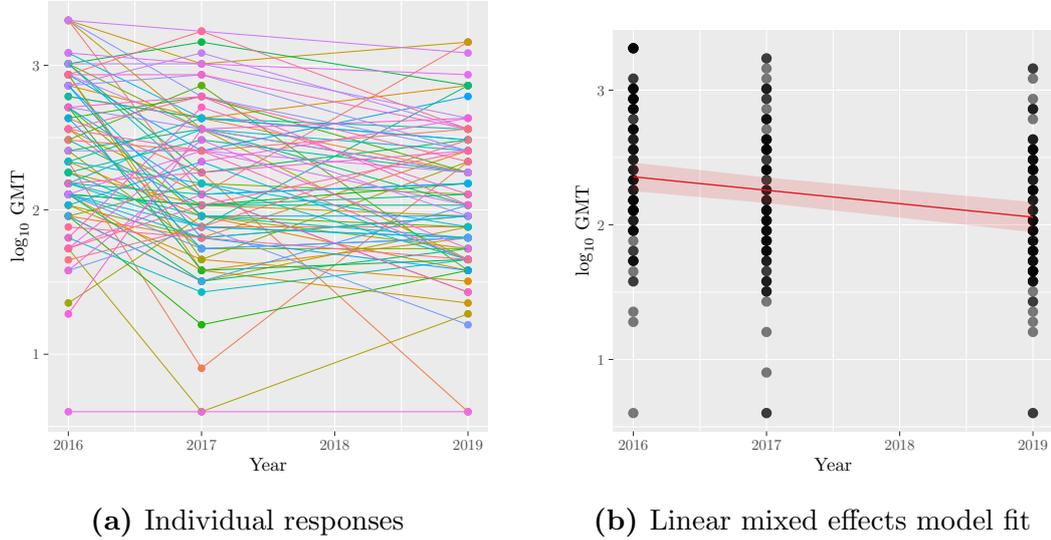
Figure 3.13: Neutralisation assay results in $\log_{10}$ GMT. (a) Individual immune responses for patients with complete data for 2016 through 2019, where joined points is data from one person. 2018 is excluded to maximise available data. 2015 is also excluded as an effective neutralising antibody response may not yet have been formed. (b) Linear mixed effects model fit for complete cases. The mixed effect was placed on the individual; the year was a fixed effect. The shaded area represents a 95% credible region for the fit.

of EBOV, it was found by Thom *et al.* (2021) that CD8 T-cells produce the majority of IFN-γ during infection. As a result, class I alleles will be investigated in conjunction with the IFN-γ ELISpot data to see if there is any association between HLA alleles and the amount of IFN-γ produced and, therefore, CD8 T-cell activity. Additionally, since the different subunits of the GP were also tested, associations within specific GP subunits and HLAs can also be investigated. In addition to the statistical analysis that may inform that a particular HLA gives a differential response in a given subunit, the Immune Epitope Database (IEDB) can be used to estimate the probability that a specific HLA can bind to the epitope associated with a viral protein to allow antigen presentation to T-cells. For each HLA allele, IEDB returns a score from 0 to 1, where 1 is the best score, representing the binding probability. This is the probability of the HLA binding to a region on the protein. For the following analysis, 108 individuals were used

from 2016 who had complete GP ELISpot results and had their HLA typed to allow maximum data.

Table 3.3 shows the alleles, with respect to each subunit of the GP, associated with a significant difference in IFN-γ production in individuals that expressed it, compared to those who do not. The reported p-value is before correction. The SP subunit has five alleles that are statistically significant before correction. These alleles are B42:01, C17:01, A23:01, A01:02 and A34:02. All alleles except A01:02 show higher median levels of IFN-γ in individuals who express this allele compared to those without. From IEDB, binding predictions for these alleles, with respect to the SP subunit genome, found that all the HLA A alleles were found to be poor binders to the SP proteins with scores of < 0.40, while B42:01 and C17:01 are both in the top 3 B and C alleles, respectively (scores of 0.89 and 0.83, respectively).

For the GP1-1 and GP1-2 subunits, individuals who expressed A80:01 have higher levels of IFN-γ from ELISpot assays with medians of 85.83 and 158.33 SFU for GP1-1 and GP1-2 respectively. However, binding predictions suggest that this allele is a poor binder for GP1-2 with a score of 0.26. While for GP1-1, it is in the top 20 potential binding alleles with a score of 0.52. GP1-1 additionally has A26:01, which is not only a top binder with a score of 0.71 but also has individuals who express this allele having three times higher median SFU, 85.83 SFU compared to 22.50 SFU. GP1-2 also has an additional allele B15:03 which results in an almost 50% increase in IFN-γ (62.50 SFU versus 42.22 SFU) compared to those who do not present this allele. The GP1-3 subunit has alleles B15:16, C14:02 and A33:01 that are all associated with a higher level of IGN-γ in ELISpot assays compared to those who do not express the allele. Within the HLA C type, C14:02 is the best-predicted binder for the region with a score of 0.79, while A33:01 is in the top 10 with a score of 0.92, and B15:16 is only in the top 20 with a score of 0.59. A23:01 is a poor binder with a score of 0.4, resulting in almost a three times reduction in IFN-γ production compared to people lacking this allele. GP1-4, like GP1-3, has three alleles that are associated with better IFN-γ responses: C03:02 (92.5 SFU versus 35 SFU), B58:01 (95.69 SFU versus 35.83) and A26:01 (61.94 versus 37.07). Meanwhile, C16:01 is associated with a lower level of IFN-γ compared to those without. From binding predictions, B58:01 is

3. STATISTICAL ANALYSIS OF PATIENT DATA FROM VIRAL INFECTIONS

Region	Allele	Median with	Median without	p-value	IEDB Score
SP	B42:01	51.67	5.69	**1.6×10^{-4}**	0.83
	C17:01	27.78	6.25	0.005	0.42
	A23:01	20.83	5.56	0.009	0.41
	A01:02	0.28	9.72	0.014	0.13
	A34:02	53.89	7.36	0.017	0.36
GP1-1	A26:01	69.70	22.36	0.021	0.71
	A80:01	85.83	22.50	0.025	0.52
GP1-2	A80:01	158.33	42.50	0.013	0.26
	B15:03	62.50	42.22	0.017	0.58
GP1-3	B15:16	165.56	17.50	0.004	0.59
	C14:02	165.56	17.50	0.004	0.92
	A33:01	39.17	14.44	0.005	0.40
	A23:01	8.89	22.5	0.021	0.79
GP1-4	C03:02	92.5	35.00	0.001	0.35
	B58:01	95.69	35.83	0.002	0.28
	C16:01	27.36	47.64	0.004	0.95
	A26:01	61.94	37.08	0.038	0.18
GP1-5	C07:01	1.67	0.83	0.028	0.12
	A30:02	2.78	0.83	0.046	0.07
	B51:09	12.78	0.83	0.046	0.06
GP1-6	A26:01	63.89	5.00	0.016	0.52
	A03:01	0.83	6.67	0.026	0.54
	A30:01	19.17	5.00	0.029	0.05
GP1-7	B:49:01	19.44	7.50	0.005	0.39
	A02:05	21.39	7.50	0.008	0.46
	C17:01	17.78	7.50	0.015	0.41
	A34:02	2.22	9.44	0.035	0.49
	C06:02	4.79	10.00	0.045	0.05
GP2-1	B58:01	30.42	7.92	0.005	0.99
	B07:02	3.33	11.67	0.0499	0.26
GP2-2	A29:02	82.78	8.06	0.037	0.07

Table 3.3: Alleles found to have significant difference before correction by Mann-Whitney U test for ELISpot IFN-γ, with respect to each GP subunit. Median ELISpot SFU expression with and without the allele is presented. Bold font indicates significance after correction for multiple comparisons using Bonferroni's method at a 5% significance level.

the second best for the B alleles with a score of 0.95, while the A and C alleles are all poor binders with scores < 0.4. This supports the idea that expressing C16:01 negatively impacts IFN-γ production since it cannot bind to the GP1-4 region and express this for T-cell recognition.

Within the GP1-5 region, there are three alleles identified. However, as seen from Table 3.3, the SFU expression is low compared to many other GP subunits. All three of these alleles are poor binders to the GP1-5 subunit with scores < 0.2; hence, this could explain the poor responses in this region. Three alleles are significant before correction with respect to the GP1-6 region: A30:01, A03:01 and A26:01. While A30:01 and A03:01 are average binders with scores ≈ 0.5, A26:01 is a poor binder with a score of 0.05. However, individuals who possess A26:01 express over ten times the SFU for IFN-γ compared to those who do not. There could be additional biological reasons for this that the HLA does not explain. GP1-7 has three alleles associated with higher responses in individuals who express it. B49:01, A02:05 and C17:01 are all in the top 20 binders for each HLA categorisation. C06:02 is identified as a poor binder with a score of 0.05, which explains the lower expression than those who do not express it. Meanwhile, A34:02 had a low binding score of 0.4, which explains the poor response in individuals who express this allele. GP2-1 has one allele, B58:01, associated with increased responses by those who possess it with a median value of 30.42 SFU compared to 7.92 SFU. IEDB binding predictions suggest this allele is a suitable binder with a score of 0.99. Meanwhile, B07:02 is a poor binder with a score of 0.26. As in several other cases, individuals who express this allele have lower IFN-γ production than those who do not possess the allele (median 3.33 SFU with versus 11.67 without). GP2-2 has one allele that is associated with an almost ten times increase in IFN-γ (82.78 SFU for individuals with compared to 8.06 SFU without A29:02). However, this allele was identified to be a poor binder, so, as with A26:01 there could be additional biological reasons for the high SFUs. While these relations are significant before correction after applying Bonferonni's correction method, only one allele association with IFN-γ in ELISpot assays remains significant, and that is for the SP subunit and allele B42:01. However, with a larger sample size, it may be possible that associations

found before correction hold. Finally, additional data would allow one to study the association of pairs of HLAs and their effect on immune responses.

HLA frequency in different regions

I now examine differences in allele frequencies between the two locations. These are summarised below in Table 3.4. This table shows upon initial inspection that several alleles are expressed in higher frequencies in one population compared to another. In particular, A23:01, A30:01, B53:01 and C17:01 and C06:02. A two-proportion Z-test to examine whether these observed frequency proportions can be considered different. This test determines whether the two proportions are different with a null hypothesis that the proportions are identical. Further details on calculating this statistic can be found in Agresti (2012). Performing the two-proportion Z test on the previously mentioned alleles, with a significance level of $\alpha = 0.05$, shows A30:01 (*p*-value 0.0065) and B42:01 (*p*-value 0.045) are the only two alleles whose differences are statistically significant. The remaining alleles A23:01 (p-value 0.26), B53:01 (p-value 0.30), C17:01 (p-value 0.061) and C06:02 (p-value 0.11) are all statistically insignificant, and hence their proportion of expressed alleles are marginally different.

3.1.5 Discussion

In this subsection, I investigated the immune responses of EBOV survivors and close contacts from the 2014 West Africa outbreak in a longitudinal cohort study undertaken by Public Health England (now known as the UK Health Security Agency). Previous studies have investigated the humoral responses to EBOV, such as that presented by Baize *et al.* (1999) for survivors and fatalities. Furthermore, several studies have investigated the differences in T-cell response between fatalities and survivors and found that a broad T-cell diversity correlates with survival (Speranza *et al.*, 2018). However, few studies have investigated the longitudinal responses of survivors after infection. Public Health England endeavoured to carry out a longitudinal study of immune responses for survivors of EBOV infection by recruiting participants from EBOV treatment centres to

HLA-A	C	G	HLA-B	C	G	HLA-C	C	G
A02:01	0.143	0.163	B53:01	0.174	0.123	C04:01	0.197	0.235
A23:01	0.136	0.098	B35:01	0.129	0.180	C16:01	0.190	0.227
A03:01	0.041	0.065	B42:01	0.100	0.025	C07:01	0.127	0.109
A30:01	0.109	0.024	B15:03	0.100	0.066	C17:01	0.092	0.033
A30:02	0.068	0.114	B78:01	0.075	0.066	C02:10	0.070	0.059
A02:02	0.061	0.049	B58:01	0.075	0.066	C06:02	0.070	0.025
A33:03	0.061	0.049	B07:02	0.053	0.098	C03:02	0.056	0.059
A34:02	0.054	0.049	B08:01	0.038	0.033	C03:04	0.035	0.034
A68:02	0.054	0.089	B45:01	0.045	0.057	C07:02	0.035	0.076
A33:01	0.048	0.073	B18:01	0.045	0.008	C01:02	0.035	0.000
A74:01	0.048	0.041	B49:01	0.045	0.008	C05:01	0.028	0.008
A26:01	0.041	0.016	B56:01	0.030	0.000	C02:02	0.014	0.025
A01:02	0.027	0.024	B44:03	0.015	0.049	C14:02	0.014	0.008
A68:01	0.027	0.033	B27:03	0.023	0.057	C18	0.014	0.017
A02:05	0.020	0.024	B52:01	0.023	0.025	C03:03	0.007	0.025
A01:01	0.020	0.024	B57:03	0.023	0.000	C15:05	0.007	0.025
A66:01	0.014	0.008	B51:01	0.015	0.016	C08:04	0.007	0.000
A66:02	0.007	0.008	B15:16	0.015	0.008	C16:112	0.000	0.008
A24:02	0.007	0.000	B51:09	0.015	0.000	C08:02	0.000	0.025
A29:02	0.007	0.008	B82:01	0.015	0.008			
A31:01	0.007	0.000	B27:05	0.008	0.000			
A80:01	0.000	0.033	B50:01	0.008	0.016			
A32:01	0.000	0.008	B40:02	0.008	0.000			
			B51:02	0.008	0.000			
			B15:10	0.008	0.016			
			B41:02	0.000	0.008			
			B14:02	0.000	0.008			
			B44:10	0.000	0.016			
			B35:08	0.000	0.008			
			B07:05	0.000	0.008			
			B57:02	0.000	0.008			
			B78:02	0.000	0.008			
			B14:01	0.000	0.008			

Table 3.4: Summary of relative proportions of which HLA alleles appear amongst the two regions in the cohort. C=Coyah, G=Gueckedou.

provide repeated yearly blood samples to investigate both antibody response and IFN-γ production, a marker of T-cell responses.

Within this study, there are two regions of interest where samples were collected. This first location is Gueckedou which, as discussed by Simon-Loriere *et al.* (2015), was the site of the incident case that resulted in the outbreak, and Coyah, another city located 385 miles to the west. Due to the proximity of these cities to the border with Sierra Leone and because there were three distinct lineages of EBOV circulating during the outbreak (Carroll *et al.*, 2015), it is possible that individuals in these two regions mount different responses to EBOV. To investigate this, the IFN-γ response to the glycoprotein was investigated between the two regions. It was shown for 2015 that the responses from Gueckedou were higher than those from Coyah. However, as the longitudinal data is examined, it was revealed that the only statistically significant differences in responses corresponded to the GP subunit GP1-6 for 2016 and 2017. The most likely reason for the difference in responses in 2015 is that the infections from Coyah are more recent than Gueckedou, and hence the adaptive immune response has not yet matured since IFN-γ is a marker of T-cell response. However, the limited information on first infections makes it impossible to confirm this hypothesis fully. The difference in GP1-6 subunit response in 2016 and 2017 could be explained by better sequence similarity between the Kitwit variant used for ELISpot experimental testing of immune responses and the circulating strain present in Coyah. However, without sequences from the individuals in both Coyah and Gueckedou, it is impossible to identify similarities between strains using sequence alignment methods and determine whether they were exposed to different strains that may explain the difference in response to this GP subunit.

Within the cohort, individuals are classified according to infection status: survivor, contact, and negatives. Survivors were confirmed from either treatment centre databases or by the individuals supplying their EBOV survivor certificates. Contacts are individuals who have come into close contact with an infected person but presented with no symptoms. The principal component analysis showed that six individuals could be potentially asymptomatic contacts, with four lying outside the 95% confidence ellipse. Upon closer inspection of these individual responses, it was found that these six individuals had not only IgG antibodies in response to the

whole inactivated EBOV but also expressed neutralising antibodies. Since negative controls do not express any antibodies or a meagre amount, this expression by these six individuals is evidence of an asymptomatic immune response to EBOV. In total, these 6 individuals account for 3.7% of individuals with results in 2015 that are asymptomatic. These results are consistent with Glynn *et al.* (2017), who also found evidence of asymptomatic individuals in a separate cohort.

As mentioned in the introduction to this section, the main goal was to assess the longitudinal immune responses of individuals who have been infected with EBOV and survived. A previous study by Rimoin *et al.* (2018) found detectable antibody response 40 years after infection with EBOV, which provides evidence that antibodies can respond to repeated exposure to EBOV. Within the cohort presented here with data from survivors, it was shown that antibody responses were decreasing in magnitude, which is an indicator of antibody decay. It was demonstrated for both neutralising antibodies and the IgG isotype that the gradient at the population level was negative. In particular, the results for the Marburg ELISA to test for IgG antibodies had a gradient of -0.023, which correlated with a decrease of 0.94 AEU per year. Meanwhile, neutralising antibodies had a gradient of -0.0998, which gives a reduction of 0.79 GMT per year. This supports the idea that antibody decay and immune responses will weaken over time, citeadaken2021ebola and Davis *et al.* (2019) both found similar results. However, since the data is only from several years of observations, it is not apparent whether this decay is linear or follows a different curve. Furthermore, it was also shown that there was an increase in IFN-γ responses over time, with a rise of 1.28 SFU per year. Since it was demonstrated that IFN-γ is predominantly associated with CD8 T-cells (Thom *et al.*, 2021), this could indicate that CD8 T-cells are increasing. However, due to the scale being in years, it is more likely that the T-cell memory repertoire is growing rather than the T-cell effector pool.

Finally, a genetic component of immune responses was also examined, particularly HLA association with IFN-γ responses. Thom *et al.* (2021) showed that in EBOV infections, it was mainly CD8 T-cells that expressed IFN-γ, the Type I HLA alleles were investigated against the ELISpot data for each GP subunit. Before correction for multiple comparisons, various HLA alleles were associated with either increased IFN-γ production by those who expressed the allele compared to

those who did not. After correction, however, there remained only one allele that was statistically significant in response to stimulation of the SP region. Individuals who expressed B42:01 expressed a higher median level of IFN-γ at 51.67 SFU, compared to 5.69 SFU for those who do not possess the allele. This indicates that individuals possessing this allele may have a better T-cell response than those who do not. With extra data, it is possible that associations found before correction would hold after correction, but this cannot be confirmed without additional data. Furthermore, it was shown that there is a regional difference in the alleles expressed. A30:01 and B42:01 are expressed in higher proportions in Coyah than in Gueckedou. However, the small sample size of the cohort is worth mentioning, which may not represent the actual distribution of alleles in different regions.

There are several limitations of the study that has been presented here that should be mentioned. The first issue is the inconsistency in recorded data. Many individuals have missing fields within the data, such as age, date of entry into a treatment clinic, missing genders and date of first positive result. A lot of this demographic and date information would have allowed for detailed investigation into the longitudinal responses since, for example, knowing all patients' first positive sample date would have allowed experimental results to be aligned for better comparisons considering time from diagnosis. Furthermore, many patients' missed one or more time points throughout the study, which led to reduced data, and hence a reduction in the conclusions' power since sample sizes were reduced and inconsistent from year to year. In addition to missing data, there was a change in experimental design between 2015 and 2016, where the standard of IgG antibody detection for the GP was changed to meet WHO standards. Hence, this data is not comparable due to different units. Another limitation of this study is the lack of fatal patient data. While there have been many studies that investigate the difference in fatal and survivor data, such as those by Speranza et al. (2018) and Colavita et al. (2019), it would have been interesting to have consistent experimental methods to allow comparison of the two cases within the one cohort study. A final limitation of this study is the overall population size. Since immune response are heterogeneous, and there are a wide array of alleles for HLA, the size of the population needs to be large such that the statistical analysis holds more weight. An increase in the population size would allow for

more powerful conclusions to be made and, in the case of the HLA analysis, would possibly show more associations to be statistically significant after correction. Additionally, more data would allow comparing pairs of HLA alleles to see if pair expression affects immune responses.

3.2 St Jude Children's Research Hospital SARS-CoV-2 trace study

In March 2020, the World Health Organization declared SARS-CoV-2 a "public health emergency of international concern" after cases rapidly spread to 113 countries. This outbreak was soon reclassified as a global pandemic leading many countries to implement lockdown strategies to slow the spread of the virus, reduce cases and ease pressure on health and social care providers (Tobías, 2020). However, many countries delayed implementation of lockdown procedures, supply shortages, and poor knowledge of transmission dynamics led to most social and healthcare settings being overwhelmed (Li *et al.*, 2020a). Additional stress was added by staff becoming infected, isolating due to exposure, or being overworked. These compounding factors led many hospitals to implement strict testing regimens to reduce transmission among staff and patients. This prompted St Jude Children's Research Hospital in Memphis, Tennessee, United States of America, to begin a longitudinal study to understand host responses to SARS-CoV-2. Much of this work is already published in Lin *et al.* (2022).

SARS-CoV-2 induces a wide range of responses from infected individuals, with some experiencing mild to moderate disease or being asymptomatic while others have severe respiratory distress that can lead to death. One of the unanswered questions is the impact of previous exposure to human common coronaviruses (hCCoV) on susceptibility and disease severity of SARS-CoV-2. As discussed in Chapter 1 there are four circulating strains of hCCoVs that account for $\approx 30\%$ of all "common cold" infections and share a $\approx 30\%$ sequence homology in the spike protein (Hicks *et al.*, 2021). Other studies have found cross-reactive antibodies between hCCoV and SARS-CoV-2; however, little investigation has taken place to understand whether this is beneficial or detrimental to SARS-CoV-2 infection (Ng *et al.*, 2020; Sealy & Hurwitz, 2021). Therefore, it is essential to investigate the effects of pre-existing antibody responses of hCCoV on SARS-CoV-2.

Current investigations into pre-existing antibody responses have yielded mixed and often opposing results, which requires clarification. Some studies, such as those from Dugas *et al.* (2021), showed an increase in hCCoV antibodies was associated with milder disease, while Shrock *et al.* (2020) and Aydillo *et al.* (2021)

found the opposite. These studies, however, lacked baseline measurements, so they did not account for any change in antibody levels after infection. For these reasons, St Jude Children's Research Hospital wanted to investigate antibody levels for SARS-CoV-2 and hCCoV with baseline samples presented. In the following work, I will explore the relationships between hCCoV and SARS-CoV-2 to try and infer whether pre-existing responses affect the severity and SARS-CoV-2 antibody responses. Additionally, I will examine how demographic differences could correlate with differences in hCCoV antibodies, such as race, gender and age. Finally, I will examine antibody response differences between vaccinated and infected individuals.

3.2.1 Data summary

A summary of the data is as follows. The St. Jude tracking of viral and host factors associated with COVID-19 study (SJTRC), is a longitudinal study of adult employees at St Jude Children's Research Hospital. Participants provided written consent for their participation and a completed questionnaire relating to their demographic information and medical history. Upon enrolment, a baseline blood sample was taken along with a nasal swab to determine infection status by PCR. Subsequent weekly nasal swabs were taken when participants were on campus to check for new infections. For individuals that tested positive, additional blood samples were taken within two weeks (acute sample) and then three to eight weeks (convalescent sample). Newly infected individuals also answered a questionnaire on their symptoms every two weeks post-infection. Some individuals would also provide post-convalescence samples for extra analysis sometime after eight weeks. These samples allowed for investigating immune responses before and after infection with SARS-CoV-2. Vaccinated individuals provided additional samples after the first vaccination and then three to eight weeks after the second vaccination. Due to the study's timing, very few individuals received an additional booster vaccine. For any analysis that examines differences in natural infection against vaccination, individuals who had a vaccination and an infection were removed, and in total, 1266 individuals participated in the study with varying degrees of patient interaction.

3. STATISTICAL ANALYSIS OF PATIENT DATA FROM VIRAL INFECTIONS

Blood samples allowed the analysis of several different immunological and genetic markers. An enzyme-linked immunosorbent assay (ELISA) was undertaken for every individual's blood sample to detect antibodies against human common coronaviruses (hCCoV) spike protein. This was undertaken with respect to the four current endemic strains of hCCoV:, hCCoV-NL63, hCCoV-HKU1, hCCoV-229E and hCCoV-OC43, with three types of antibodies to be detected. These antibody types are immunoglobulin (Ig) M, IgG and IgA. The same antibodies were typed for SARS-CoV-2, but responses were measured for the spike protein, receptor binding domain (RBD), and nucleocapsid protein (N). All antibody results were reported in normalised optical density (OD). However, very few samples from baseline measurements were investigated against SARS-CoV-2; most data comes from either post-vaccine or post-infection samples. A more detailed explanation of the experimental methods can be found in Lin *et al.* (2022). Different cytokines were also investigated using a cytometric bead array (CBA), a flow cytometry application that allows users to quantify multiple proteins simultaneously. However, this data will not be analysed in the work that follows. Many of the samples failed to reach the threshold detection value. As such many samples were reported with a minimum detection threshold value, which does not provide reliable information for analysis.

Before investigating the immunological data, it is worth providing a summary of the demographic data collected by St Jude. Overall, there were 1266 participants, with 932 (74%) being female and 334 (26%) being males. The average age of all participants is 43.8 (SD 12.02) years, with the most senior being 83 years and the youngest 20 years. Amongst females, the average age is 43.5 (SD 11.9) years, with the most senior female participant being 78 years, for male the average age is 44.5 (SD 12.3) years, with the oldest male 83 years. For both genders, the youngest participant was 20 years. The cohorts average body mass index (BMI) is 27.56 (SD 6.07), with males having an average of 27.28 (SD 4.89) and females 27.67 (SD 6.45), the largest recorded BMI amongst females is 58.35 compared to the males maximum 54.64. Within this cohort, 979 reported themselves as non-Hispanic, white, Caucasians, with 32 reporting as Hispanic, white, Caucasians and five labelled as "Other", white Caucasian. There are also 116 Asian individuals, 103

black individuals, one person of American Indian or Alaska Native descent and 32 individuals of unknown race. 41 people self-reported as immunocompromised.

3.2.2 Antibody response

In Chapter 1, it was stated that SARS-CoV-2 is a member of the same family as human common Coronaviruses (hCCoV) and, in particular, is of the same linage of hCCoV-HKU1 and hCCoV-OC43 which are all β-coronaviruses. It is unclear whether pre-existing immune responses to hCCoV protect against SARS-CoV-2 infection (Sealy & Hurwitz, 2021). One may expect that antibodies to hCCoV could provide additional support in eliciting an effective immune response to SARS-CoV-2 infection if antibodies express cross-reactivity amongst viral strains. However, if there is insufficient cross-reactivity of antibodies, expansion of the hCCoV antibodies could hinder the formation of effective SARS-CoV-2 antibodies. Previous studies have reported conflicting information on the hCCoV antibody responses. Some studies, such as those from Dugas *et al.* (2021), showed no increase in hCCoV antibodies, while Shrock *et al.* (2020) and Aydillo *et al.* (2021) reported an increase in one or more hCCoV antibodies after SARS-CoV-2 infection. The previously mentioned studies, however, may express inconsistencies in their results due to not having a baseline measurement before infection with SARS-CoV-2. This factor can be overcome in the study presented here since both a baseline and post-infection or vaccination exist.

Before investigating the differences in endemic hCCoV, it is worth examining if there is cross-reactivity between the four hCCoV viruses. Figure 3.14 illustrates correlations between the four hCCoV viruses studied and three antibody isotypes: IgG, IgA and IgM. What is interesting to note is that there are not only positive correlations between alpha and beta coronaviruses but also, in general, correlations between all four hCCoV in a given isotype. For example, OC43 for IgM has a strong positive correlation with HKU1 and 229E and NL63. These correlations within isotypes, rather than between viruses, suggest there is cross-reactivity among hCCCoV-specific antibodies, with a higher degree of cross-reactivity in the IgM response followed by IgA then IgG. This is in agreement with the findings of Poston *et al.* (2021).
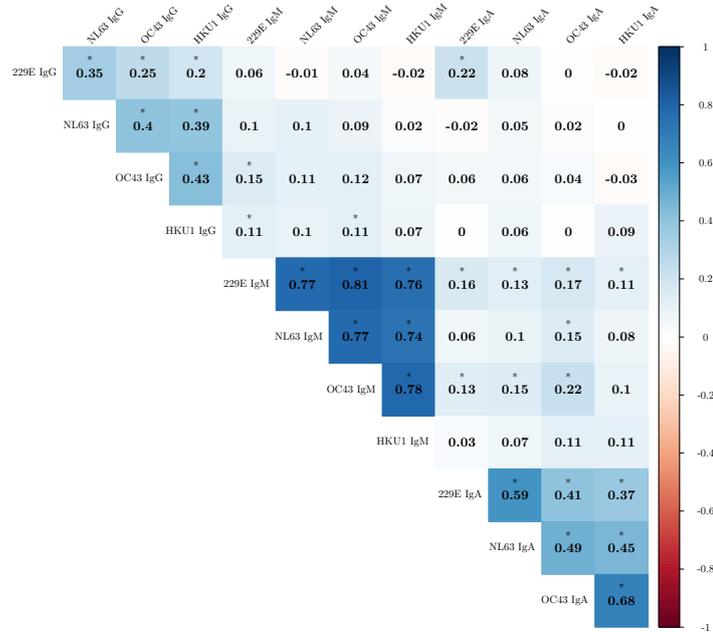
Figure 3.14: Correlations between endemic hCCoV viruses for isotypes IgG, IgM, and IgA using baseline samples. Correlations calculated using Spearman correlation correction adjusted using Bonferroni method ($*p < 5 \times 10^{-2}$)

Endemic hCCoV antibody isotypes in different demographics

As previously mentioned, one of the key differences between this study and others is the availability of baseline samples to examine hCCoV responses before and after infection and vaccination. Therefore, it is interesting to investigate the differences in these baseline responses between different demographics. It is well known that as we age, our immune systems go into decline; as such, one can expect a difference in immune responses between those of an older age compared to a younger age. As such, the cohort was classified into two groups "Younger" whose age was < 43 and "Older" with ages ≥ 43. The age of 43 was chosen as this is the median age in the cohort. To examine the total variability caused by each antibody isotype with
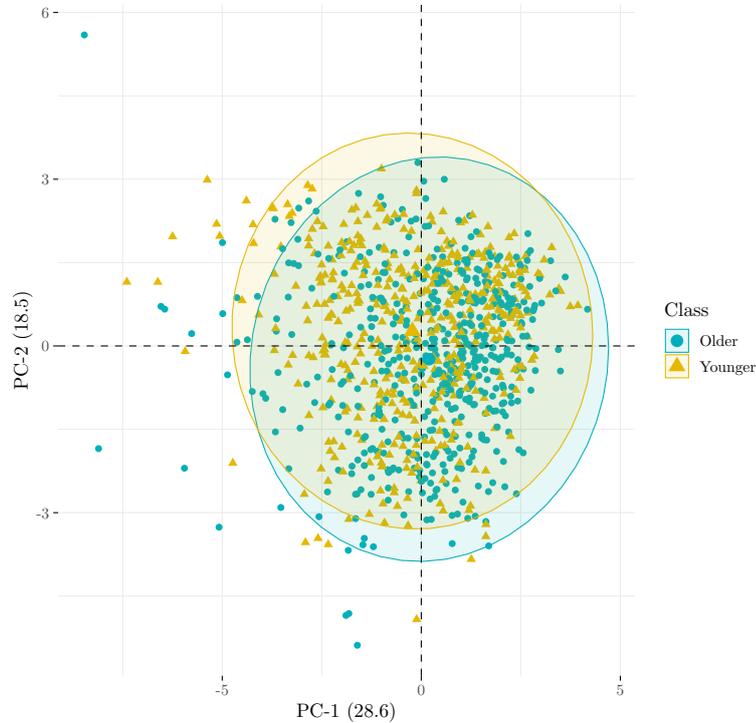
Figure 3.15: PCA for endemic hCCoV antibody isotypes for "Younger" and "Older" groups. The first PC accounts for 28.6% of the total variance, while the second PC accounts for 18.5%.

respect to each endemic strain of hCCoV, principal component analysis was used to reduce the original set of 12 experimental variables into a new set of 12 principal components. This will allow one to visualise any difference within responses of the age classes. Figure 3.15 represents the first two PCs, accounting for 28.6% and 18.5% of the variance, respectively. Using Kaiser's rule, the first four PC would be kept, accounting for 72.4% of the total variance in the data. Figure 3.15 shows that the two groups do not appear to differ significantly, with the "Older" classification having a shift to the right in PC-1 and a shift downwards in PC-2. This may imply some differences in the responses between these groups. However, it is important to understand what these two PC represent. Figure 3.16 shows the proportion of each original hCCoV antibody isotype that each PC represents. As can be seen from Figure 3.16, the first principal component predominantly

3. STATISTICAL ANALYSIS OF PATIENT DATA FROM VIRAL INFECTIONS
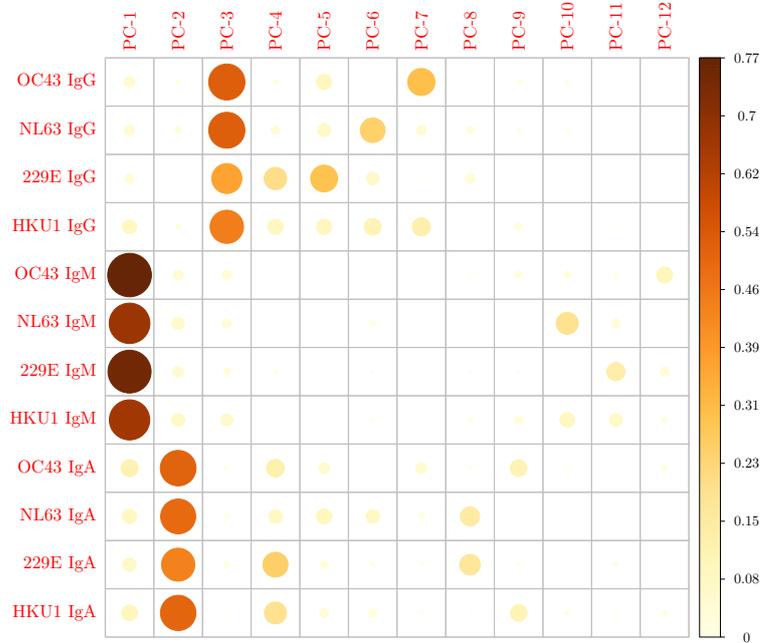
Figure 3.16: Representation proportion for each baseline hCCoV antibody isotype experiment with respect to each PC.

represents hCCoV antibodies with respect to isotype IgM, while PC-2 and PC-3 represent isotypes IgA and IgG, respectively. Therefore, the shifts exhibited in Figure 3.15 could represent differences in responses to these isotypes.

To investigate these differences, it may be more beneficial to consider these sets of experiments at the uni-variate level. Figure 3.17 shows box plots comparing the differences between the immune responses of the two age groups for each isotype. From Figure 3.17, we conclude that older individuals have significantly higher responses for antibody IgA for strains 229E (p-value < 0.001), HKU1 (p-value < 0.0001), and NL64 (p-value < 0.0001) but not OC43. Conversely, younger individuals show higher levels of isotype IgM for all hCCoV viral strains. This could be explained by the fact that IgM is the first isotype produced (Janeway Jr *et al.*, 2001), so younger individuals who have not had repeated exposure to hCCoV will have higher levels of IgM than of other isotypes. It has also previously

been shown by Gorse *et al.* (2020) that older individuals express higher levels of antibodies for hCCoV than younger individuals due to repeated exposure. This also would help explain the differences in hCCoV antibody responses.

Gender and race may also be contributing factors in antibody responses to hCCoV. Figure 3.18 presents box plots comparing responses between female and male individuals in the cohort with baseline hCCoV samples. The plots in Figure 3.18 show that females presented significantly higher titres of antibodies of isotype IgM for all four hCCoV strains. Additionally, females also expressed higher levels of IgA antibodies with respect to strains OC43 and NL63. Males presented higher levels of antibodies for 229E (p-value < 0.01). Appendix Figure A.1 shows the difference in antibody responses between people of different races. The predominant differences in responses are for 229E and OC43 for IgG antibodies and HKU1, NL63, and OC43 for IgA. Most notably, responses for Black individuals for isotypes IgG for 229E and OC43 are lower than White or Asian, while IgA antibodies for HKU1, NL63, and OC43 are higher than Asian or white individuals.

Since the participants of this study work in a paediatric hospital, and their interactions are with children, this may lead to increased exposure to hCCoVs. It is interesting to see if there is any relation between patient contact and levels of antibodies. Figure 3.19 illustrates the differences in antibody responses compared to the status of patient contact. It shows that participants with direct contact had higher levels of IgM antibodies for all four strains of hCCoV, with a more significant amount than those who had no patient contact whatsoever. Figure 3.19 also shows that direct contact could result in higher levels of OC43 IgA than those without contact. However, there was no significant difference between individuals with indirect contact. If this data is taken together with data on gender response, it would suggest that younger female participants with direct contact are more likely to have elevated levels of hCCoV IgM antibodies.

Endemic hCCoV immune response after SAR-CoV-2 infection and vaccination

Previous studies such as those presented by Ng *et al.* (2020) and Ladner *et al.* (2021) have shown that hCCoV antibodies cross-react with SAR-CoV-2, which
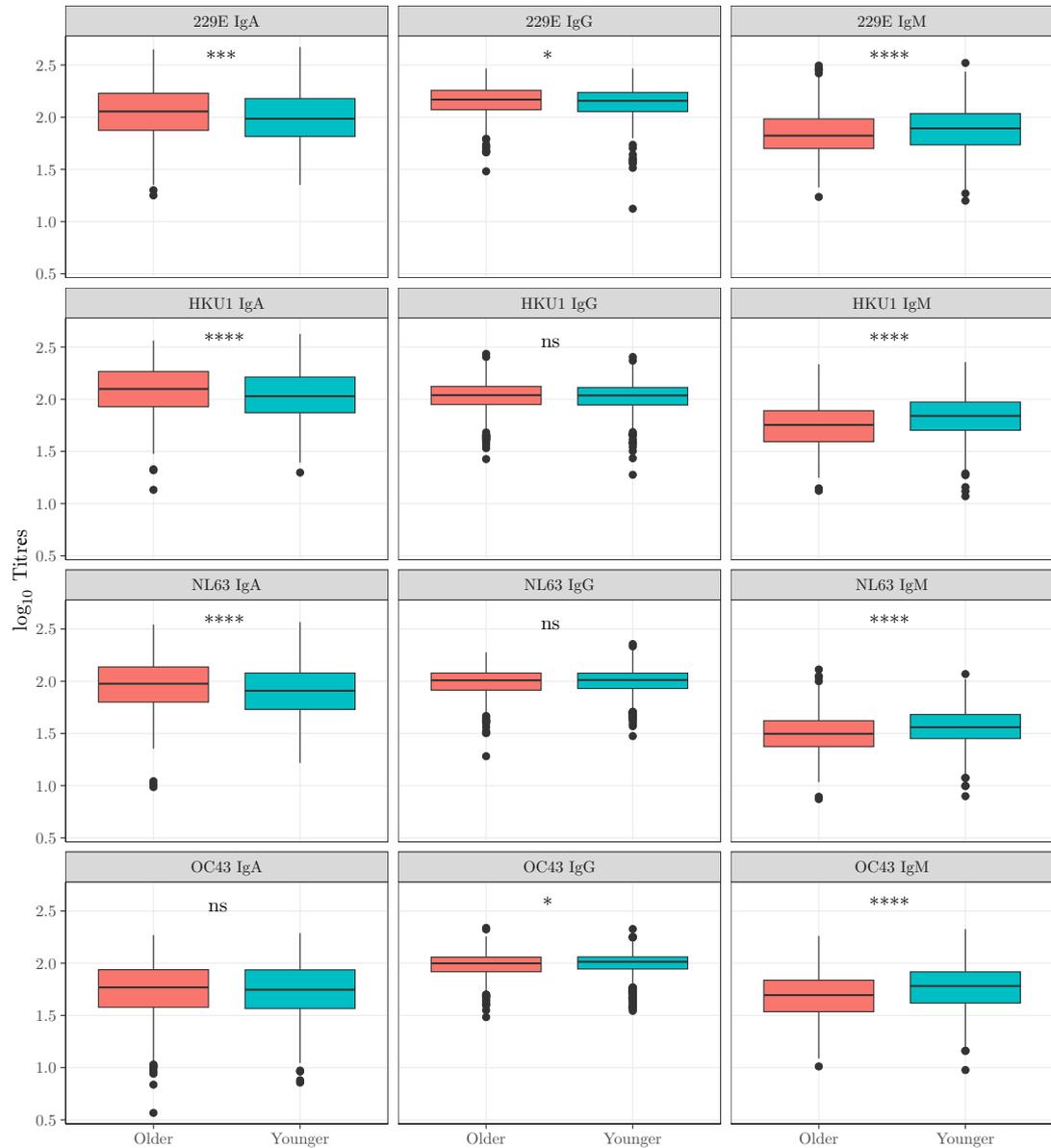
Figure 3.17: Box plots of hCCoV antibody isotype response compared between younger (< 43 years) and older (≥ 43 years). Statistical significance was determined by the Wilcoxon–Mann–Whitney test with Bonferroni adjustment (ns, not significant; $^*p < 5 \times 10^{-2}$; $^{**}p < 10^{-2}$; $^{***}p < 10^{-3}$; $^{****}p < 10^{-4}$).
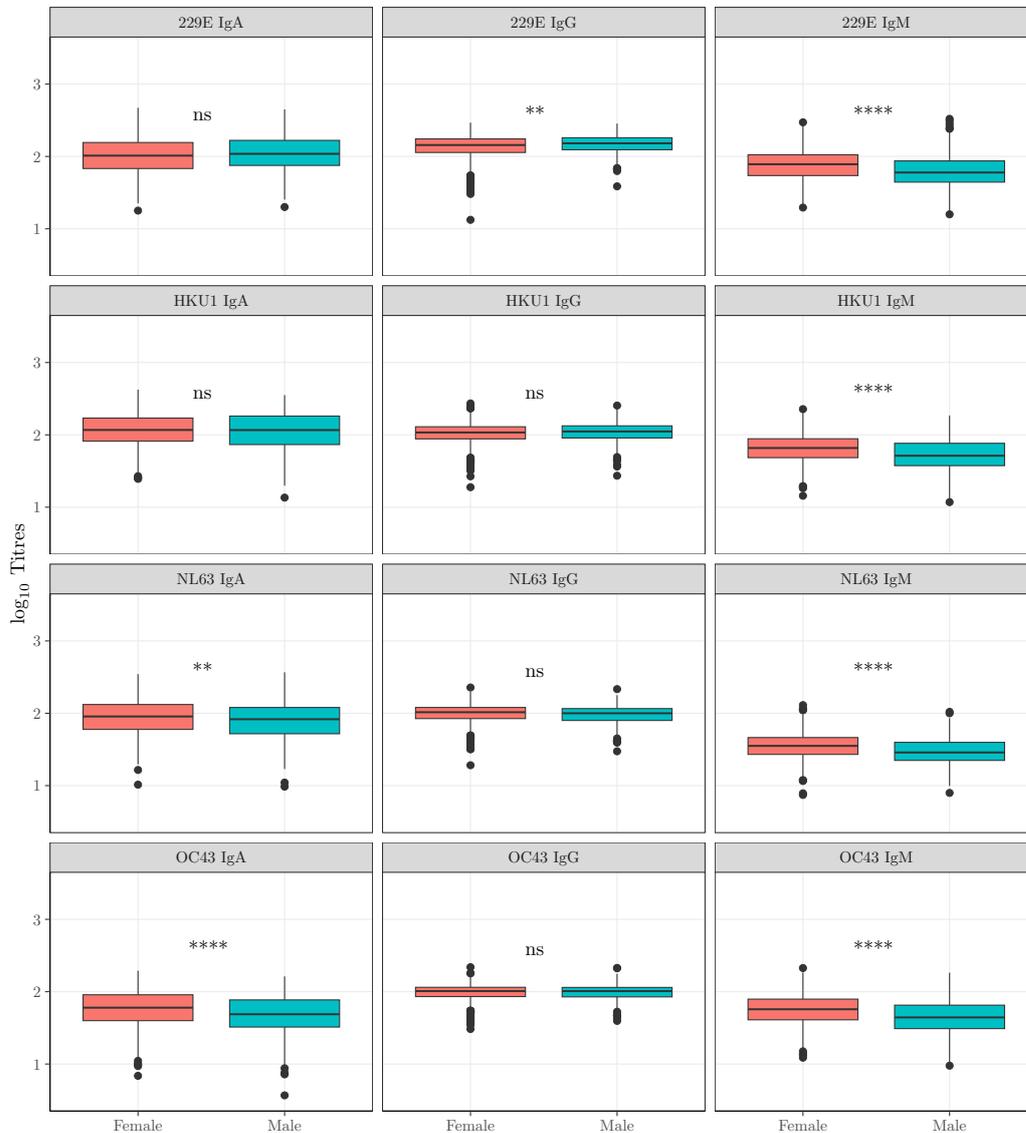
Figure 3.18: Box plots of hCCoV antibody isotype response compared between female and male individuals. Statistical significance was determined by the Wilcoxon–Mann–Whitney test with Bonferroni adjustment (ns, not significant; $*p < 5 \times 10^{-2}$; $**p < 10^{-2}$; $***p < 10^{-3}$; $****p < 10^{-4}$).
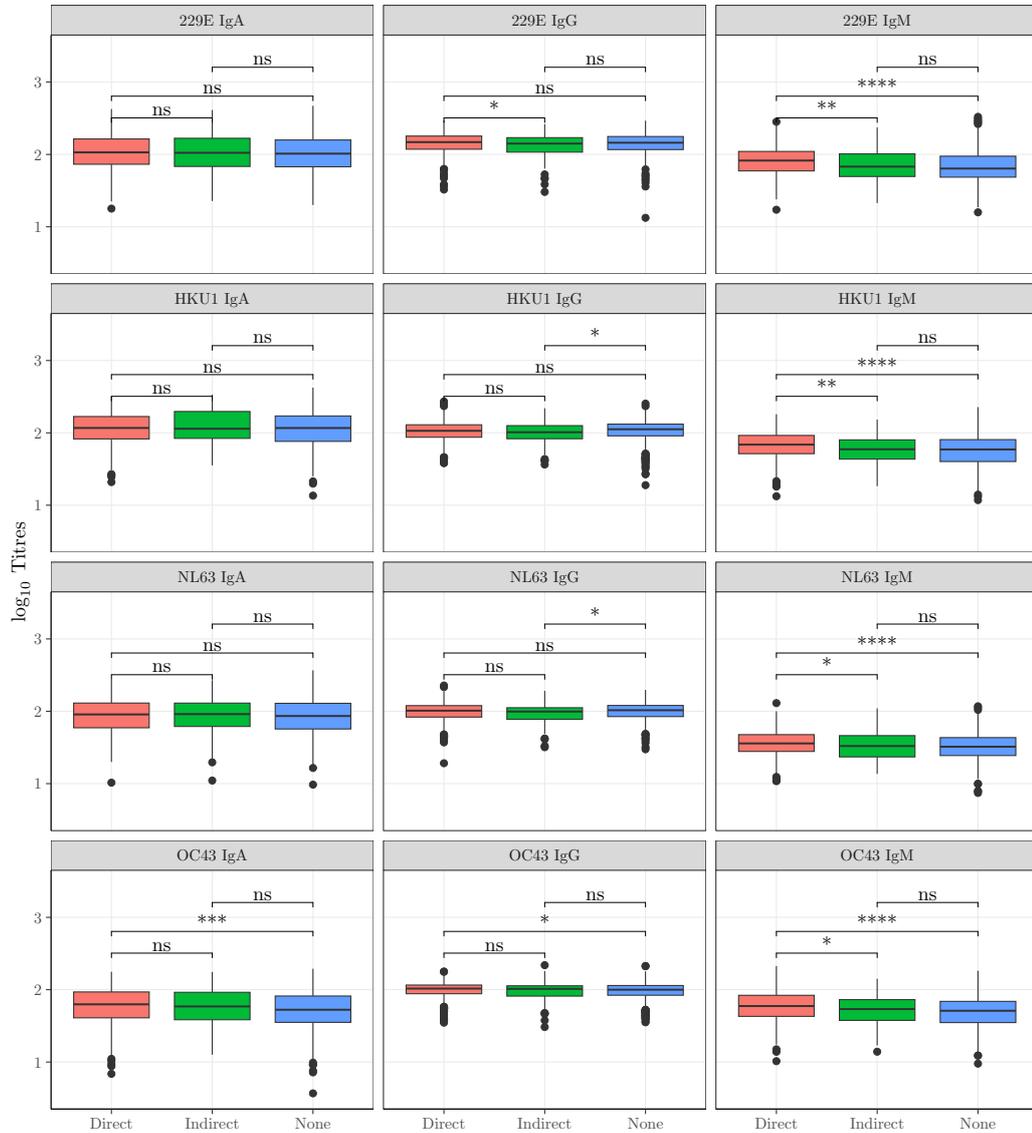
Figure 3.19: Box plots of antibody responses between individuals with direct, indirect or no patient contact. Statistical significance was determined by the Wilcoxon–Mann–Whitney test with Bonferroni adjustment (ns, not significant; $*p < 5 \times 10^{-2}$; $**p < 10^{-2}$; $***p < 10^{-3}$; $****p < 10^{-4}$).

is illustrated as an increase in antibodies early after infection. Therefore it is important to see whether there is any change in levels of hCCoV antibodies before infection with SARS-CoV-2 and hence any further cross-reactivity and potential protection from infection. To examine this, baseline samples of individuals with no previous SARS-CoV-2 infections ($n = 1100$) were compared against those who became infected during the study ($n = 151$). Figure 3.20 shows box plots for these two groups. From Figure 3.20, it can be seen that for most hCCoV antibodies, there is no significant change in the expressed levels after infection compared to before. However, both HKU1 and OC43 for IgG are expressed in higher titres after infection than before. Since both OC43 and HKU1 are β-coronaviruses, this could highlight the potential cross-reactivity seen by others (Ng *et al.*, 2020). Furthermore, this could also mean some protection provided by hCCoV antibodies against SARS-CoV-2 infection. It is also interesting that those individuals who never became infected express higher levels of IgM for 229E than those who became infected. The fact that many of the hCCoV antibodies do not change is consistent with previous studies that show hCCoV antibodies cannot neutralise SARS-CoV-2 (Aguilar-Bretones *et al.*, 2021).

Within the cohort, any individual who became infected during the study was asked to take a questionnaire regarding their symptoms and severity. Most of the cohort reported mild-moderate disease severity, with four reporting severe or critical severity. Previous studies such as that by Garcia-Beltran *et al.* (2021) have shown that levels of IgA and IgG for SARS-CoV-2 spike and receptor binding domain (RBD) following infection correlate with disease severity. This could be because individuals who struggle to control the virus due to poor immune responses will express higher viral titres. Therefore, there is greater antigen exposure, and responses could be an additional indicator of disease severity separate from objective self-reported severity. Consequently, it seems necessary to investigate if antibody responses correlate with disease severity. People who had samples taken 16-40 days post SAR-CoV-2 infection, with no vaccinations, were used, such that only naive infections were considered. Kendall rank-correlation coefficient was used to calculate the correlation between severity scores and antigen titres and are illustrated in Figure 3.21(A). A false discovery rate corrected p-values. The data in this figure suggests, similar to other studies, that the level of IgG specific
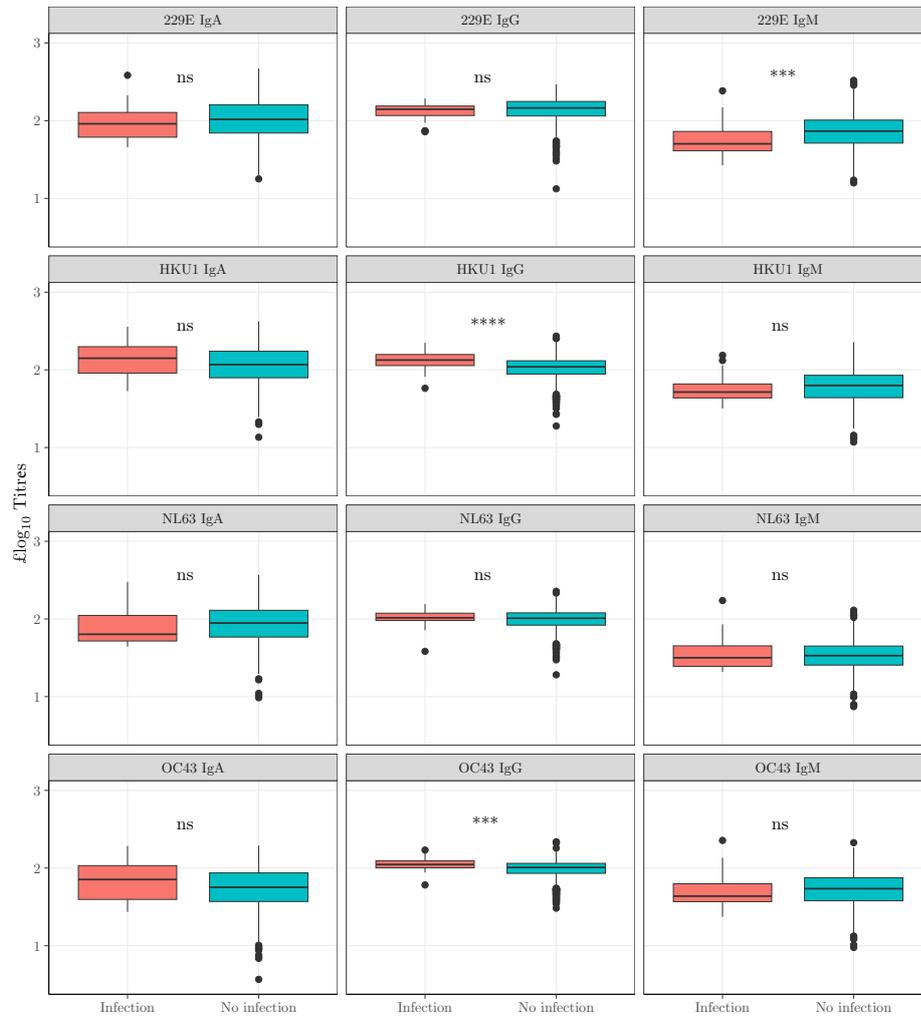
Figure 3.20: Baseline hCCCoV-normalized were compared between individuals that became infected ($n = 151$) during the study to individuals that remained SAR-CoV-2 negative ($n = 1,100$) using the Wilcoxon-Mann-Whitney test and adjusted with Bonferroni method (ns, not significant; *$p < 5 \times 10^{-2}$; **$p < 10^{-2}$; ***$p < 10^{-3}$; ****$p < 10^{-4}$).

for SAR-CoV-2 spike, RBD and N protein have a weak positive correlation with increased disease severity (p-value < 0.05 after correction). Furthermore, for IgM, higher levels with respect to spike and RBD suggest increased severity, along with levels of IgA for spike. This information indicates that isotypes IgG and IgM for SARS-CoV-2 correlate with disease severity. To further investigate this, one can examine if there are any correlations between baseline hCCoV antibody levels and SARS-CoV-2 specific antibodies 16-40 days post-infection. The result of this analysis is summarised in Figure 3.21(B). Notably, higher levels of OC43 IgG before infection positively correlate with spike IgG antibodies. This information, taken with the previous knowledge on the correlation between spike IgG and symptom severity, indicates that OC43 may be associated with increased disease severity.

In addition to studying correlations between hCCoV and SARS-CoV-2 antibody levels, one can also see if there is any relation between the magnitude change in hCCoV responses before and after infection to SARS-CoV-2. A change in hCCoV antibodies expressed in the early period of infection could be indicative of B-cell activation that results in the production of new antibodies. Therefore, the percentage change before and after infection in hCCoV antibodies was calculated for samples within the first 15 days of infection. An increase in this period would correspond to B-cell activation. Figure 3.21(C) shows an increase in β-coronavirus IgG, and IgA is associated with higher levels of SARS-CoV-2 IgG and IgM antibodies in early infection. Furthermore, since SARS-CoV-2 IgG and IgM levels are also correlated with severity, as stated earlier, this increase in magnitude could be associated with increased disease severity. However, these increases could also be due to the cross-reactivity of antibodies for hCCoV with SARS-CoV-2.

Since there exist correlations between the increase in β-coronavirus antibodies and SARS-CoV-2 in the early stages of infection, potentially due to newly generated antibodies for SARS-CoV-2 that cross-react with hCCoV, then one would expect to see a similar trend in response post-vaccination for individuals who have had no other previous exposure to the virus. If no such trend exists, it would suggest that the trends seen in infected people correlate with severity rather than cross-reactivity. Figures 3.21(D) and 3.21(E) show the relationships between SARS-CoV-2 antigen response to hCCoV baseline measurement and hCCoV percentage
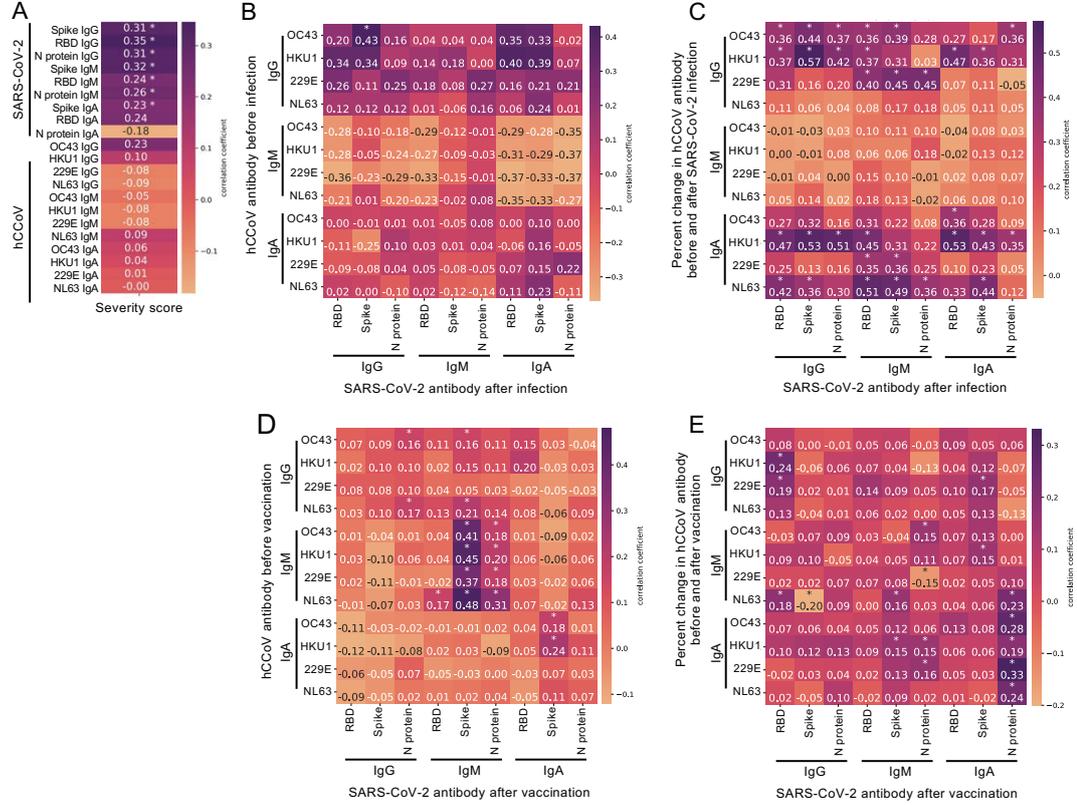
Figure 3.21: (A) The normalised OD of antibodies in samples taken 16–40 days after SARS-CoV-2 diagnosis ($n = 123$) was compared to the five severity scores. Kendall rank correlation coefficients are indicated in the heatmap. A false discovery rate corrected p-values. *$p < 0.05$. (B–E) Pearson's formulation was utilised to calculate correlation coefficients, with multiple testing corrections with the TestCor package between (B) normalised ODs of baseline hCCoV antibodies compared to normalised ODs of SARS-CoV-2 antibody in samples collected 16–40 days after infection ($n = 41$), (C) the per cent change from baseline of hCCoV antibodies compared to SARS-CoV-2 antibody in samples collected between 1 and 15 days after infection ($n = 43$), (D) baseline hCCoV-normalized ODs compared to SARS-CoV-2 antibody 20–85 days after vaccination with Pfizer/BioNTech BNT162b2 ($n = 256$), and (E) the increase in hCCoV antibodies relative to the baseline sample compared to SARS-CoV-2 antibody in samples collected after vaccination ($n = 256$).

change respectively. From Figures 3.21(D) and 3.21(E) it can be seen that neither baseline hCCoV antibodies nor an increase in hCCoV antibodies after vaccination correlated with SARS-CoV-2 antibodies after vaccination. Figures 3.21(B)-(D) highlight stark differences in the hCCoV responses to SARS-CoV-2 post-infection and vaccination. It is worth highlighting there were significant correlations between baseline hCCoV IgM response and SARS-CoV-2 spike IgM responses after vaccination. IgM antibodies are typically more cross-reactive within hCCoV compared to IgG or IgA (Chan *et al.*, 2005), and hence there could be cross-reactivity for IgM for SARS-CoV-2 since they are of the same family. However, it has been shown that vaccination does not induce a robust IgM response in many individuals (Banga Ndzouboukou *et al.*, 2021), so little is known about the effects on vaccine efficacy. Since there is no increase in vaccinated individuals for IgA or IgG responses, these results may indicate that pre-existing hCCoV antibody responses have a negative impact on SARS-CoV-2 immune response and contribute to increased severity.

SARS-CoV-2 natural infection compared to vaccination

In the previous section, relationships between pre-existing hCCoV antibodies were compared to SARS-CoV-2 antibodies both after infection and vaccination. However, there was no investigation into the difference in SARS-CoV-2 antibody levels after infection compared to vaccination. With this St Jude cohort, the main two vaccines administered to participants were Pfizer/BioNTech and Moderna vaccines, both of which are mRNA vaccines targeted at the spike (S) protein of SARS-CoV-2 (Martinez-Flores *et al.*, 2021). Since vaccines only contain a fraction of the total genome, responses to parts that are not the S protein should not exist in vaccinated individuals than in those who suffer a natural infection. To investigate this, individuals who had an infection with no prior vaccination ($n = 159$) were compared to those who had two jabs for a full vaccination ($n = 744$). The following analyses will be for isotypes IgM, IgG and IgA with respect to the S protein, RBD and N protein so that, in total, 9 variables are examined. First, to visually inspect whether there were any differences between the two groups, principal component analysis was used to reduce the number of variables.
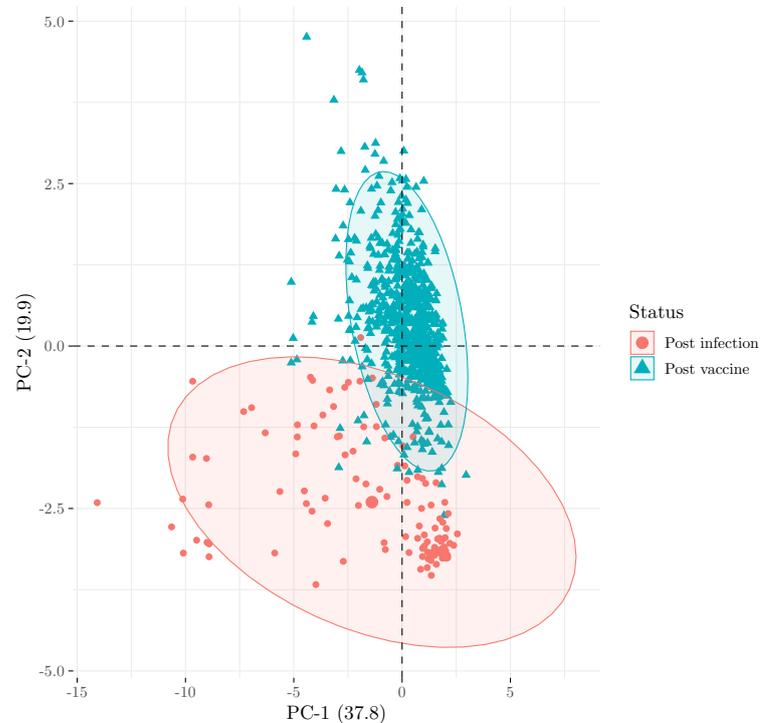
Figure 3.22: PCA for natural infection against vaccination for the first two PCs. The first PC accounts for 37.8% of the total variance, while the second PC accounts for 19.9% of the variance.

Figure 3.22 illustrates the first and second PC, which account for 37.8% and 19.9% of the variance, respectively. By Kaiser's rule, only the first three principal components would be kept and would account for 69.9% of the total variance.

Figure 3.22 shows two clear, distinguishable groups formed in the first two PCs, with a slight overlap. Infected individuals are shifted predominantly down from the origin compared to vaccinated individuals. Using the ε-test as introduced in Section 2.2.3 to investigate whether the differences in these groups on a multivariate level are significant, I compare infected against vaccinated individuals, taking into account all antibody responses. The results from this test give a p-value of 9.9×10^{-4}, which indicates the responses from these groups are statistically significantly different. A question that remains to be answered is why are these groups so different.
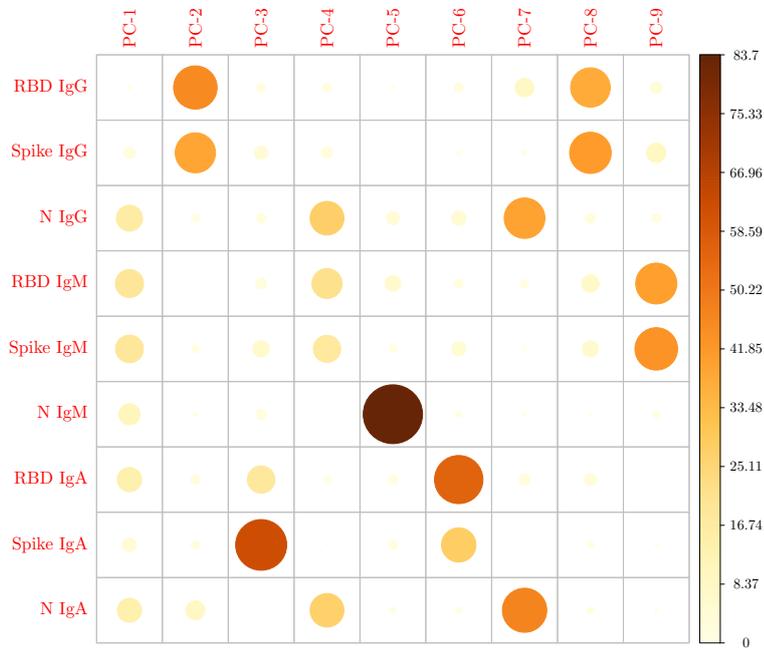
Figure 3.23: Contribution by each antibody response for isotypes IgG, IgM and IgA with respect to SARS-CoV-2 RBD, S protein and N protein to each PC.
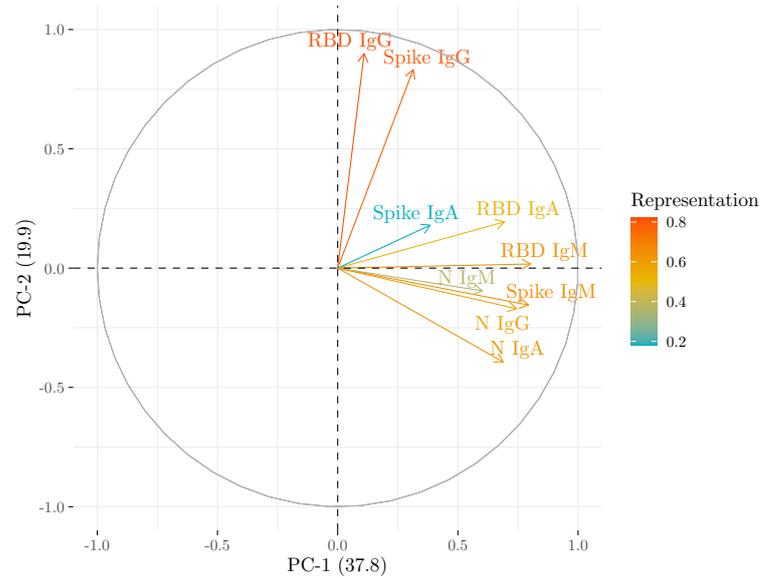
Figure 3.24: Correlation circle of each antibody isotype IgG, IgM, IgA with
respect to SARS-CoV-2 RBD, S protein and N protein. The angle between
variables represents the correlation of variables with the length of the line from the
origin, representing the quality of representation. The PCs in this figure match
those presented in Figure 3.22.

Figure 3.23 presents the contribution of each variable to each component. For
PC-1, the main contributions come from antibodies IgG, IgM for all SARS-CoV-
2 proteins and N protein for IgA. As mentioned previously, the most notable
difference in Figure 3.22 is a shift in the second PC, for which IgG antibody
responses for RBD and S protein are the main contributors.

The correlation circle plotted in Figure 3.24 shows that SARS-CoV-2 IgG
antibodies for RBD and S protein are well represented in the first two principal
components. Furthermore, this figure also shows by the small angle between RBD
IgG and Spike IgG that these two variables have a strong positive correlation
indicating high levels of IgG towards spike are expressed alongside high levels of

IgG to RBD. Spike IgM and N IgG are also highly correlated with one another, and there are also weaker positive correlations between these variables and N for all isotypes, RBD for IgA, and spike IgA. Spike IgA and N IgM have the worst quality of representation in the first two principal components. Since the second principal component is where the most significant difference between groups lies, it is worth examining on a uni-variate level the antibody responses to RBD and S for IgG.

Figure 3.25 shows responses for SARS-CoV-2 RBD, spike, and nucleoprotein for each antibody isotype for infected individuals against vaccinated individuals. It can be seen that individuals who have been immunised with no previous infection express higher levels of IgG for the RBD and spike protein (p-value< 0.001). Additionally, vaccinated individuals express higher IgA levels for the spike than infected individuals. Isotype IgM is expressed in higher amounts for natural infections against all three tested SARS-CoV-2 proteins and is also higher for IgA with respect to the N protein and RBD. The differences in responses are likely because the vaccine only contains mRNA for the spike protein in the two vaccines considered. Hence naturally infected individuals will express higher antibody levels for proteins not included in the vaccine. The differences in IgM response could be explained by the fact that vaccinated people here are considered only if they have two vaccinations. Therefore, since IgM responses are generated first before the production of IgG and IgA, it is likely that this part of the response is through the initial production stage, and the second dose of vaccine boosts IgG responses. This is in agreement with Cho *et al.* (2021) who also found lower expression of IgM and IgA after vaccination but higher levels of IgG.

In the previous section, when examining hCCoV responses, it was shown there exists a difference in responses for people of different ages and races. It is worth examining this demographic information now for people infected with SARS-CoV-2 and vaccinated individuals. When examining the responses of infected individuals separated by race shown in Appendix Figure A.2, it can be seen there are no notable differences in immune responses. However, when this is examined for vaccinated individuals against their race (Figure 3.26), it can be seen for a few isotypes, there is a difference in immune responses depending on the SARS-CoV-2 protein under consideration. For the RBD, IgA and IgM levels

are higher in Asian and Black individuals than White individuals. While for IgA, the antibody titre with respect to the N protein and Spike protein in Asian individuals is higher than in White individuals but not for Black individuals. It is interesting to note that antibody responses for IgG are similar between all races. Since this is the most elevated after vaccination (Cho *et al.*, 2021), it is reassuring to see little variation between groups suggesting an equal level of response. When examining whether age affects responses, there is no significant difference in antibody responses of infected individuals (p-value> 0.05). However, in vaccinated individuals (Figure 3.27), there is a significant difference in responses for IgM with "Younger" individuals expressing higher levels of IgM compared to "Older". This is similar to antibody responses to hCCoV and is most likely because "younger" individuals can produce a more robust IgM immune response to new infections than "older" individuals. "Younger" individuals also express higher levels of IgA for the RBD of SARS-CoV-2. This does not follow the trend seen for hCCoV and could be a result of younger individuals typically possessing stronger immune systems and hence producing more antibodies in response to vaccination.

Discussion

In this subsection, I focused on the statistical analysis of antibody responses to hCCoV and SARS-CoV-2 in a large cohort study undertaken by St. Jude Children's Research Hospital. It has been shown a virus that has antigenic similarity to a virus from a previous infection can lead to activation of memory B cells instead of a naive primary reaction (Guthmiller & Wilson, 2018). As a result, this means that for viruses with antigenic similarity, there is a chance that antibodies may cross-react and, as such, could provide some level of immunity to new novel viruses. On the other hand, cross-reactivity could also hinder the immune response since antibodies could bind to conserved sites, inhibiting newly formed antigens, and potentially better binders, from binding and taking effect (Cobey & Hensley, 2017). Since SARS-CoV-2 is a β-coronavirus and there are four endemic coronaviruses in humans, there is potential for cross-reactivity between endemic hCCoVs antibodies and SARS-CoV-2.

3.2 St Jude Children's Research Hospital SARS-CoV-2 trace study

The main goal of this work was to understand the differences in antibody responses for both hCCoV and SARS-CoV-2 after infection with SARS-CoV-2 or vaccination and investigate any potential cross-reactivity. Several studies have investigated this and found conflicting results (Aydillo *et al.*, 2021; Dugas *et al.*, 2021; Shrock *et al.*, 2020). However, these studies lacked baseline samples before either an infection or immunisation. Therefore this study allowed the unique perspective to examine changes in hCCoV antibody responses and see if these related to SARS-CoV-2 antibody response. Furthermore, since this study had 1266 participants, demographic information was collected, allowing investigation into how demographics can impact antibody responses to both SARS-CoV-2 and hCCoV.

Humans are repeatedly exposed to hCCoV (Edridge *et al.*, 2020). As such, in our cohort, the majority express antibodies for one or more hCCoV strains when tested, and as such, there is potential for cross-reactivity. Figure 3.21 shows that hCCoV boosted antibody response correlates with increased antibodies for SARS-CoV-2. Hence, this indicates that increased hCCoV antibodies correlate with disease severity. This agrees with the results presented by Dugan *et al.* (2021), which also showed that an increase in non-neutralising antibodies did not provide protection from SARS-CoV-2. However, when examining vaccination, there were no significant increases in antibody responses for IgG and IgA, only for IgM. Since IgM is highly cross-reactive and little is known about IgM responses in vaccination, this data suggested that pre-existing hCCoV immune responses have a negative impact on SARS-CoV-2 response and further cement that higher levels of hCCoV antibodies may contribute to disease severity. Furthermore, it was also shown that infected individuals expressed higher levels of IgG for HKU1 and OC43 than uninfected individuals, which supports cross-reactivity. However, this cross-reactivity does not provide additional protection when taken together with the previous information.

Age plays a role in the antibody responses for hCCoV, with "younger" individuals expressing higher titres of IgM antibodies and "older" individuals expressing higher levels of IgA and IgM. This is due to the specificity of responses increasing with age due to repeated exposure to hCCoV. Meanwhile, younger participants

have not had the same repeated exposure resulting in higher levels of IgM antibodies which are highly cross-reactive and the first antibody isotype produced. Regarding post-vaccine responses, as with hCCoV antibody responses, "younger" individuals expressed higher IgM levels than "older" individuals. This shows that "younger" people can develop higher levels of IgM due to vaccination than "older" people. Since this isotype is cross-reactive, this may provide some additional protection from infection with other strains of SARS-CoV-2. For hCCoV, it was shown that IgG antibody responses were lower for Black individuals when compared to White or Asian individuals but higher for IgA(Figure A.1). This pattern is not seen for post-infection results, where most responses are not significantly different. The post-vaccination results are similar, with Black individuals expressing higher levels of IgA for the RBD compared to White individuals. Furthermore, Asian individuals also expressed higher IgA levels compared to White individuals. This information indicated that demographic factors could play some role in the vaccine responses and responses to circulating endemic hCCoV strains.

One of the most exciting aspects of this study was the ability to investigate the antibody responses of people post-infection without any prior infection compared to vaccinated individuals without any previous virus exposure. It was shown in Figure 3.22 at a multi-variate level that there are apparent differences between responses in these two groups, most of which are explained by differences in response to the RBD and Spike protein for IgG. When investigating on a uni-variate level, vaccinated individuals expressed significantly higher levels of IgG for both the RBD and Spike along with Spike IgA. However, IgM with respect to N, Spike, and RBD are much higher in naturally infected individuals along with IgA for the N protein, and RBD in agreement with Cho *et al.* (2021). This shows that naturally infected individuals have a broader response for all three potential targets in all investigated isotypes. In contrast, vaccinated people have incredibly high responses for a select few potential targets. This data indicated that possible antigen-binding site changes could reduce vaccine effectiveness and increase severity since responses to other SARS-CoV-2 proteins are not so strong.

As previously mentioned, the unique perspective of this study was that many individuals had a baseline measurement and a secondary measurement to investigate the effects of hCCoV. However, this study also had many limitations that

may reduce the results' power. The first limitation was the ages of the participants. The youngest individual in the cohort was 20 years old, so responses from infants through teenagers are not accounted for, while there are only 41 individuals 65+ years old. Therefore, the study does not fully cover the average population's distribution of ages; thus, some conclusions drawn from this study may not hold. Another limitation, as mentioned, is severity scores. Not only did the study not include many individuals who had severe disease, so correlations on severity may be inaccurate, but also severity was self-reported, so there exists the possibility of bias being introduced. Another potential issue is that most of the cohort is female (74%), so male responses may not be well represented.

The data here demonstrates that SARS-CoV-2 infection and vaccination activate existing memory B cells, specific for hCCoVs. Baseline levels of hCCoV antibodies and the magnitude of which these antibodies increased after infection or vaccination varied dramatically among individuals. Higher baseline levels or an increase of β-coronavirus IgG and IgA after infection were associated with increased SARS-CoV-2 antibody levels, correlated with greater disease severity. Moving forward, several additional questions would be of interest to investigate. The first is the longitudinal responses of infected and post-vaccination individuals. These individuals could be repeatedly measured to study long-term and potential waning immunity and determine the antibody decay rate in both scenarios. Since there are now multiple strains of SARS-CoV-2, the breadth of the immune response given vaccination or infection to a particular strain for all these strains would be essential to investigate (CDC, 2022). This would allow further investigation into cross-reactive antibodies amongst the strains and may indicate when subsequent booster vaccines may be needed for a new strain. Finally, additional samples of people with multiple vaccinations would allow investigation of antibody boost or further specificity of antigens as seen in older individuals exposed to repeated hCCoV infections.
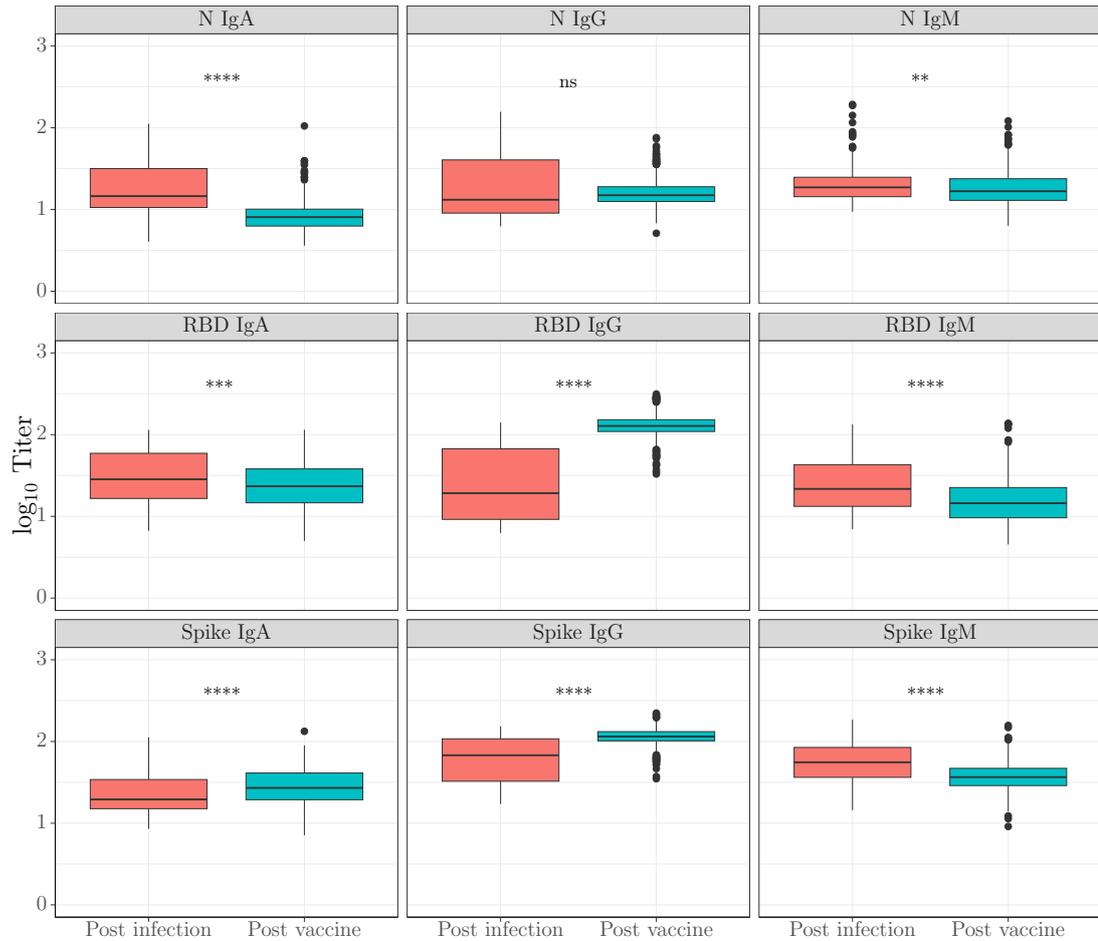
Figure 3.25: Box plots for each antibody isotype, with respect to SARS-CoV-2 RBD, S and N protein for individuals infected with no vaccine ($n = 159$) and people vaccinated with no infection ($n = 744$). Statistical significance was determined by the Wilcoxon–Mann–Whitney test with Bonferroni adjustment (ns, not significant; $*p < 5 \times 10^{-2}$; $**p < 10^{-2}$; $***p < 10^{-3}$; $****p < 10^{-4}$).
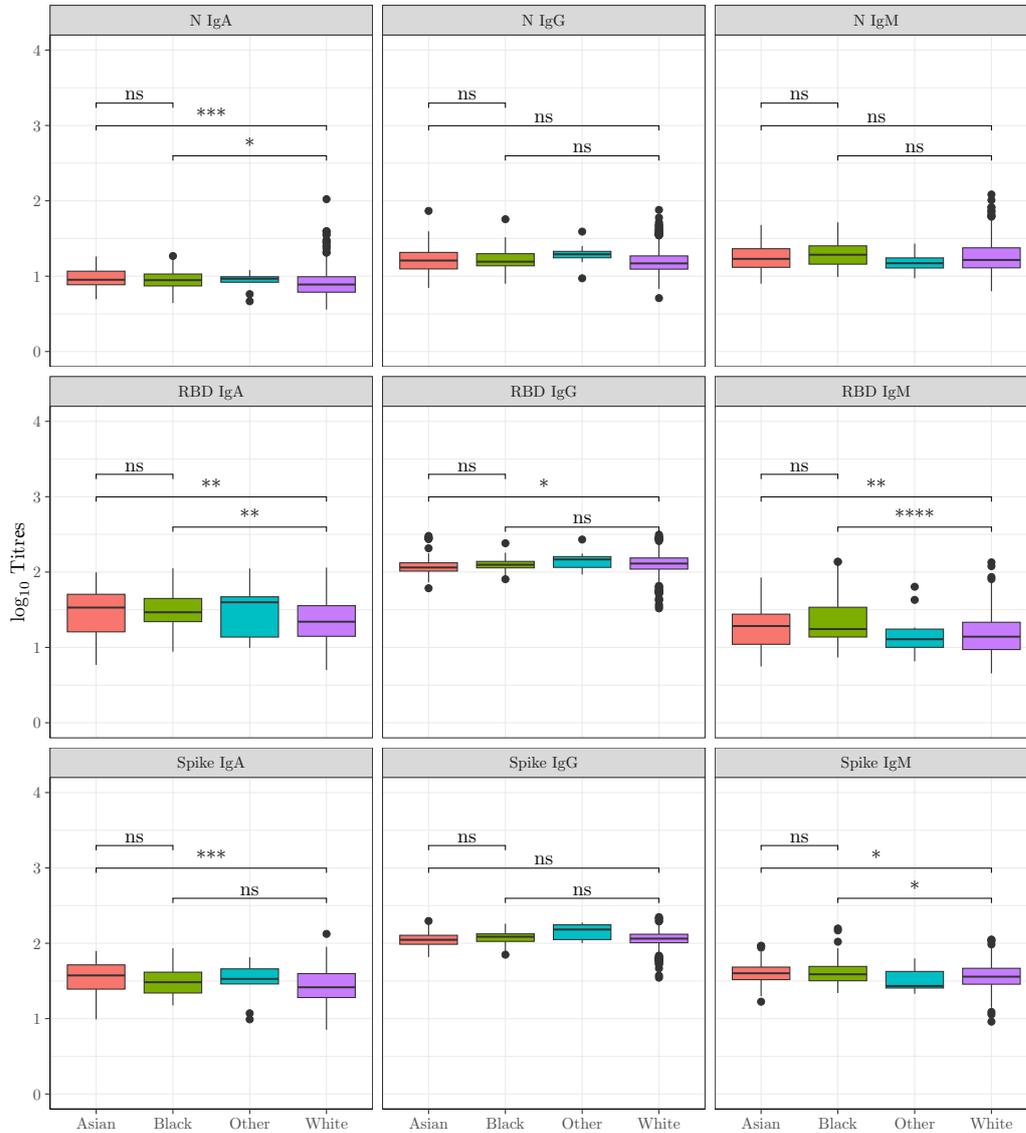
Figure 3.26: Box plots for each antibody isotype, with respect to SARS-CoV-2 RBD, S and N protein for different races. Statistical significance was determined by the Wilcoxon–Mann–Whitney test with Bonferroni adjustment (ns, not significant; $*p < 5 \times 10^{-2}$; $**p < 10^{-2}$; $***p < 10^{-3}$; $****p < 10^{-4}$).
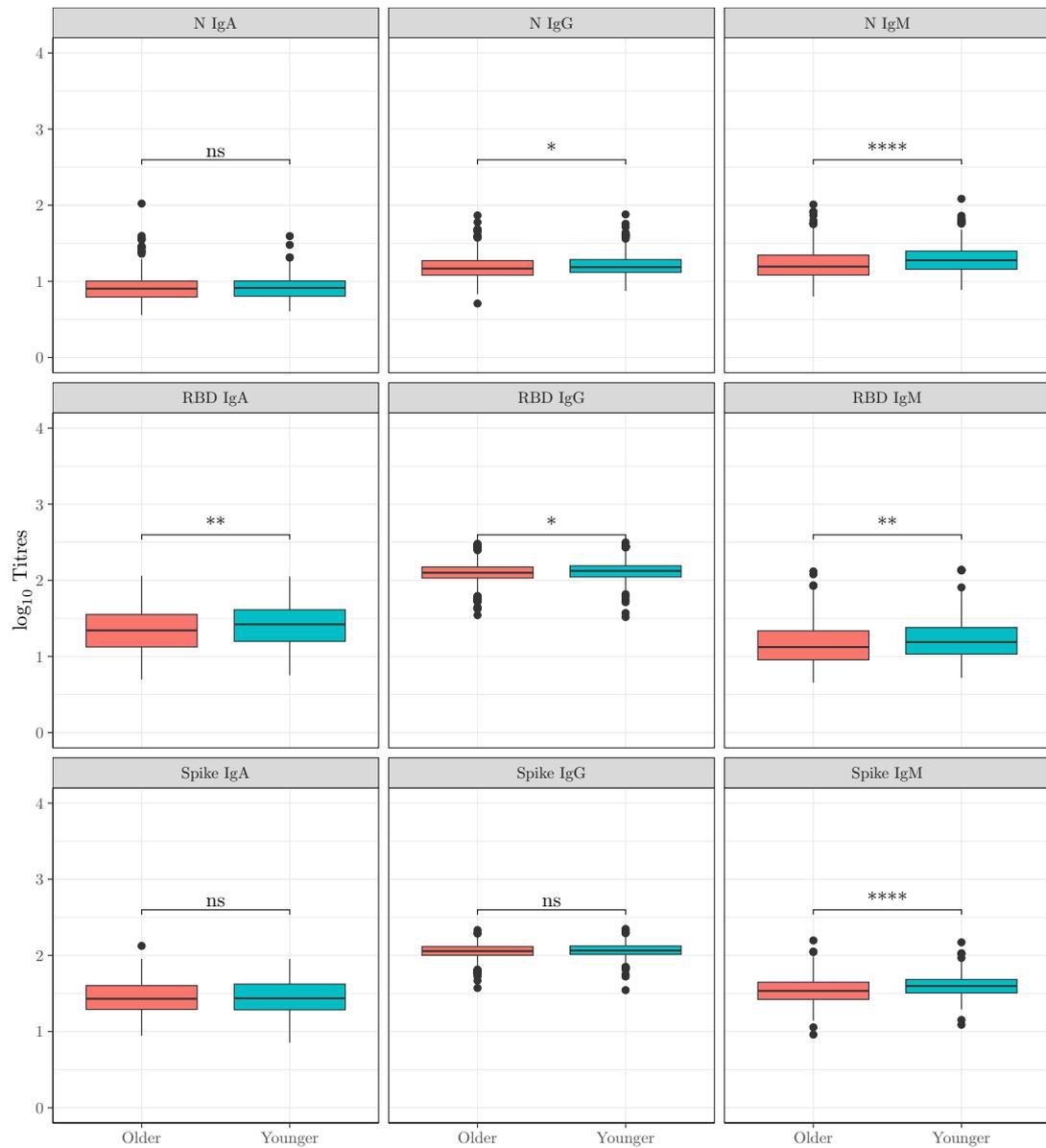
Figure 3.27: Box plots of SARS-CoV-2 antibody isotype response compared between younger (< 43 years) versus older (≥ 43 years) individuals based on the median age of the cohort. Statistical significance was determined by the Wilcoxon–Mann–Whitney test with Bonferroni adjustment (ns, not significant; *$p < 5 \times 10^{-2}$; **$p < 10^{-2}$; ***$p < 10^{-3}$; ****$p < 10^{-4}$).

Chapter 4

Mathematical model of interferon antagonism

The majority of work in this chapter is published in Locke *et al.* (2021). As highlighted in Chapter 1 we are exposed to a diversity of pathogens throughout our lives. Our bodies have evolved to provide us with intricate molecular and cellular mechanisms to sense, prevent and respond to such infections. Cells in the first line of protection, associated with the innate immune system, are equipped with pattern-recognition receptors (PRRs) that sense pathogen-associated molecular patterns (PAMPs), such as viral RNA (Ivashkiv & Donlin, 2014). Activation of PRRs in infected cells leads to the secretion of type I interferon (IFN), the main anti-viral cytokine (Escudero-Pérez & Muñoz-Fontela, 2019; Perry *et al.*, June 2005; Randall & Goodbourn, 2008; Trinchieri, 2010). The binding of type I IFN to its receptor, in turn, induces the transcription of a family of interferon-stimulated genes (ISGs), whose protein products have both anti-viral activity and immuno-modulatory effects (Audsley & Moseley, 2013; Escudero-Pérez & Muñoz-Fontela, 2019; Randall & Goodbourn, 2008).

The survival of a viral population in a host depends on viruses replicating and avoiding intra-cellular host defences. Many viruses have developed strategies to evade immune detection, thus, subvert these defences (Katze *et al.*, 2002). There exists a great diversity of such viral strategies. Here I will only consider those mechanisms that interfere with intra-cellular pathways to regulate type I IFN secretion or type I IFN signalling.

4. MATHEMATICAL MODEL OF INTERFERON ANTAGONISM

Filoviruses, such as EBOV and Marburg virus, encode viral proteins that can counteract type I IFN responses to replicate efficiently and minimise the therapeutic anti-viral power of IFNs. Filoviruses possess viral proteins to antagonise the IFN response due to their natural reservoir in bats. Bats possess highly effective innate immune responses, and without the antagonism of viral proteins, the virus would not be able to replicate and persist in bats (**?**). However humans who have a less robust innate immune system, these type I IFN antagonist proteins, or viral antagonistic proteins (VAPs), are essential to guarantee viral replication, prevent the type I IFN-induced anti-viral state in infected and bystander cells, as well as impair the ability of antigen-presenting cells to initiate adaptive immune responses (Messaoudi *et al.*, 2015). This ability of filoviruses to replicate *"unchecked"* by the host's innate anti-viral response can partly account for their lethality. Early innate immune evasion facilitates fast and excessive viral replication, which results in a delayed and damaging host immune response (Escudero-Pérez & Muñoz-Fontela, 2019). Unfortunately, filoviruses are not the only viral family to actively avoid immune surveillance. Other examples include influenza A virus, hepatitis B virus, and Bunyaviruses, such as Crimean-Congo haemorrhagic fever (CCHFV) or Rift Valley fever viruses (Weber *et al.*, 2002).

It is important to note that there is a difference, and even conflicting evidence, between responses to *in vivo* and *in vitro* infection models. In EBOV infection, for example, type I IFN production is abrogated after three days post-infection *in vitro*, yet for *in vivo* infection, type I IFN cytokines are secreted during the entire infective period (Bosworth *et al.*, 2017; Geisbert *et al.*, 2000; Kotliar *et al.*, 2020). It should also be highlighted that data availability in vivo is a challenge. Patients do not seek treatment until they are symptomatic, and as such, the innate immune response is typically missed in sampling. While mouse models exist, they are a poor substitute compared to non-human primates; however, non-human primates come with various ethical, cost, and bio-safety issues that must be considered.

In this chapter, I develop mathematical models of the intra-cellular molecular processes known to antagonise type I IFN production by viral proteins (Audsley & Moseley, 2013; Escudero-Pérez & Muñoz-Fontela, 2019). Existing mathematical models of intra-cellular production of type I IFN have many parameters that are difficult to estimate or do not account for viral protein antagonism of PRR

pathways (Rand *et al.*, 2012; Rinas, 2016; Zou *et al.*, 2010). There are also models that describe inter-cellular interactions via IFN α receptors (Schmid *et al.*, 2015). The goal is to model upstream and downstream mechanisms of viral protein antagonism and to provide a case study applied to EBOV.

Three stochastic models formulated as continuous-time birth-and-death Markov processes, are introduced for viral antagonism of type I interferon expression, with each model considering a different molecular mechanism of antagonism. Stochastic models can be relevant as some proteins may be expressed in low numbers, and I will additionally consider competition for resources as introduced by Reuter (1961). This will allow the investigation of stochastic descriptors, such as quantifying the time scales to reach a particular level of signalling molecule and the overall long-term behaviour of the system.

These proposed models will have several unidentified parameters that will require determination. It is easier to fit a deterministic model than a stochastic one. Therefore the linear noise approximation will determine a mean-field approximation for each model. This mean-field approximation will be used for parameter calibration. However, we must evaluate our models' sensitivity to variation in parameter values before determining these values. Sobol sensitivity analysis will identify which parameters require careful calibration within a given model to minimise the variance in the model output. Furthermore, assessment of parameter identifiability given the limited available data will inform which parameters can be obtained from the current data, covered in section 4.3. In section 4.4, the proposed models are calibrated with clinical data from *in vivo* EBOV infection of rhesus macaques (Kotliar *et al.*, 2020), where model selection allows comparison of the different biological hypotheses. The approach that will be presented could be used to quantify inhibition of type I IFN secretion by other pathogens, such as SARS-CoV-2 virus (Acharya *et al.*, 2020; Schultze & Aschenbrenner, 2021), Dengue and West Nile viruses (Chan & Gack, 2016), and Bunyaviruses (Elliott & Weber, 2009; Vaheri *et al.*, 2013).

4.1 Stochastic models

In this section, I introduce three potential stochastic models that aim to characterise the antagonism of type I IFN production, each incorporating a different biological hypothesis. Matrix analytical methods are presented in section 4.1.1 to investigate stochastic descriptors and how these can be constructed as levels in block matrix form.

4.1.1 A first model of Type I IFN inhibition by viral antagonistic proteins

A first mathematical model can be introduced to characterise the inhibition of type I IFN secretion by a viral antagonistic protein (VAP), which considers the role of the following proteins: RIG-I, viral RNA, VAP and TBK1, a protein kinase which coordinates the activation of interferon-regulatory factor (IRF) proteins (Audsley & Moseley, 2013; Escudero-Pérez & Muñoz-Fontela, 2019). The specific VAP will depend on the virus under consideration; for instance, if the virus is EBOV, then VAP is VP35, and in the case of Bunyaviruses, VAP is the non-structural protein NS (Weber *et al.*, 2002). Denote RIG-I by R, viral RNA by D, VAP by V, and TBK1 by B. Driven by current experimental evidence taken from literature, the following reactions (Audsley & Moseley, 2013; Escudero-Pérez & Muñoz-Fontela, 2019) are proposed as shown in Figure 4.1.

I will assume mass action kinetics in what follows. The first reaction describes RIG-I and viral RNA binding to form a RIG-I:RNA complex ($R : D$) with rate k_R, and unbinding with rate q_R. The second reaction describes VAP and viral RNA binding to form a VAP:RNA complex ($V{:}D$) with rate k_V, and unbinding with rate q_V. The last reaction describes activation, *i.e.*, phosphorylation, of TBK1 with rate k_B, and de-activation (de-phosphorylation) with rate q_B. Denote the activated B molecule by B^*. This model will now be referred to as *model 1*. Denote by n_R, n_D, n_V, and n_B, the per cell total number of RIG-I proteins, viral RNA molecules, VAP proteins and TBK1 proteins, respectively. Protein degradation or synthesis is neglected, so that for the timescale considered, the total number of molecules of a given species is conserved. Consider a continuous-time Markov
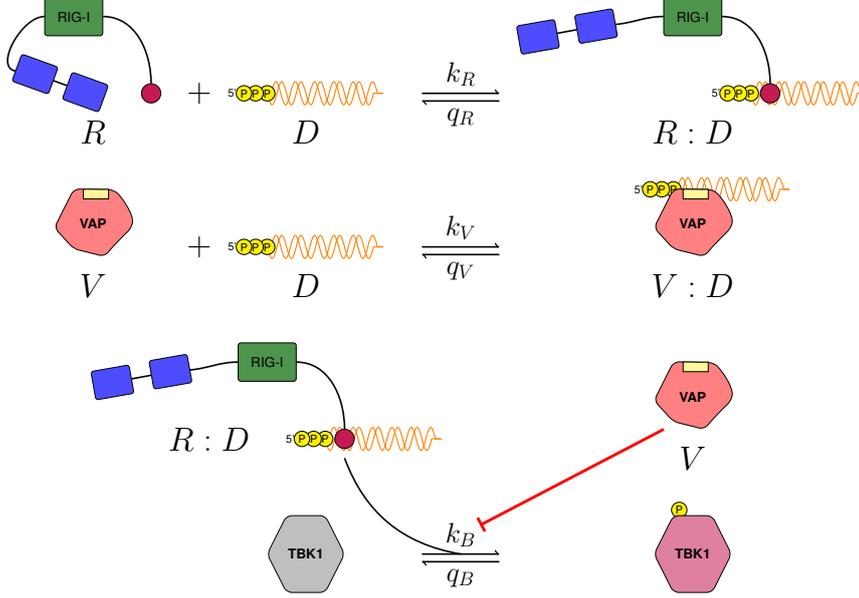
Figure 4.1: A first model of type I IFN inhibition by VAP (model 1). Model 1 considers the following molecules: RIG-I denoted R, viral RNA, denoted D, VAP denoted V, and TBK1 denoted B. In this model, there are six reactions and three molecular complexes.

process $\mathfrak{X}_1 = \{\boldsymbol{X}(t) = (n_{RD}(t), n_{VD}(t), n_{B^*}(t)) : t \geq 0\}$ where, the number of RIG-I:RNA (or $R:D$) complexes at time $t \geq 0$ is denoted by $n_{RD}(t)$, the number of VAP:RNA (or V:D) complexes by $n_{VD}(t)$, and the number of activated TBK1 (or B^*) complexes by $n_{B^*}(t)$ (Allen, 2007). Conservation of molecules implies that the number of free RIG-I molecules at any given time is given by $R(t) = n_R - n_{RD}(t)$, the number of free viral RNA is given by $D(t) = n_D - n_{RD}(t) - n_{VD}(t)$, the number of free VAP molecules is given by $V(t) = n_V - n_{VD}(t)$, and the number of un-phosphorylated TBK1 molecules is given by $B(t) = n_B - n_{B^*}(t)$. From the condition that $n_{RD}(t), n_{VD}(t), n_{B^*}(t) \geq 0$, and from the previous comments it can be shown that,

$$D(t) \geq 0 \implies n_{RD}(t) + n_{VD}(t) \leq n_D(t)$$
$$B(t) \geq 0 \implies n_{B^*}(t) \leq n_B(t),$$

for all $t \geq 0$, which specify the state space $\mathcal{S}_1$ of $\mathfrak{X}_1$. Given that $(n_{RD}(t), n_{VD}(t), n_{B^*}(t)) = (n_1, n_2, n_3)$ at time $t \geq 0$ then two situations arise $n_R + n_V \leq n_D$ or $n_R + n_V > n_D$
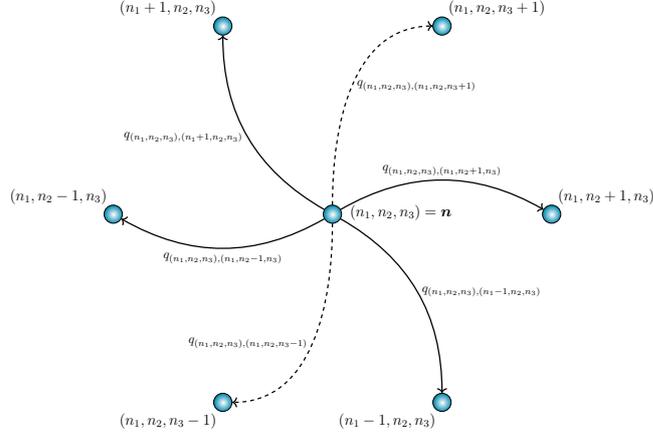
Figure 4.2: Diagram of a process $\mathfrak{X}_1$, showing possible states which the process can move to from a general state (n_1, n_2, n_3) and associated transition rates.

that result in two possible state spaces. In particular:

- if $n_R + n_V \leq n_D$, then $\mathcal{S}_1 = \{(n_1, n_2, n_2) \subset (\mathbb{N} \cup \{0\})^3 : n_1 \leq n_R, n_2 \leq n_V, n_3 \leq n_B\}$,

- if $n_R + n_V > n_D$, then $\mathcal{S}_1 = \{(n_1, n_2, n_2) \subset (\mathbb{N} \cup \{0\})^3 : n_1 + n_2 \leq n_D, n_3 \leq n_B\}$.

Which state space is necessary will depend on the values of n_R, n_V and n_D, which will be found. The overall dynamics of complex formation and dissociation, phosphorylation, and de-phosphorylation can be represented as transitions, sometimes referred to as *jumps* between states. The notation $(n_1, n_2, n_3) \rightarrow (n_1', n_2', n_3')$ implies a one-step transition from state (n_1, n_2, n_3) to state (n_1', n_2', n_3') as illustrated in Figure 4.2. From Definition 2.3.4, Figure 4.2 and reactions depicted in Figure 4.1, by the assumption of mass action kinetics, the transition rates are

given by

$$q_{(\mathbf{n},\mathbf{n'})} = \begin{cases} k_R(n_D - n_1 - n_2)(n_R - n_1), & \text{if } \mathbf{n'} = (n_1 + 1, n_2, n_3), \\ q_R n_1, & \text{if } \mathbf{n'} = (n_1 - 1, n_2, n_3), \\ k_V(n_D - n_1 - n_2)(n_V - n_2), & \text{if } \mathbf{n'} = (n_1, n_2 + 1, n_3), \\ q_V n_2, & \text{if } \mathbf{n'} = (n_1, n_2 - 1, n_3), \\ \frac{k_B(n_B - n_3)n_1}{\kappa_V + (n_V - n_2)}, & \text{if } \mathbf{n'} = (n_1, n_2, n_3 + 1), \\ q_B n_3, & \text{if } \mathbf{n'} = (n_1, n_2, n_3 - 1), \\ 0, & \text{otherwise}, \end{cases}$$

where $\mathbf{n} = (n_1, n_2, n_3)$. These rates encode both upstream and downstream viral antagonism in the type I IFN secretion pathway (Audsley & Moseley, 2013; Escudero-Pérez & Muñoz-Fontela, 2019). Firstly RNA silencing is carried out by VAP (or V). This results in a competition process for viral RNA with RIG-I, since both VAP and RIG-I can bind to viral RNA. This is the upstream mechanism of viral antagonism to inhibit type I IFN expression. The birth rate related to n_3 also describes activation of B in the presence of RIG-I:RNA complexes, n_1. This equation includes the antagonistic effect of VAP in the phosphorylation of TBK1, encoded in the denominator, $\kappa_V + (n_V - n_2)$, which implies free VAP lowers the *effective* rate of TBK1 phosphorylation, with a carrying capacity κ_V. In this way, model 1 incorporates a downstream inhibitory mechanism as well.

These rates allow the infinitesimal transition matrix from Definition 2.3.5 to be defined. Model 1 considers a three-dimensional process; as a result, the state space $\mathcal{S}_1$ can be organised in terms of levels such that,

$$\mathcal{S}_1 = \bigcup_{k=0}^{n_B} L(k) \,,$$

where $L(k) = \{(n_1, n_2, n_3) \in \mathcal{S}_1 : n_3 = k\}$ for $0 \leq n_3 \leq n_B$. Consequently, the three-dimensional construction of this process allows sub-levels to be constructed within $L(k)$. Define these sub-levels to be,

$$L(k) = \bigcup_{r=0}^{\min(n_V, n_D)} l(k; r) \,,$$

with $l(k; r) = \{(n_1, n_2, n_3) \in \mathcal{S} : n_2 = r, n_3 = k\}$, for $0 \leq r \leq \min(n_V, n_D)$ and $0 \leq k \leq n_B$. These sub-levels $l(k; r)$ are ordered as follows,

$$l(k; r) = \{(0, r, k), (1, r, k), \ldots, (\min(n_R, n_D - r), r, k)\} \,,$$

for $0 \leq r \leq \min(n_V, n_D)$, $0 \leq k \leq n_B$. This ordering and using levels and sub-levels give rise to an infinitesimal generator as in Definition 2.3.5, containing the transition rates defined earlier. The infinitesimal transition generator matrix, which follows a tri-diagonal block matrix structure, is defined as;

$$\mathbf{Q} = \begin{pmatrix} \mathbf{A}_{00} & \mathbf{A}_{01} & & & & \\ \mathbf{A}_{10} & \mathbf{A}_{11} & \mathbf{A}_{12} & & \mathbf{0} & \\ & \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} & & \\ & & \ddots & \ddots & \ddots & \\ & \mathbf{0} & & & \ddots & \mathbf{A}_{n_B-1 n_B} \\ & & & & \mathbf{A}_{n_B n_B-1} & \mathbf{A}_{n_B n_B} \end{pmatrix}.$$

The blocks within this matrix are structured as:

- for $0 \leq k \leq n_B$,

$$\mathbf{A}_{kk} = \begin{pmatrix} \mathbf{B}_{00}^{kk} & \mathbf{B}_{01}^{kk} & & & & \\ \mathbf{B}_{10}^{kk} & \mathbf{B}_{11}^{kk} & \mathbf{B}_{12}^{kk} & & \mathbf{0} & \\ & \mathbf{B}_{21}^{kk} & \mathbf{B}_{22}^{kk} & \mathbf{B}_{23}^{kk} & & \\ & & \ddots & \ddots & \ddots & \\ & \mathbf{0} & & & \ddots & \mathbf{B}_{r^*-1 r^*}^{kk} \\ & & & & \mathbf{B}_{r^* r^*-1}^{kk} & \mathbf{B}_{r^* r^*}^{kk} \end{pmatrix},$$

  where $r^* = \min(n_V, n_D)$.

- for $0 \leq k \leq n_B - 1$,

$$\mathbf{A}_{kk+1} = \begin{pmatrix} \mathbf{0} & & & & \\ \mathbf{B}_{10}^{kk+1} & \mathbf{0} & & \mathbf{0} & \\ & \mathbf{B}_{21}^{kk+1} & \mathbf{0} & & \\ & & \ddots & \ddots & \\ \mathbf{0} & & & \ddots & \\ & & & \mathbf{B}_{r^* r^*-1}^{kk+1} & \mathbf{0} \end{pmatrix},$$

104

- finally for $1 \leq k \leq n_B$,

$$\mathbf{A}_{kk-1} = \begin{pmatrix} 0 & \mathbf{B}_{01}^{kk-1} & & & & \\ & 0 & \mathbf{B}_{12}^{kk-1} & & \Large 0 & \\ & & 0 & \mathbf{B}_{23}^{kk-1} & & \\ & & & \ddots & \ddots & \\ & \Large 0 & & & \ddots & \mathbf{B}_{r^*-1r^*}^{kk-1} \\ & & & & & 0 \end{pmatrix}.$$

These matrices allow for a description of the different levels $L(k)$. Matrix $\mathbf{A}_{kk}$ describes transitions for sub-levels $l(k;r)$ for a fixed level $l(k)$, while the matrices $\mathbf{A}_{kk-1}$ and $\mathbf{A}_{kk+1}$ represent jumps from $L(k)$ to $L(k-1)$ and $L(k+1)$, respectively. The matrices $\mathbf{B}$ within the above matrices are described as follows, for $0 \leq r \leq \min(n_V, n_D)$, $0 \leq k \leq n_B$,

$$(\mathbf{B}_{rr}^{kk})_{ij} = \begin{cases} k_R(n_D - i - r)(n_R - i), & \text{for } j = i + 1, \\ -\omega(i, r, k), & \text{for } j = i, \\ q_R i, & \text{for } j = i - 1, \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

where $0 \leq i, j \leq \min(n_R, n_D - r)$. For $0 \leq r \leq \min(n_V, n_D) - 1$, $0 \leq k \leq n_B$

$$(\mathbf{B}_{rr+1}^{kk})_{ij} = \begin{cases} k_V(n_D - i - r)(n_V - r), & \text{for } j = i - 1, \\ 0, & \text{otherwise} \end{cases} \tag{4.2}$$

where $1 \leq i \leq \min(n_R, n_D - r)$ and $0 \leq j \leq \min(n_R, n_D - r) - 1$. For $1 \leq r \leq \min(n_D, n_V)$, $0 \leq k \leq n_B$

$$(\mathbf{B}_{rr-1}^{kk})_{ij} = \begin{cases} q_V r, & \text{for } j = i + 1, \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

where $0 \leq i \leq \min(n_R, n_D - r) - 1$ and $1 \leq j \leq \min(n_R, n_D - r)$. For $1 \leq r \leq \min(n_D, n_V)$, $0 \leq k \leq n_B - 1$

$$(\mathbf{B}_{rr-1}^{kk+1})_{ij} = \begin{cases} \frac{k_B(n_B - k)i}{\kappa_V + (n_V - r)}, & \text{for } j = i, \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

105

where $0 \leq i,j \leq \min(n_R, n_D - r)$. For $0 \leq r \leq \min(n_V, n_D) - 1$, $1 \leq k \leq n_B$

$$(\mathbf{B}_{rr+1}^{kk-1})_{ij} = \begin{cases} q_B k, & \text{for } j = i, \\ 0, & otherwise \end{cases}, \qquad (4.5)$$

where $0 \leq i,j \leq \min(n_R, n_D - r)$. $\omega(i,r,k)$ in Equation (4.1) are defined as

$$\omega(i,r,k) = k_R(n_D - i - r)(n_R - i) + q_R i + k_V(n_D - i - r)(n_V - r)$$
$$+ q_V r + \frac{k_B(n_B - k)i}{\kappa_V + (n_V - r)} + q_B k. \qquad (4.6)$$

From this definition of a generator matrix, it is possible to determine the stationary probability distribution of our given model by solving (Lopez-Garcia *et al.*, 2016),

$$\mathbf{Q}\boldsymbol{\pi} = \mathbf{0}, \ \boldsymbol{\pi}^T \mathbf{e} = \mathbf{1} , \qquad (4.7)$$

where $\boldsymbol{\pi}$ is made up of row vectors $\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{n_B}$ which contain ordered probabilities $\boldsymbol{\pi}_{(n_1,n_2,n_3)}$ arranged to match levels of $L(k)$. Depending on the state space, which will be determined by initial protein numbers n_D, n_V, n_R and n_B, matrix analytic methods can be utilised to determine the stationary distribution. However, if initial protein numbers lead to a large state space, analytical methods will be hindered due to computational limitations, primarily memory availability. Consequently, stochastic simulation algorithms may be used to investigate the steady-state distribution. In addition to the stationary probability distribution, studying the model's early time dynamics will be interesting. Particularly of interest is the following random variable,

$$T_{(n_1,n_2,n_3)}(N_3) = \text{``time to reach a number } N_3 \text{ of activated TBK1 } (n_3)\text{''} ,$$

for $N_3 > 0$, and some initial state $\mathbf{n} = (n_1, n_2.n_3)$. To study this stochastic descriptor, only values of n_3 such that $n_3 < N_3$ need be considered, because if $n_3 \geq N_3$ then $T_{(n_1,n_2,n_3)} = 0$. From our original continuous-time Markov process $\mathcal{X}_1$, a Markov process that considers the values of N_3 needs to be defined. Consider, $\mathcal{X}_1(N_3)$ with state space $\mathcal{S}_1(N_3)$, such that $n_3 \geq N_3$ is one absorption set $N_3' = \bigcup_{k=N_3}^{n_B} L(k)$ and a further set $\mathcal{S}_u = \bigcup_{k=0}^{N_3-1} L(k)$ that contains all the

transient states before a jump from $N_3 - 1$ to N_3 in n_3. Thus, the state space can be written as,

$$\mathcal{S}_1(N_3) = \mathcal{S}_u \cup N_3' .$$

To determine these times along with the higher moments of the distribution the use of Laplace-Stieltjes transforms is beneficial. Its definition follows;

Definition 4.1.1. If X is a continuous random variable with a probability density function $f(z)$, then the Laplace-Stieltjes transform is given by the expectation

$$\mathcal{L}f(z) = E[e^{-zX}] = \int_0^{+\infty} e^{-zX} f(z) ds, \ \mathbb{R}(z) \geq 0 ,$$

where z is a complex number (Kijima, 1997).

Given proposed model 1, the Laplace-Stieltjes transform of $T_{(n_1,n_2,n_3)(N_3)}$ is defined as,

$$\mathcal{L}_{(n_1,n_2,n_3)}(z) = E[e^{-zT_{(n_1,n_2,n_3)}(N_3)}], \ \mathbb{R}(z) \geq 0 . \tag{4.8}$$

As a result, moments can be obtained by successive differentiation of equation (4.8) with respect to z and evaluated at $z = 0$. In general, the moments can be expressed as,

$$m^l_{(n_1,n_2,n_3)} = E[T^l_{(n_1,n_2,n_3)}(N_3)] = (-1)^l \frac{d^l}{dz^l} \mathcal{L}_{(n_1,n_2,n_3)}(z)|_{z=0}, \ l \geq 1. \tag{4.9}$$

Using first-step arguments the Laplace-Stieltjes transform is as follows,

$$
\begin{aligned}
\mathcal{L}_{(n_1,n_2,n_3)} = {} & (1 - \delta_{n_1 n_R}) \frac{k_R(n_D - n_1 - n_2)(n_R - n_1)}{z + \omega(n_1, n_2, n_3)} \mathcal{L}_{(n_1+1,n_2,n_3)}(Z) \\
& + (1 - \delta_{n_1 0}) \frac{q_R n_1}{z + \omega(n_1, n_2, n_3)} \mathcal{L}_{(n_1-1,n_2,n_3)} \\
& + (1 - \delta_{n_2 n_V}) \frac{k_V(n_D - n_1 - n_2)(n_V - n_2)}{z + \omega(n_1, n_2, n_3)} \mathcal{L}_{(n_1,n_2+1,n_3)}(Z) \\
& + (1 - \delta_{n_2 0}) \frac{q_V n_2}{z + \omega(n_1, n_2, n_3)} \mathcal{L}_{(n_1,n_2-1,n_3)}(Z) \\
& + (1 - \delta_{n_3 N_3-1}) \frac{k_B(n_B - n_3)n_1}{(z + \omega(n_1, n_2, n_3))(\kappa_V + (n_V - n_2))} \delta_{n_3 N_3-1} \mathcal{L}_{(n_1,n_2,n_3+1)}(Z) \\
& + (1 - \delta_{n_3 0}) \frac{q_B n_3}{z + \omega(n_1, n_2, n_3)} \mathcal{L}_{(n_1,n_2,n_3-1)}(Z) \\
& + \delta_{n_3 N_3-1} \frac{k_B(n_B - n_3)n_1}{(z + \omega(n_1, n_2, n_3))(\kappa_V + (n_V - n_2))} ,
\end{aligned}
\tag{4.10}
$$

4. MATHEMATICAL MODEL OF INTERFERON ANTAGONISM

where $\omega(n_1, n_2, n_3)$ is as defined in Equation. (4.6) with i, r and k exchanged for n_1, n_2 and n_3, respectively. Here δ_{ij} represent a Kronecker δ. These equations describe the possible jumps between states of the state space $\mathcal{S}_1(N_3)$ for the Markov Chain $\mathcal{X}_1(N_3)$. Like the infinitesimal generator matrix, the Laplacian transforms in Equations 4.10 can be written in a block-matrix vector form. Assume a vector, $\mathbf{g}(z) = (\mathbf{g}_1(z), \mathbf{g}_2(z), \ldots, \mathbf{g}_{N_3-1}(z))$. For each $\mathbf{g}_k(z) = (\mathbf{g}_1^k(z), \mathbf{g}_2^k(z), \ldots, \mathbf{g}_{r*}^k(z))$ it is constructed to match levels $L(k)$ for $0 \leq k \leq N_3$. This can be written as

$$\mathbf{g}(z) = \mathbf{A}(z)\mathbf{g}(z) + \mathbf{c}(z), \tag{4.11}$$

where,

$$\mathbf{A}(z) = \begin{pmatrix} \mathbf{A}_{00}(z) & \mathbf{A}_{01}(z) & & & & \\ \mathbf{A}_{10}(z) & \mathbf{A}_{11}(z) & \mathbf{A}_{12}(z) & & \mathbf{0} & \\ & \mathbf{A}_{21}(z) & \mathbf{A}_{22}(z) & \mathbf{A}_{23}(z) & & \\ & & \ddots & \ddots & \ddots & \\ & \mathbf{0} & & & \ddots & \mathbf{A}_{N_3-2 N_3-1}(z) \\ & & & & \mathbf{A}_{N_3-1 N_3-2}(z) & \mathbf{A}_{N_3-1 N_3-1}(z) \end{pmatrix},$$

and,

$$\mathbf{c}(z) = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{c}_{N_3-1}(z) \end{pmatrix}.$$

The vector $\mathbf{c}$ contains probabilities for jumps from states with $n_3 = N_3 - 1$ to N_3. In the same way as the generator matrix $\mathbf{Q}$, matrix $\mathbf{A}(z)$ can be further broken down and can be defined as matrices $\mathbf{B}(z)$ which are described as follows, for $0 \leq r \leq \min(n_V, n_D)$, $0 \leq k \leq N_3 - 1$,

$$(\mathbf{B}_{rr}^{kk})_{ij} = \begin{cases} \frac{k_R(n_D-i-r)(n_R-i)}{(z+\omega(i,r,k))^{p+1}}, & \text{for } j = i+1, \\ \frac{q_R i}{(z+\omega(i,r,k))^{p+1}}, & \text{for } j = i-1, , \\ 0, & \text{otherwise} \end{cases} \tag{4.12}$$

where $0 \leq i, j \leq \min(n_R, n_D - r)$. For $0 \leq r \leq \min(n_V, n_D) - 1$, $0 \leq k \leq N_3 - 1$,

$$(\mathbf{B}_{rr+1}^{kk})_{ij} = \begin{cases} \frac{k_V(n_D-i-r)(n_V-r)}{(z+\omega(i,r,k))^{p+1}}, & \text{for } j = i-1, \\ 0, & \text{otherwise} \end{cases}, \tag{4.13}$$

where $1 \leq i \leq \min(n_R, n_D - r)$ and $0 \leq j \leq \min(n_R, n_D - r) - 1$. For $1 \leq r \leq \min(n_D, n_V)$, $0 \leq k \leq N_3 - 1$,

$$(\mathbf{B}_{rr-1}^{kk})_{ij} = \begin{cases} \frac{q_V r}{(z+\omega(i,r,k))^{p+1}}, & \text{for } j = i + 1, \\ 0, & otherwise \end{cases}, \qquad (4.14)$$

where $0 \leq i \leq \min(n_R, n_D - r) - 1$ and $1 \leq j \leq \min(n_R, n_D - r)$. For $1 \leq r \leq \min(n_D, n_V)$, $0 \leq k \leq N_3 - 2$,

$$(\mathbf{B}_{rr-1}^{kk+1})_{ij} = \begin{cases} \frac{k_B(n_B-k)i}{(\kappa_V+(n_V-r))(z+\omega(i,r,k))^{p+1}}, & \text{for } j = i, \\ 0, & otherwise \end{cases}, \qquad (4.15)$$

where $0 \leq i, j \leq \min(n_R, n_D - r)$. For $0 \leq r \leq \min(n_V, n_D) - 1$, $1 \leq k \leq N_3 - 1$,

$$(\mathbf{B}_{rr+1}^{kk-1})_{ij} = \begin{cases} \frac{q_B k}{(z+\omega(i,r,k))^{p+1}}, & \text{for } j = i, \\ 0, & otherwise \end{cases}, \qquad (4.16)$$

where $0 \leq i, j \leq \min(n_R, n_D - r)$. The vector $\mathbf{c}_{N_3-1}(z) = (\mathbf{c}_1^{N_3-1}(z), \mathbf{c}_2^{N_3-1}(z), \dots$, $\mathbf{c}_{\min(n_V, n_D)}^{N_3-1}(z))$ is broken into the levels $L(k)$ and can be further broken down into levels $l(k; r)$ such that,

$$\mathbf{c}_i^{N_3-1,r} = \frac{k_B(n_B - (N_3 - 1))i}{(z + \omega(i, r, k))(\kappa_V + (n_V - r))}, \text{ for } 0 \leq i \leq \min(n_R, n_D - r).$$

Use of definition (4.9) applied to (4.11) allows higher moments to be found.

$$\mathbf{m}^l = \sum_{p=0}^{l} (-1)^p \binom{l}{p} \frac{d^p}{dz^p} \mathbf{A}(z)|_{z=0} (-1)^{l-p} \frac{d^{l-p}}{dz^{l-p}} \mathbf{g}(z)|_{z=0} + (-1)^l \frac{d^l}{dz^l} \mathbf{c}(z)|_{z=0}$$

$$= \mathbf{A}(0)\mathbf{m}^l + \sum_{p=1}^{l} (-1)^p \binom{l}{p} \mathbf{A}^p(0)\mathbf{m}^{l-p} + (-1)^l \mathbf{c}^l \qquad (4.17)$$

The entries in derivatives of $\mathbf{A}(z)$ for $0 \leq r \leq \min(n_V, n_D)$, $0 \leq k \leq N_3 - 1$ are defined as

$$(\mathbf{B}_{rr}^{kk})_{ij}^p(z) = \begin{cases} (-1)^p p! \frac{k_R(n_D-i-r)(n_R-i)}{(z+\omega(i,r,k))^{p+1}}, & \text{for } j = i + 1, \\ (-1)^p p! \frac{q_R i}{(z+\omega(i,r,k))^{p+1}}, & \text{for } j = i - 1, , \\ 0, & otherwise \end{cases}$$

where $0 \leq i, j \leq \min(n_R, n_D - r)$. For $0 \leq r \leq \min(n_V, n_D) - 1$, $0 \leq k \leq N_3 - 1$,

$$(\mathbf{B}_{rr+1}^{kk})_{ij}^p(z) = \begin{cases} (-1)^p p! \frac{k_V(n_D - i - r)(n_V - r)}{(z + \omega(i,r,k))^{p+1}}, & \text{for } j = i - 1, \\ 0, & \text{otherwise} \end{cases},$$

where $1 \leq i \leq \min(n_R, n_D - r)$ and $0 \leq j \leq \min(n_R, n_D - r) - 1$. For $1 \leq r \leq \min(n_D, n_V)$, $0 \leq k \leq N_3 - 1$,

$$(\mathbf{B}_{rr-1}^{kk})_{ij}^p(z) = \begin{cases} (-1)^p p! \frac{q_V r}{(z + \omega(i,r,k))^{p+1}}, & \text{for } j = i + 1, \\ 0, & \text{otherwise} \end{cases},$$

where $0 \leq i \leq \min(n_R, n_D - r) - 1$ and $1 \leq j \leq \min(n_R, n_D - r)$. For $1 \leq r \leq \min(n_D, n_V)$, $0 \leq k \leq N_3 - 2$,

$$(\mathbf{B}_{rr-1}^{kk+1})_{ij}^p(z) = \begin{cases} (-1)^p p! \frac{k_B(n_B - k)i}{(z + \omega(i,r,k))^{p+1}(\kappa_V + (n_V - r))}, & \text{for } j = i, \\ 0, & \text{otherwise} \end{cases},$$

where $0 \leq i, j \leq \min(n_R, n_D - r)$. For $0 \leq r \leq \min(n_V, n_D) - 1$, $1 \leq k \leq N_3 - 1$,

$$(\mathbf{B}_{rr+1}^{kk-1})_{ij}^p(z) = \begin{cases} (-1)^p p! \frac{q_B k}{(z + \omega(i,r,k))^{p+1}}, & \text{for } j = i, \\ 0, & \text{otherwise} \end{cases},$$

where $0 \leq i, j \leq \min(n_R, n_D - r)$. The derivative of vector $\mathbf{c}(z)$ will have the entries

$$(\mathbf{c}^{N_3-1,r})_i^p = (-1)^p p! \frac{k_B(n_B - (N_3 - 1))i}{(z + \omega(i,r,k))^{p+1}(\kappa_V + (n_V - r))},$$

for $0 \leq i \leq \min(n_R, n_D - r)$ and $0 \leq r \leq \min(n_V, n_D)$. The proposed model only includes two mechanisms of inhibition and one particular signalling pathway; therefore, investigating additional mechanisms is of interest.

A second model of type I IFN inhibition by VAP: PACT protein

The model shown in Figure. 4.1 characterises two key aspects of the role of VAP: upstream antagonism with RIG-I when binding to viral RNA and downstream antagonism in the activation of TBK1. Yet, viruses explore additional mechanisms to inhibit type I IFN secretion. The protein activator of the interferon-induced protein kinase (PACT) has been identified as a secondary activator of RIG-I during viral infections (Ho *et al.*, 2016; Luthra *et al.*, 2013). This molecule
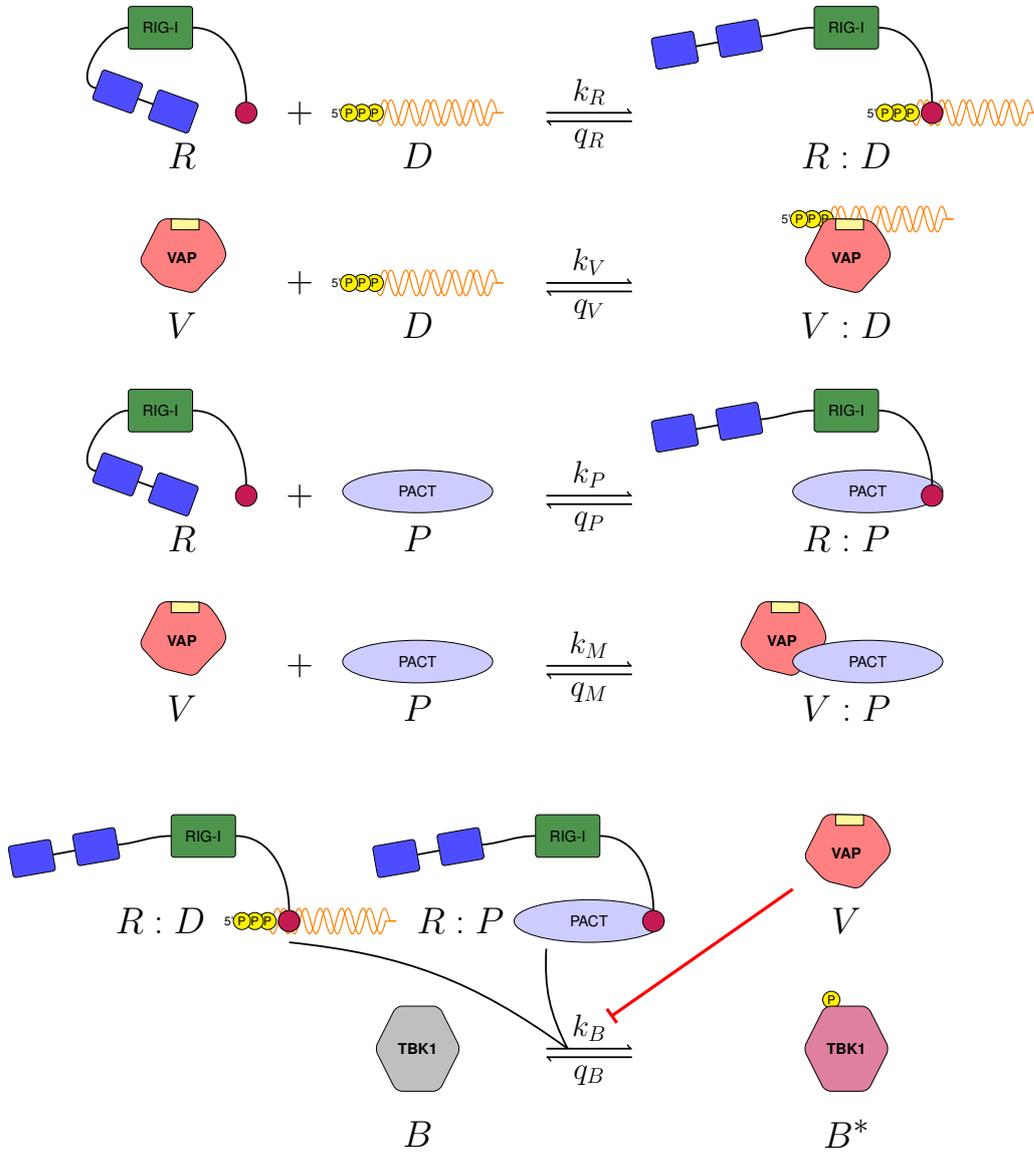
Figure 4.3: A second model of type I IFN inhibition by viral protein (model 2). Model 2 includes the role of protein activator of the interferon-induced protein kinase (PACT) molecules, since PACT has been identified as a secondary activator of RIG-I during viral infections (Ho *et al.*, 2016; Luthra *et al.*, 2013). In this model, there are ten reactions and five molecular complexes.

provides an additional activation route for RIG-I, and thus, a boost to type I IFN induction (see Figure. 4.3). However, just like RNA silencing, many viruses have mechanisms to inhibit the interaction of PACT with RIG-I. Influenza virus protein NS1, EBOV VP35 and MERS-CoV protein 4a (p4a) have all been identified to cause antagonism of the interaction between PACT and RIG-I (Luthra *et al.*, 2013; Siu *et al.*, 2014; Tawaratsumida *et al.*, 2014). Therefore, to include this second viral strategy of innate immunity inhibition, I propose a second model which includes PACT, and all the other molecular species and reactions of model 1.

R, D, V and B retain their previous definitions (see Section 4.1.1), and PACT is denoted by P. Based on this experimental evidence (Ho *et al.*, 2016; Luthra *et al.*, 2013), the following set of reactions are proposed, illustrated in Figure. 4.3. The first, second and final reactions remain unchanged from model 1 (see Figure. 4.3). The third reaction includes the binding of RIG-I and PACT to form a RIG-I:PACT complex ($R : P$) with rate k_P, and dissociation rate q_P. The final new reaction (fourth reaction) includes the binding of VAP to PACT to form a VAP:PACT complex ($V : P$) with binding rate k_M, and unbinding rate q_M. From here, this model will be referred to as *model 2*.

Denote by n_R, n_D, n_P, n_V, and n_B, the per cell total number of RIG-I, viral RNA molecules, PACT, VAP and TBK1, respectively. As previously described neglects protein degradation and synthesis, so that the total number of molecules for each protein species is conserved. Denote the total number of RIG-I:RNA ($R : D$) complexes at time $t \geq 0$ by $n_{RD}(t)$, VAP:RNA ($V : D$) by $n_{VD}(t)$, RIG-I:PACT ($R : P$) by $n_{RP}(t)$, VAP:PACT ($V : P$) by $n_{VP}(t)$ and activated TBK1 (B^*) by $n_{B^*}(t)$. As with *model 1*, conservation of molecules is assumed such that at a given time $R(t) = n_R - n_{RD}(t) - n_{RP}(t)$, $V(t) = n_V - n_{VD}(t) - n_{VP}(t)$, $P(t) = n_P - n_{RP}(t) - n_{VP}(t)$, $D(t) = n_D - n_{RD}(t) - n_{VD}(t)$ and $B(t) = n_B - n_{B^*}(t)$ for RIG-I, VAP, PACT, viral RNA and TBK1 respectively.

Let $(n_{RD}, n_{VD}, n_{RP}, n_{VP}, n_{B^*}) = (n_1, n_2, n_3, n_4, n_5) = \mathbf{n}$ and consider a CTMC $\mathcal{X}_2 = \{\mathbf{n}(t) : t \geq 0\}$ where state vector $\mathbf{n} \in \mathcal{S}_2 \subset (\mathbb{N} \cup \{0\})^5$ is a collection of random variables representing the number of each molecule at time t. The state space $\mathcal{S}_2$ can be identified from the conditions imposed by conserved proteins

numbers $(n_R, n_V, n_P, n_D, n_B)$ and Figure 4.3. The non-null infinitesimal transition rates can be defined as

$$
q_{(\mathbf{n},\mathbf{n'})} = \begin{cases}
k_R(n_D - n_1 - n_2)(n_R - n_1 - n_3), & \text{if } \mathbf{n'} = (n_1 + 1, n_2, n_3, n_4, n_5), \\
q_R n_1, & \text{if } \mathbf{n'} = (n_1 - 1, n_2, n_3, n_4, n_5), \\
k_V(n_D - n_1 - n_2)(n_V - n_2 - n_4), & \text{if } \mathbf{n'} = (n_1, n_2 + 1, n_3, n_4, n_5), \\
q_V n_2, & \text{if } \mathbf{n'} = (n_1, n_2 - 1, n_3, n_4, n_5), \\
k_P(n_P - n_3 - n_4)(n_R - n_1 - n_3), & \text{if } \mathbf{n'} = (n_1, n_2, n_3 + 1, n_4, n_5), \\
q_P n_3 & \text{if } \mathbf{n'} = (n_1, n_2, n_3 - 1, n_4, n_5), \\
k_M(n_P - n_3 - n_4)(n_V - n_2 - n_4) & \text{if } \mathbf{n'} = (n_1, n_2, n_3, n_4 + 1, n_5), \\
q_M n_4, & \text{if } \mathbf{n'} = (n_1, n_2, n_3, n_4 - 1, n_5), \\
\frac{k_B(n_B - n_5)(n_1 + n_3)}{\kappa_V + (n_V - n_2 - n_4)}, & \text{if } \mathbf{n'} = (n_1, n_2, n_3, n_4, n_5 + 1), \\
q_B n_5, & \text{if } \mathbf{n'} = (n_1, n_2, n_3, n_4, n_5 - 1), \\
0, & \text{otherwise.}
\end{cases}
$$

Unlike model 1, the five populations under consideration lead to matrices with extremely large dimensions. Consequently, studying stochastic descriptors using matrix analytic methods is computationally limited due to the large amounts of memory that would be required. Therefore, Gillespie simulations will be implemented instead, using the Algorithm 2.3.1 to investigate this model.

A third model of type I IFN inhibition by VAP: PKR signalling pathway

So far, two models have been introduced which examine the effects of VAP on RIG-I induced type I interferon induction. However, as discussed in the introduction of this section, other PRRs exist. One such alternative pathway is protein kinase R (PKR), which also binds to viral RNA, and the resulting bound complex, *A:D* in Figure 4.4, induces the type I IFN secretion pathway (McAllister *et al.*, 2010, 2012). Viruses can also target this pathway. Influenza, herpes simplex 1 and Ebola viruses have been observed to inhibit the PKR pathway (George *et al.*, 2009; Schümann *et al.*, 2009). From current experimental evidence, I introduce a third and final mathematical model. In this model, the complex, *A:D* (see Figure 4.4), plays the role performed by phosphorylated TBK1 in models 1 and 2, as the downstream element in the pathway to induce type I IFN secretion. This model does not consider the RIG-I pathway but describes the PKR one. Yet, I

consider the role of RIG-I in sequestering viral RNA from both VAP and PKR, and the role of RIG-I in sequestering PACT from VAP. This allows for competition between molecular species which may impact the overall dynamics,

Consider molecules R, D, V and P, as in models 1 and 2. Let's now introduce A to represent PKR. From model 2, I retain the first four reactions shown in Figure 4.3 to keep competition for viral RNA between VAP, RIG-I and PKR. Previous reactions that involve free VAP molecules are also kept. The final set of reactions considered, with rates k_A and q_A, respectively, are presented in Figure 4.4. A fifth reaction is PKR binding to viral RNA, with rate k_A, resulting in the formation of a phosphorylated PKR:RNA (A:D) complex. This complex can become un-phosphorylated and disassociate with rate q_A. In the presence of VAP, PKR can be actively de-phosphorylated and disassociated with rate q_{AV}. Thus, in this model, a new viral strategy exists to inhibit innate recognition via type I IFN. Variables $n_{RD}(t), n_{VD}(t), n_{RP}(t)$ and $n_{VP}(t)$ represent the same complexes as in model 2. The new variable, which describes complex A:D, $n_{AD}(t)$, now represents the number of phosphorylated PKR:RNA complexes. This model is referred to as *Model 3*. Similar to model 1 and model 2, conservation of proteins is assumed such that $D(t) = n_D - n_{RD} - n_{VD} - n_{AD}$, $R(t) = n_R - n_{RD} - n_{RP}$, $V(t) = n_V - n_{VD} - n_{VP}$, $P(t) = n_P - n_{RP} - n_{VP}$ and $A(t) = n_A - n_{AD}$ for dsRNA, RIG-I, VAP, PACT and PKR respectively.

Let $(n_{RD}, n_{VD}, n_{RP}, n_{VP}, n_{AD}) = (n_1, n_2, n_3, n_4, n_5) = \mathbf{n}$ and consider a CTMP $\mathcal{X}_3 = \{\mathbf{n}(t) : t \geq 0\}$ where state vector $\mathbf{n} \in \mathcal{S}_3 \subset (\mathbb{N} \cup \{0\})^5$ is a collection of random variables representing the number of each molecule at time t. The state space $\mathcal{S}_3$ can be identified from the conditions imposed by conserved proteins numbers $(n_R, n_V, n_P, n_D, n_A)$ and Figure 4.4. The-non null infinitesimal transition

rates can be defined as

$$q_{(\mathbf{n},\mathbf{n}')} = \begin{cases} k_R(n_R - n_1 - n_3)(n_D - n_1 - n_2 - n_5), & \text{if } \mathbf{n}' = (n_1 + 1, n_2, n_3, n_4, n_5), \\ q_R n_1, & \text{if } \mathbf{n}' = (n_1 - 1, n_2, n_3, n_4, n_5), \\ k_V(n_V - n_2 - n_4)(n_D - n_1 - n_2 - n_5), & \text{if } \mathbf{n}' = (n_1, n_2 + 1, n_3, n_4, n_5), \\ q_V n_2, & \text{if } \mathbf{n}' = (n_1, n_2 - 1, n_3, n_4, n_5), \\ k_P(n_R - n_1 - n_3)(n_P - n_3 - n_4), & \text{if } \mathbf{n}' = (n_1, n_2, n_3 + 1, n_4, n_5), \\ q_P n_3, & \text{if } \mathbf{n}' = (n_1, n_2, n_3 - 1, n_4, n_5), \\ k_M(n_V - n_2 - n_4)(n_P - n_3 - n_4), & \text{if } \mathbf{n}' = (n_1, n_2, n_3, n_4 + 1, n_5), \\ q_M n_4, & \text{if } \mathbf{n}' = (n_1, n_2, n_3, n_4 - 1, n_5), \\ k_A(n_A - n_5)(n_D - n_1 - n_2 - n_5), & \text{if } \mathbf{n}' = (n_1, n_2, n_3, n_4, n_5 + 1), \\ [q_A + q_{AV}(n_V - n_2 - n_4)]n_5, & \text{if } \mathbf{n}' = (n_1, n_2, n_3, n_4, n_5 - 1), \\ 0, & \text{otherwise.} \end{cases}$$

It is worth remembering that all reactions in this model are described by mass action kinetics, except the one proportional to q_{AV}. In this case, to model the de-phosphorylation and disassociation enhancement caused by the VAP, I have added a term proportional to the number of free VAP molecules, $n_V - n_{VD} - n_{VP}$. Finally, note that in this model, TBK1 is assumed to be either non-functional as part of the signalling pathway, or insufficiently stimulated, to contribute to the production of type I IFN.

4.2   Linear noise approximation

The dynamics of each proposed model can be described by the master equation (ME) otherwise known as the forward Kolmogorov equation in Definition 2.3.6 (Allen, 2010),

$$\frac{dp_{\mathbf{n}}(t)}{dt} = \sum_{\mathbf{n}' \in \mathcal{S}, \mathbf{n}' \neq \mathbf{n}} q_{(\mathbf{n}',\mathbf{n})} p_{\mathbf{n}'}(t) - \sum_{\mathbf{n}' \in \mathcal{S}, \mathbf{n}' \neq \mathbf{n}} q_{(\mathbf{n},\mathbf{n}')} p_{\mathbf{n}}(t), \qquad \forall \mathbf{n} \in \mathcal{S}$$

where $\mathcal{S}$ represents the state space for a given model under consideration, with initial condition $p_{\mathbf{0}}(0) = 1$. This equation is known to be difficult to solve analytically and therefore is typically investigated using alternative methods. As previously mentioned, Gillespie simulations can be used to study and simulate stochastic models (Section 2.3.1). However, the linear noise approximation can
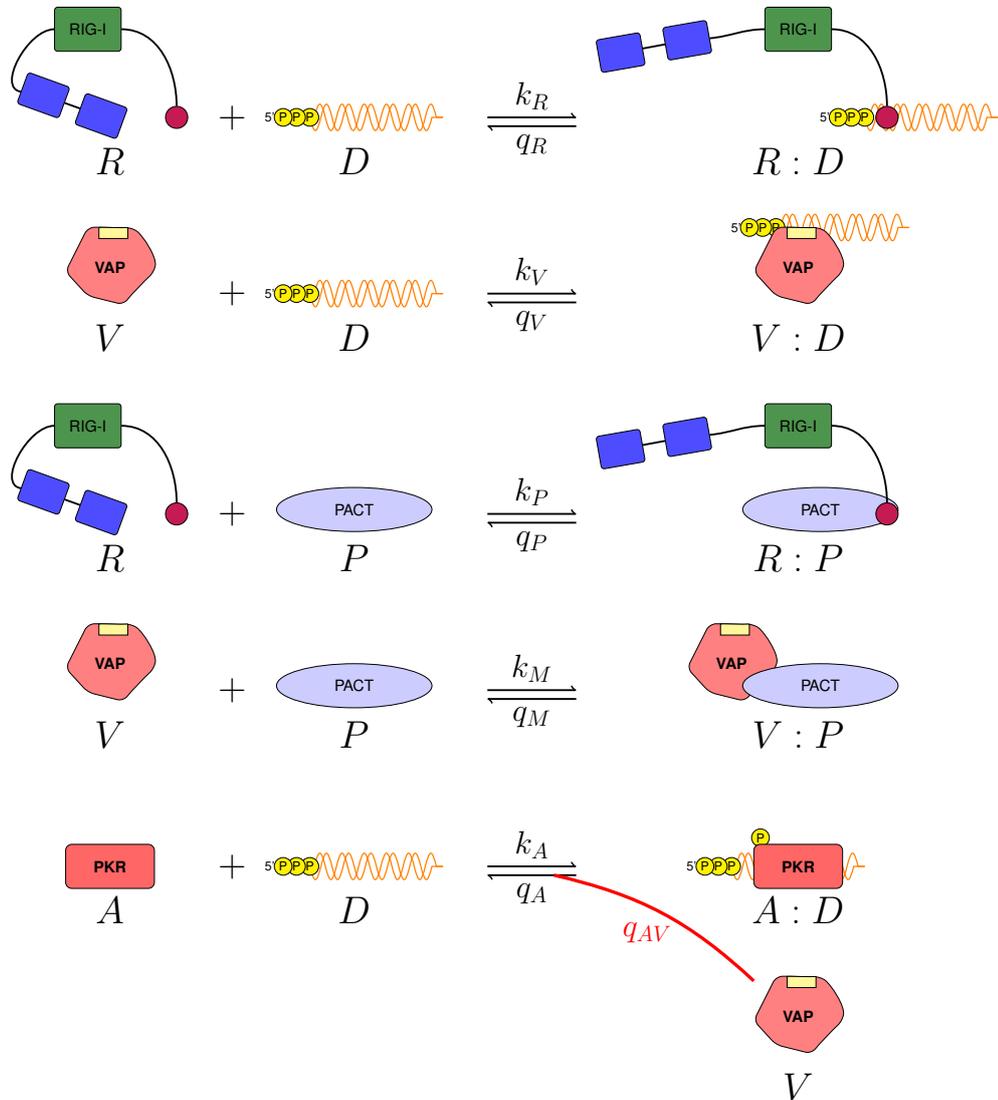
Figure 4.4: A third model of type I IFN inhibition by viral protein (model 3). It includes the contribution of the PKR pathway. In this model, there are ten reactions and five molecular complexes. In the presence of free VAP, phosphorylated PKR is actively de-phosphorylated, as indicated by the red reverse reaction arrow with rate q_{AV}.

also allow investigation of these models (sometimes referred to as the system size (Ω-) expansion); see Van Kampen (1992). The linear noise approximation means that the master equation is expanded around a local steady-state in terms of $1/\sqrt{\Omega}$ where Ω is a specified volume, such as within a cell. This expansion allows conversion from a macroscopic analysis to a mesoscopic analysis to study fluctuations. First-order terms of $1/\sqrt{\Omega}$ provide a deterministic approximation that describes "mean-field" rates of change for each protein characterised by a set of ordinary differential equations. Meanwhile, second-order terms explain fluctuations for each protein around the mean-field steady-state expressed by the linear Fokker-Planck equation (Peralta & Toral, 2018). I aim to apply this method to each model proposed in Section 4.1.

First, it is useful to define the step operator $E_i^{\pm k}$ which when applied to a function f of i, changes the value of i by $\pm k$ such that in general,

$$E_i^k[f(i)] = f(i+k), \qquad E_i^{-k}[f(i)] = f(i-k).$$

Therefore, from the rates defined in Section 4.1.1 the ME in terms of step operator $E_i^{\pm k}$ for Model 1 are,

$$
\begin{aligned}
\frac{dp_{\mathbf{n}}}{dt} &= (E_{n_1}^{-1} - 1)[k_R(n_D - n_1 - n_2)(n_R - n_1)p_{\mathbf{n}}] + (E_{n_1}^1 - 1)[q_R n_1 p_{\mathbf{n}}] \\
&\quad + (E_{n_2}^{-1} - 1)[k_V(n_D - n_1 - n_2)(n_V - n_2)p_{\mathbf{n}}] + (E_{n_2}^1 - 1)[q_V n_2 p_{\mathbf{n}}] \\
&\quad + (E_{n_3}^{-1} - 1)\left[\frac{k_B(n_B - n_3)n_1}{\kappa_V + (n_V - n_2)}p_{\mathbf{n}}\right] + (E_{n_3}^1 - 1)[q_B n_3 p_{\mathbf{n}}],
\end{aligned}
\tag{4.18}
$$

where t is omitted for brevity. *Model 1* will be investigated as a worked example. Next, the macroscopic variables must be re-scaled such that any discrete variable n_i is written as a sum of the "mean" number of type $i \in 1, 2, 3$ particles and linear noise (fluctuations) ξ_i of order $\Omega^{1/2}$ such that,

$$n_i = \Omega m_i + \Omega^{\frac{1}{2}}\xi_i. \tag{4.19}$$

As a result, the probability density and step operator must undergo a change variable. The probability density $p_{\mathbf{n}}(t)$ changes variables to $\Pi(\boldsymbol{\xi}, t)$ which must be differentiated with respect to time t to determine the time evolution,

$$\frac{dp_{\mathbf{n}}}{dt} = \partial_t \Pi(\boldsymbol{\xi}, t) + \partial_{\boldsymbol{\xi}}\Pi(\boldsymbol{\xi}, t)\frac{d\boldsymbol{\xi}}{dt}. \tag{4.20}$$

Then, since the values of n_i are fixed for the derivative of $\frac{d}{dt}p_{\mathbf{n}}$ then from Equation (4.19) it can be shown that,

$$0 = \frac{dn_i}{dt} = \Omega\frac{dm_i}{dt} + \Omega^{\frac{1}{2}}\frac{d\xi_i}{dt},$$

and hence

$$\frac{dm_i}{dt} = -\Omega^{-\frac{1}{2}}\frac{d\xi_i}{dt}. \tag{4.21}$$

Substituting expression (4.21) into (4.20) gives the time evolution of the probability density in terms of variables m_i, ξ_i and the volume Ω as

$$\frac{d}{dt}p_{\mathbf{n}} = \partial t\Pi(\boldsymbol{\xi}, t) - \Omega^{\frac{1}{2}}\sum_i \frac{dm_i}{dt}\partial_{\xi_i}\Pi(\boldsymbol{\xi}, t). \tag{4.22}$$

The right hand side of the ME also need to be rewritten in terms of the new variables. Therefore it is useful to approximate the step operator by using Taylor's expansion, where

$$E_{n_i}^{\pm 1} = 1 \pm \partial_{n_i} + \frac{1}{2}\partial_{n_i}^2 + \ldots \qquad .$$

Applying the change of variable takes the step operator from $n_i + k$ to $\xi_i + \Omega^{-\frac{1}{2}}k$, hence the step operator becomes

$$E_{n_i}^{\pm 1} = 1 \pm \Omega^{-\frac{1}{2}}\partial_\xi + \frac{1}{2}\Omega^{-1}\partial_{\xi_i}^2 + \ldots \qquad . \tag{4.23}$$

From here Equations (4.19), (4.20) and (4.23) can be substituted into Equa-

tion (4.18) to give,

$$
\begin{aligned}
\partial_t \Pi - \Omega^{\frac{1}{2}} \sum_{i=1}^{3} \frac{dm_i}{dt} \partial_{\xi_i} \Pi \;=\; & -\tilde{k}_R \Omega^{\frac{1}{2}} (m_R - m_1)(m_D - m_1 - m_2) \partial_{\xi_1} \Pi \\
& - \tilde{k}_V \Omega^{\frac{1}{2}} (m_V - m_2)(m_D - m_1 - m_2) \partial_{\xi_2} \Pi \\
& - k_B \Omega^{\frac{1}{2}} \frac{m_1 (m_B - m_3)}{\tilde{\kappa}_V + m_V - m_2} \partial_{\xi_3} \Pi \\
& + q_R \Omega^{\frac{1}{2}} m_1 \partial_{\xi_1} \Pi + q_V \Omega^{\frac{1}{2}} m_2 \partial_{\xi_2} \Pi + q_B \Omega^{\frac{1}{2}} m_3 \partial_{\xi_3} \Pi \\
& + \tilde{k}_V (m_V - m_2)(m_D - m_1 - m_2)(\tfrac{1}{2} \partial_{\xi_2}^2) \Pi \\
& + \tilde{k}_V (m_V - m_2) \partial_{\xi_2} [(\xi_1 + \xi_2) \Pi] \\
& + \tilde{k}_V (m_D - m_1 - m_2) \partial_{\xi_2} (\xi_2 \Pi) \\
& + \tilde{k}_R (m_R - m_1)(m_D - m_1 - m_2)(\tfrac{1}{2} \partial_{\xi_1}^2) \Pi \\
& + \tilde{k}_R (m_R - m_1) \partial_{\xi_1} [(\xi_1 + \xi_2) \Pi] \qquad\qquad (4.24) \\
& + \tilde{k}_R (m_D - m_1 - m_2) \partial_{\xi_1} (\xi_1 \Pi) \\
& + k_B \frac{m_1 (m_B - m_3)}{\tilde{\kappa}_V + m_V - m_2} (\tfrac{1}{2} \partial_{\xi_3}^2) \Pi \\
& - k_B \left[\frac{m_1 (m_B - m_3)}{(\tilde{\kappa}_V + m_V - m_2)^2} \partial_{\xi_3} (\xi_2 \Pi) \right] \\
& - k_B \left[\frac{(m_B - m_3)}{(\tilde{\kappa}_V + m_V - m_2)} \partial_{\xi_3} (\xi_1 \Pi) \right] \\
& + k_B \left[\frac{m_1}{(\tilde{\kappa}_V + m_V - m_2)} \partial_{\xi_3} (\xi_3 \Pi) \right] \\
& + q_R \left[\partial_{\xi_1} (\xi_1 \Pi) + \tfrac{1}{2} m_1 \partial_{\xi_1}^2 \Pi \right] \\
& + q_V \left[\partial_{\xi_2} (\xi_2 \Pi) + \tfrac{1}{2} m_2 \partial_{\xi_2}^2 \Pi \right] \\
& + q_B \left[\partial_{\xi_3} (\xi_3 \Pi) + \tfrac{1}{2} m_3 \partial_{\xi_3}^2 \Pi \right] ,
\end{aligned}
$$

where we simplify $\Pi(\boldsymbol{\xi}, t)$ to Π in the notation and,

$$
\begin{aligned}
k_R &= \Omega^{-1} \tilde{k}_R , & n_R &= \Omega m_R , \\
k_V &= \Omega^{-1} \tilde{k}_V , & n_V &= \Omega m_V , \\
n_D &= \Omega m_D , & \kappa_V &= \Omega \tilde{\kappa} .
\end{aligned}
$$

4. MATHEMATICAL MODEL OF INTERFERON ANTAGONISM

Collecting terms of order $\Omega^{\frac{1}{2}}$ in Equation (4.24) and equating coefficients gives the following ordinary differential equations,

$$
\begin{aligned}
\frac{dm_1}{dt} &= \tilde{k}_R(m_R - m_1)(m_D - m_1 - m_2) - q_R m_1 , \\
\frac{dm_2}{dt} &= \tilde{k}_V(m_V - m_2)(m_D - m_1 - m_2) - q_V m_2 , \\
\frac{dm_3}{dt} &= k_B \frac{m_1(m_B - m_3)}{\tilde{\kappa}_V + m_V - m_2} - q_B m_3 ,
\end{aligned}
\qquad (4.25)
$$

where m_1, m_2 and m_3 correspond to RIG-1:viral RNA complexes, VAP:viral RNA complexes and activated TBK1 respectively. The ODEs in Equation (4.25) could be obtained by considering Figure 4.1 as a purely deterministic system. Collecting and equating terms of order Ω^0 from Equation (4.24) provides the linear Fokker-Planck equation of the form,

$$
\partial_t \Pi = -\sum_{i,j=1}^{N} A_{ij}\, \partial_{\xi_i}(\xi_j \Pi) + \frac{1}{2}\sum_{i,j=1}^{N} B_{ij}\, \partial^2_{\xi_i \xi_j} \Pi ,
\qquad (4.26)
$$

where there are N random variables. The matrix $\boldsymbol{B}$ is symmetric but the matrix $\boldsymbol{A}$, in general, is not, with their entries for model 1 as follows,

$$
\begin{aligned}
B_{11} &= q_R m_1^* + \tilde{k}_R(m_R - m_1^*)(m_D - m_1^* - m_2^*) , \\
B_{22} &= q_V m_2^* + \tilde{k}_V(m_V - m_2^*)(m_D - m_1^* - m_2^*) , \\
B_{33} &= q_B m_3^* + k_B \frac{m_1^*(m_B - m_3^*)}{\tilde{\kappa}_V + m_V - m_2^*} , \\
-A_{11} &= q_R + \tilde{k}_R(m_R - m_1^*) + \tilde{k}_R(m_D - m_1^* - m_2^*) , \\
-A_{12} &= \tilde{k}_R(m_R - m_1^*) , \\
-A_{21} &= \tilde{k}_V(m_V - m_2^*) , \\
-A_{22} &= q_V + \tilde{k}_V(m_V - m_2^*) + \tilde{k}_V(m_D - m_1^* - m_2^*) , \\
-A_{31} &= -k_B \left[\frac{(m_B - m_3^*)}{(\tilde{\kappa}_V + m_V - m_2^*)} \right] , \\
-A_{32} &= -k_B \left[\frac{m_1^*(m_B - m_3^*)}{(\tilde{\kappa}_V + m_V - m_2^*)^2} \right] , \\
-A_{33} &= q_B + k_B \left[\frac{m_1^*}{(\tilde{\kappa}_V + m_V - m_2^*)} \right] ,
\end{aligned}
$$

where (m_1^*, m_2^*, m_3^*) represents the steady-state solution of Equations (4.25). ODEs that describe the first and second-order moments for the fluctuations can be obtained by multiplying Equation (4.26) by ξ_i for $i \in \{1, 2, 3\}$ and integrating. These moments are defined as

$$\frac{d}{dt}\langle\xi_j\rangle = \sum_{k=1}^{3} A_{jk}\langle\xi_k\rangle , \tag{4.27}$$

$$\frac{d}{dt}\langle\xi_i\xi_j\rangle = \sum_{k=1}^{3} A_{ik}\langle\xi_k\xi_j\rangle + \sum_{k=1}^{3} A_{jk}\langle\xi_i\xi_k\rangle + B_{ij} , \tag{4.28}$$

where $\langle\cdot\rangle = \mathbb{E}[\cdot]$ for $k, i, j \in \{1, 2, 3\}$. From the definitions for first and second moments in Equations (4.27) and (4.28) respectively, steady-state correlations can be defined as,

$$C_{ij} = \langle\xi_i\xi_j\rangle - \langle\xi_i\rangle\langle\xi_j\rangle. \tag{4.29}$$

This method can also be applied to model 2, to provide the deterministic mean-field approximation as follows,

$$\begin{aligned}
\frac{dm_1}{dt} &= \tilde{k}_R(m_R - m_1 - m_3)(m_D - m_1 - m_2) - q_R m_1 , \\
\frac{dm_2}{dt} &= \tilde{k}_V(m_V - m_2 - m_4)(m_D - m_1 - m_2) - q_V m_2 , \\
\frac{dm_3}{dt} &= \tilde{k}_P(m_R - m_1 - m_3)(m_P - m_3 - m_4) - q_P m_3 , \\
\frac{dm_4}{dt} &= \tilde{k}_M(m_V - m_2 - m_4)(m_P - m_3 - m_4) - q_M m_4 , \\
\frac{dm_5}{dt} &= k_B\frac{(m_1 + m_3)(m_B - m_5)}{\tilde{\kappa}_V + m_V - m_2 - m_4} - q_B m_5,
\end{aligned} \tag{4.30}$$

where

$$\begin{aligned}
k_R &= \Omega^{-1}\tilde{k}_R , & n_R &= \Omega m_R , & n_P &= \Omega m_P , \\
k_V &= \Omega^{-1}\tilde{k}_V , & n_V &= \Omega m_V , & k_P &= \Omega^{-1}\tilde{k}_P , \\
n_D &= \Omega m_D , & \kappa_V &= \Omega\tilde{\kappa}_V , & k_M &= \Omega^{-1}\tilde{k}_M .
\end{aligned}$$

The entries in matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ for the general linear Fokker-Planck equation,

given the steady-state solution $(m_1^*, m_2^*, m_3^*, m_4^*, m_5^*)$ are,

$$
\begin{aligned}
B_{11} &= q_R m_1^* + \tilde{k}_R(m_R - m_1^* - m_3^*)(m_D - m_1^* - m_2^*) \,, \\
B_{22} &= q_V m_2^* + \tilde{k}_V(m_V - m_2^* - m_4^*)(m_D - m_1^* - m_2^*) \,, \\
B_{33} &= q_P m_3^* + \tilde{k}_P(m_R - m_1^* - m_3^*)(m_P - m_3^* - m_4^*) \,, \\
B_{44} &= q_M m_4^* + \tilde{k}_M(m_V - m_2^* - m_4^*)(m_P - m_3^* - m_4^*) \,, \\
B_{55} &= q_B m_5^* + k_B \frac{(m_1^* + m_3^*)(m_B - m_5^*)}{\tilde{\kappa}_V + m_V - m_2^* - m_4^*} \,, \\
-A_{11} &= q_R + \tilde{k}_R(m_R - m_1^* - m_3^*) + \tilde{k}_R(m_D - m_1^* - m_2^*) \,, \\
-A_{12} &= \tilde{k}_R(m_R - m_1^* - m_3^*) \,, \\
-A_{13} &= \tilde{k}_R(m_D - m_1^* - m_2^*) \,, \\
-A_{21} &= \tilde{k}_V(m_V - m_2^* - m_4^*) \,, \\
-A_{22} &= q_V + \tilde{k}_V(m_V - m_2^* - m_4^*) + \tilde{k}_V(m_D - m_1^* - m_2^*) \,, \\
-A_{24} &= \tilde{k}_V(m_D - m_1^* - m_2^*) \,, \\
-A_{31} &= \tilde{k}_P(m_P - m_3^* - m_4^*) \,, \\
-A_{33} &= q_P + \tilde{k}_P(m_P - m_3^* - m_4^*) + \tilde{k}_P(m_R - m_1^* - m_3^*) \,, \\
-A_{34} &= \tilde{k}_P(m_R - m_1^* - m_3^*) \,, \\
-A_{42} &= \tilde{k}_M(m_P - m_3^* - m_4^*) \,, \\
-A_{43} &= \tilde{k}_M(m_V - m_2^* - m_4^*) \,, \\
-A_{44} &= q_M + \tilde{k}_M(m_V - m_2^* - m_4^*) + \tilde{k}_M(m_P - m_3^* - m_4^*) \,, \\
-A_{51} &= \frac{(m_B - m_5^*)}{\tilde{\kappa}_V + m_V - m_2^* - m_4^*} \,, \\
-A_{52} &= \frac{-(m_B - m_5^*)(m_1^* + m_3^*)}{(\tilde{\kappa}_V + m_V - m_2^* - m_4^*)^2} \,, \\
-A_{53} &= \frac{(m_B - m_5^*)}{\tilde{\kappa}_V + m_V - m_2^* - m_4^*} \,, \\
-A_{54} &= \frac{-(m_B - m_5^*)(m_1^* + m_3^*)}{(\tilde{\kappa}_V + m_V - m_2^* - m_4^*)^2} \,, \\
-A_{55} &= q_B + \frac{(m_1^* + m_3^*)}{\tilde{\kappa}_V + m_V - m_2^* - m_4^*} \,.
\end{aligned}
$$

Similarly for model 3 the deterministic mean-field approximation is given by,

$$
\begin{aligned}
\frac{dm_1}{dt} &= \tilde{k}_R(m_R - m_1 - m_3)(m_D - m_1 - m_2 - m_5) - q_R m_1 \ , \\
\frac{dm_2}{dt} &= \tilde{k}_V(m_V - m_2 - m_4)(m_D - m_1 - m_2 - m_5) - q_V m_2 \ , \\
\frac{dm_3}{dt} &= \tilde{k}_P(m_R - m_1 - m_3)(m_P - m_3 - m_4) - q_P m_3 \ , \\
\frac{dm_4}{dt} &= \tilde{k}_M(m_V - m_2 - m_4)(m_P - m_3 - m_4) - q_M m_4 \ , \\
\frac{dm_5}{dt} &= \tilde{k}_A(m_A - m_5)(m_D - m_1 - m_2 - m_3) - (q_A - \tilde{q}_{AV}(m_V - m_2 - m_4)m_5)
\end{aligned}
\tag{4.31}
$$

where

$$
\begin{array}{lll}
k_R = \Omega^{-1}\tilde{k}_R \ , & n_R = \Omega m_R \ , & n_P = \Omega m_P \ , \\[2mm]
k_V = \Omega^{-1}\tilde{k}_V \ , & n_V = \Omega m_V \ , & k_P = \Omega^{-1}\tilde{k}_P \ , \\[2mm]
n_D = \Omega m_D \ , & \kappa_V = \Omega\tilde{\kappa}_V \ , & k_M = \Omega^{-1}\tilde{k}_M \ . \\[2mm]
n_A = \Omega m_A \ , & q_{AV} = \Omega\tilde{q}_{AV} \ , & k_A = \Omega^{-1}\tilde{k}_A \ .
\end{array}
$$

The entries in matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ in the general linear Fokker-Planck equation,

given the steady-state solution $(m_1^*, m_2^*, m_3^*, m_4^*, m_5^*)$ are,

$$
\begin{aligned}
B_{11} &= q_R m_1^* + \tilde{k}_R(m_R - m_1^* - m_3^*)(m_D - m_1^* - m_2^* - m_5^*) \,, \\
B_{22} &= q_V m_2^* + \tilde{k}_V(m_V - m_2^* - m_4^*)(m_D - m_1^* - m_2^* - m_5^*) \,, \\
B_{33} &= q_P m_3^* + \tilde{k}_P(m_R - m_1^* - m_3^*)(m_P - m_3^* - m_4^*) \,, \\
B_{44} &= q_M m_4^* + \tilde{k}_M(m_V - m_2^* - m_4^*)(m_P - m_3^* - m_4^*) \,, \\
B_{55} &= (q_A + \tilde{q}_{AV}(m_V - m_2^* - m_4^*))m_5^* + \tilde{k}_A(m_A - m_5^*)(m_D - m_1^* - m_2^* - m_5^*) \,, \\
-A_{11} &= q_R + \tilde{k}_R(m_R - m_1^* - m_3^*) + \tilde{k}_R(m_D - m_1^* - m_2^* - m_5^*) \,, \\
-A_{12} &= \tilde{k}_R(m_R - m_1^* - m_3^*) \,, \\
-A_{13} &= \tilde{k}_R(m_D - m_1^* - m_2^* - m_5^*) \,, \\
-A_{15} &= \tilde{k}_R(m_R - m_1^* - m_3^*) \,, \\
-A_{21} &= \tilde{k}_V(m_V - m_2^* - m_4^*) \,, \\
-A_{22} &= q_V + \tilde{k}_V(m_V - m_2^* - m_4^*) + \tilde{k}_V(m_D - m_1^* - m_2^* - m_5^*) \,, \\
-A_{24} &= \tilde{k}_V(m_D - m_1^* - m_2^*) \,, \\
-A_{25} &= \tilde{k}_V(m_V - m_2^* - m_4^*) \,, \\
-A_{31} &= \tilde{k}_P(m_P - m_3^* - m_4^*) \,, \\
-A_{33} &= q_P + \tilde{k}_P(m_P - m_3^* - m_4^*) + \tilde{k}_P(m_R - m_1^* - m_3^*) \,, \\
-A_{34} &= \tilde{k}_P(m_R - m_1^* - m_3^*) \,, \\
-A_{42} &= \tilde{k}_M(m_P - m_3^* - m_4^*) \,, \\
-A_{43} &= \tilde{k}_M(m_V - m_2^* - m_4^*) \,, \\
-A_{44} &= q_M + \tilde{k}(m_V - m_2^* - m_4^*) + \tilde{k}_M(m_P - m_3^* - m_4^*) \,, \\
-A_{51} &= \tilde{k}_A(m_A - m_5^*) \,, \\
-A_{52} &= \tilde{q}_{AV}m_5 + \tilde{k_{MA}}(m_A - m_5) \,, \\
-A_{54} &= \tilde{q}_{AV}m_5^* \,, \\
-A_{55} &= (q_A + \tilde{q}_{AV}(m_V - m_2^* - m_4^*)) + k_A(m_D - m_1^* - m_2^* - m_5^*) + \tilde{k}_A(m_A - m_5^*) \,.
\end{aligned}
$$

The deterministic mean-field approximations above will allow the calibration of the model with reference to data. Also, these can be used to perform parameter identifiability and sensitivity analysis. From there, fixed parameter values allow for the determination of the steady-state values and correlations.

4.3 Sensitivity analysis

In what follows, I will restrict my study to EBOV. In the case of EBOV, the antagonistic viral protein to type I IFN secretion pathways is VP35 (Basler *et al.*, 2000; Dilley *et al.*, 2017; Kimberlin *et al.*, 2010). In this case, RIG-1 binds to dsRNA. VP35 also binds to dsRNA, forming a competition process for this resource. Given the limited experimental data available to parameterise these models, it
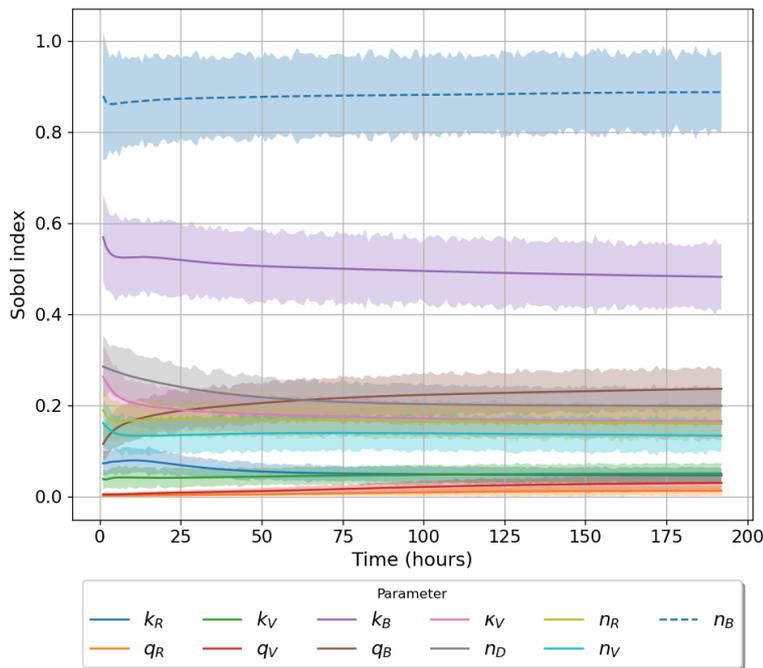
Figure 4.5: Time evolution of total-order Sobol sensitivity indices (model 1). Model output for this model is activated TBK1, n_{B^*}. The shaded region accounts for a 95% confidence interval.

is vital to understand the significance of each parameter on the corresponding model output. To this end, Sobol sensitivity analysis can be used as presented in Section 2.5.1. The choice of parameter ranges is summarised in Table 4.3 and uses the Python package "*SALib*". For each model, $10^4(2N + 2)$ parameter sets were generated using a Satelli sampler, where N is the number of parameters for the model being considered (Homma & Saltelli, 1996; Wainwright *et al.*, 2014). For model 1 (Figure. 4.1) and model 2 (Figure. 4.3), I examine how variation in

parameter values affects the output, phosphorylated TBK1 ($B^\star$ in figures). For model 3 (Figure. 4.4), I instead examine the effect of changes in parameter values on the output, phosphorylated PKR:dsRNA complexes ($A : D$ in figures).

	Model 1		Model 2		Model 3	
	Parameter	Index	Parameter	Index	Parameter	Index
Most important	n_B	0.887	n_B	0.932	n_A	0.868
	k_B	0.482	k_B	0.462	n_D	0.698
	q_B	0.236	q_B	0.289	k_R	0.266
	n_D	0.199	κ_V	0.225	n_R	0.241
	κ_V	0.165	n_R	0.201	k_A	0.238
	n_R	0.159	n_V	0.142	k_M	0.228
	n_V	0.133	n_D	0.135	q_{AV}	0.227
	k_V	0.050	n_P	0.117	n_V	0.226
	k_R	0.046	k_R	0.100	k_V	0.226
	q_V	0.029	k_P	0.093	k_P	0.226
	q_R	0.012	k_M	0.074	n_P	0.182
			k_V	0.071	q_A	0.07
			q_P	0.053	q_M	0.058
			q_M	0.051	q_P	0.053
			q_V	0.038	q_V	0.030
Least important			q_R	0.026	q_R	0.014

Table 4.1: Total-order Sobol sensitivity indices for each proposed model. Parameters are listed from most important to least, according to the sensitivity index. 10^4 samples were generated with sensitivity to n_{B*} for model 1, n_{B*} for models 2, and n_{AD} for model 3.

Table 4.1 lists the total-order of Sobol sensitivity indices for each of the three proposed mathematical models. For models 1 and 2, the analysis indicates that the most critical parameter is n_B, the total number of TBK1 molecules. Changes in this parameter result in significant fluctuations in the chosen model output. The TBK1 phosphorylation rate, k_B, exhibits a large Sobol index of 0.482 and 0.462 for models 1 and 2, respectively. The binding rates k_V and k_R, along with their associated unbinding rates, carry an insignificant contribution to variation in
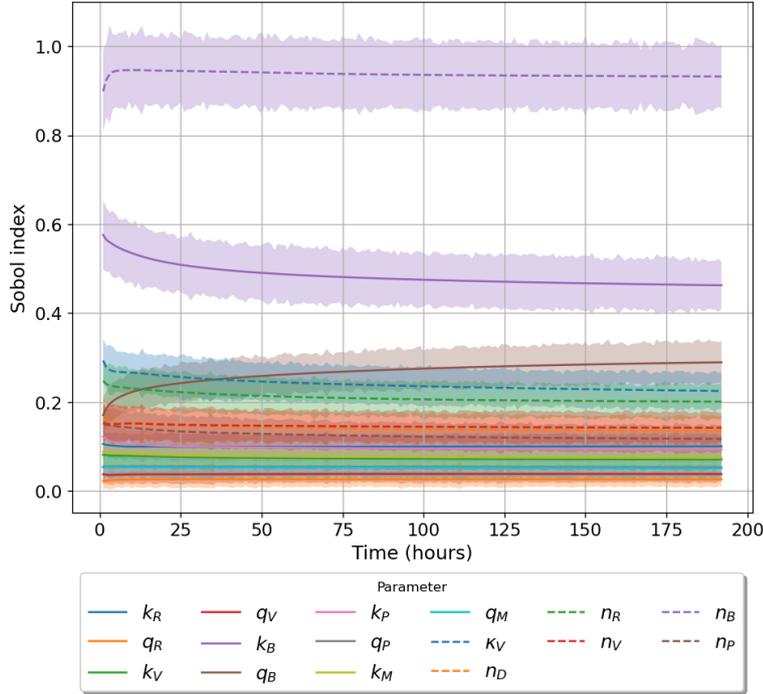
Figure 4.6: Time evolution of total-order Sobol sensitivity indices (model 2). Model output for this model is activated TBK1, n_{B^*}. The shaded region accounts for a 95% confidence interval.

model output (index ≤ 0.1) for both models 1 and 2. A similar trend is observed for the binding and unbinding rates involving the protein PACT in model 2. When considering the sensitivities of model 3, it is shown that unbinding rates are insignificant, with a low index (< 0.1). Parameters n_A and n_D are the most important ones in model 3. These have indices of 0.868 and 0.698, indicating any change in these parameters results in large model output fluctuations. The remaining parameters have a roughly equal level of importance. It is important to note that the total number of molecules for each model output has the largest sensitivity index in all three cases. In models 1 and 2, this is followed by the TBK1 phosphorylation rate k_B, while for model 3, it is the total number of viral RNA, n_D. Figures 4.5, 4.6 and 4.7 illustrate the time evolution of the sensitivity indices for each model. As observed, each model's two most important parameters remain so for the entire time course. In models 1 and 2, the importance of q_B
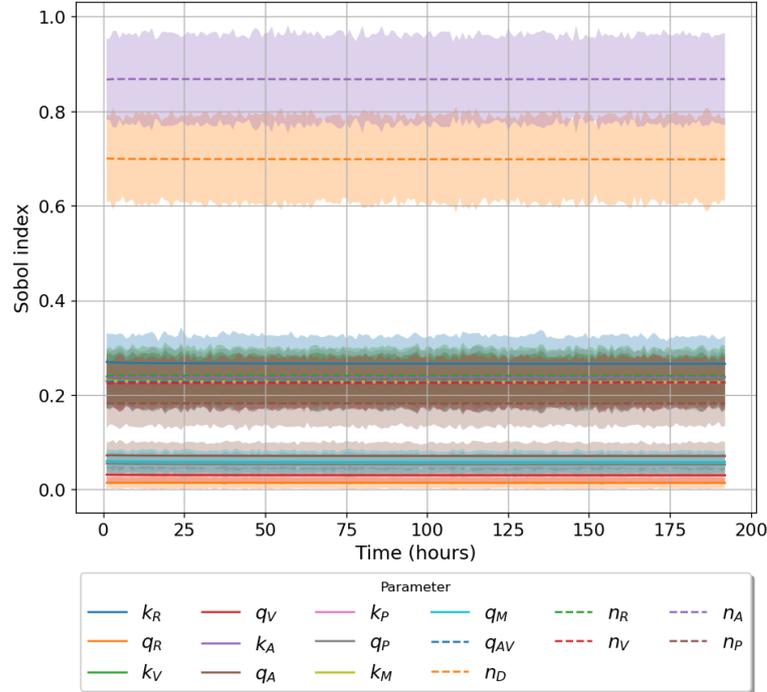
Figure 4.7: Time evolution of total-order Sobol sensitivity indices (model 3). Model output for this model is phosphorylated PKR, n_{AD}. The shaded region accounts for a 95% confidence interval.

increased with time. Other parameters show slight variation over time in their index value. Figure 4.7 shows that the Sobol sensitivity indices for model 3 remain constant with respect to time.

4.4 Parameter calibration and model selection

4.4.1 Kotliar *et al.* (2020) data

Kotliar *et al.* used single-cell transcriptomics and CyTOF-based single-cell protein quantification to characterise peripheral immune cells during EBOV infection in rhesus monkeys. Their analysis allowed them to conclude that the interferon response is suppressed in infected cells. Here their transcriptomic data set will be used for the parameter calibration of the mathematical models introduced in the previous sections (see Section 4.1). In particular, approximate Bayesian

computation sequential Monte Carlo will be implemented using the algorithm in Section 2.4.2. Eighteen non-human primates (NHPs) were exposed to the EBOV/Kikwit isolate (*Kikwit-9510621*) diluted to a target concentration of 10^3 plaque forming units (PFU) in a volume of 1 mL/dose. Two baseline blood samples were collected between 0-14 and 14-30 days before infection. Post-infection (PI) clinical observations and whole blood collection were carried out until eight days PI (see Ref. (Kotliar *et al.*, 2020, Figure 1)). 1.52×10^5 genes were tested with single-cell RNA-sequencing. Transcript counts (10^4) for IFN-β were extracted from the data to perform parameter calibration for each mathematical model. The data are summarised in Table 4.2.

Day	Mean counts per 10^4 (SD)
0	0 (0)
3	1 (0)
5	3 (2.83)
6	2.6 (1.26)
8	0 (0)

Table 4.2: Mean transcript counts (with standard deviation) for IFN-β from a longitudinal study of EBOV infection in macaques (Kotliar *et al.*, 2020).

4.4.2 Parameter identifiability

Before carrying out model calibration, it is important to study the structural identifiability of the parameters. Since many of the parameters of these mathematical models have not been previously determined, it is worthwhile to determine whether an estimate of their values is given the limited data set (Castro & de Boer, 2020). I have made use of Structural Identifiability Analyzer (SIAN), and the results are as follows; for model 1, q_B, k_B, and n_B are locally identifiable parameters. This is because the variable $n_{B^*}(t)$, phosphorylated TBK1, is the model output compared to data. Similarly, SIAN also showed that q_V and q_R are also locally identifiable (Ilmer *et al.*, 2021). Since q_R and k_R have been previously determined (Lou *et al.*, 2017), these were omitted when considering the identifiability

of our model (Lu *et al.*, 2010). The remaining parameters are all unidentifiable. Thus, it is reasonable to conclude that model 1 is structurally unidentifiable.

This analysis is repeated for model 2, where a similar trend is observed. Disassociation rates for complexes $R : D$, $V : D$, $R : P$ and $V : P$ are all locally identifiable, as well as k_B, q_B and n_B. The rest of the parameters of this model are unidentifiable, so model 2 is also unidentifiable. Finally, for model 3, it is shown that all its parameters are locally identifiable. Therefore while carrying out parameter calibration, it is important to remember that with the data at hand, for models 1 and 2, there may be a limit to what knowledge can be obtained regarding posterior distributions of their parameters. On the other hand, since parameters for model 3 are locally identifiable, I should be able to characterise the posterior distributions from this limited data set. However, multi-modal distributions may be obtained since this is only local identifiability.

4.4.3   Parameter calibration

Three different mathematical models which require parameter calibration have been introduced. Data presented in Table 4.2 will be used to calibrate these models using approximate Bayesian computation sequential Monte Carlo (ABC-SMC) as described in Section 2.4.2. The rates associated with RIG-I, k_R and q_V, are fixed using values obtained from the literature (Lou *et al.*, 2017). VAP rates, k_V and q_V, are chosen to remain within the value of its dissociation constant (Basler *et al.*, 2000; Edwards *et al.*, 2016). As described by Toni *et al.* (2009) $K = 18$ iterations are implemented of the ABC-SMC method with $n = 2,500$ accepted parameter sets. Assume all parameters follow a uniform prior distribution as defined in Table 4.3. Uniform distributions are taken from the exponent base 10 to maximise the exploration of parameter ranges. Given a set of parameters, $\boldsymbol{\theta}$, for any of the three mathematical models introduced in the previous sections, I define a Euclidean distance measure to be

$$d(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) = \sqrt{\sum_{t \in \mathcal{T}} (x(t) - y(t))^2} \, ,$$

where $\mathcal{T}$ is the set of time points in the data set, $x(t)$ denotes the output from the mathematical model at time t for parameters $\boldsymbol{\theta}$, and $y(t)$ represents experimental data at time t. Note that in this case, $x(t)$ will be the variable representing phosphorylated TBK1 for models 1 and 2 ($n_{B^*}(t)$, n_3 in model 1 and n_5 in model 2). In comparison, for model 3, I consider phosphorylated PKR ($n_{AD}(t)$, n_5 in model 3). I assume a linear relationship between transcripts and protein numbers. The first iteration threshold ε_1 is defined as the median of 10^6 initial realisations via ABC with prior samples obtained from the distributions in Table 4.3. Then define threshold ε_k as the median distance from iteration $k-1$. The perturbation kernel will be uniform and used to perturb the parameter values during each iteration.

Figures 4.8, 4.10 and 4.11 present posterior histograms from the final iteration along with the model median and the 95% credible intervals. Tables 4.4, 4.5, and 4.6 present posterior median and mean values for each parameter, with 95% credible intervals also reported. For models 1 and 2, $n_{B^*}(t)$ (n_3 and n_5, respectively) is fitted to data, while for model 3, it is $n_{AD}(t)$ (n_5).

Posterior histograms in Figure 4.8 for model 1 illustrate inference can characterise the value q_B, with a narrow posterior compared to its prior distribution. This is particularly important since sensitivity analysis indicated this was the third most important parameter, as shown in Table 4.1. Additionally, narrower posterior distributions were obtained for k_V, q_V and n_V. Sensitivity analysis indicated that k_B and n_B were the two most important parameters to control within this model. However, if the correlations are considered between these two parameters, it shows a strong negative correlation between them ($r = -0.83$). Therefore, it is only possible to learn about their ratio. There also exists a positive correlation ($r = 0.65$) between n_D and n_V as shown in Figure. 4.9, so that with the data set and Bayesian inference, only the ratio of these two parameters can be estimated.

Inference for model 2 (see Figure. 4.10) indicates that very little is learnt for most parameters, including the rates associated with PACT binding to VAP, k_M, and its corresponding disassociation rate, q_M. Sensitivity analysis revealed the rates for TBK1 phosphorylation (k_B) and de-phosphorylation (q_B), along with n_B, the number of TBK1 molecules, is essential to minimise variation in model

131

(a) Model 1 posterior histograms

(b) Model 1 ABC-SMC time course

Figure 4.8: (a) Posterior histograms obtained from the ABC-SMC algorithm. Blue histograms indicate prior distributions, and purple histograms illustrate posteriors. (b) Model fit from accepted parameter sets obtained during the final iteration of the ABC-SMC algorithm. Blue triangles represent data presented in Table 4.2 plotted with its standard deviation. The black line illustrates the point-wise median value from the accepted parameter sets (shaded in green) with a 95% credible interval. These results represent 18 iterations with 2500 accepted parameter sets for model 1.

Parameter	Search range	Units	Reference	Model
k_R	0.04836	h^{-1} *per molecule pair*	Lou *et al.* (2017)	1,2,3
q_R	7.632	h^{-1} *per molecule*	Lou *et al.* (2017)	1,2,3
k_V	$U[-7,2]$	h^{-1} *per molecule pair*	Basler *et al.* (2000); Edwards *et al.* (2016)	1,2,3
q_V	$U[-6,2]$	h^{-1} *per molecule*	Basler *et al.* (2000); Edwards *et al.* (2016)	1,2,3
k_P	$U[-7,-2]$	h^{-1} *per molecule pair*		2,3
q_P	$U[-6,0]$	h^{-1} *per molecule*		2,3
k_M	$U[-7,-2]$	h^{-1} *per molecule pair*		2,3
q_M	$U[-6,0]$	h^{-1} *per molecule*		2,3
k_B	$U[-6,8]$	h^{-1} *per molecule*	Zhang & Zou (2013); Zou *et al.* (2010)	1,2
q_B	$U[-6,3]$	h^{-1}*per molecule*	Zhang & Zou (2013); Zou *et al.* (2010)	1,2
k_A	$U[-7,2]$	h^{-1} *per molecule pair*		3
q_A	$U[-6,2]$	h^{-1} *per molecule*		3
q_{AV}	$U[-6,2]$	h^{-1} *per molecule*		3
κ_V	$U[1,7]$			1,2
n_D	$U[0,7]$			1,2,3
n_R	$U[0,6]$			1,2,3
n_V	$U[0,6]$			1,2,3
n_B	$U[0,6]$			1,2
n_A	$U[0,6]$			3

Table 4.3: Parameter ranges used for Bayesian inference. Ranges were taken based on the published data, with two to three orders of magnitude taken on either side of these values (Basler *et al.*, 2000; Bosworth *et al.*, 2017; Edwards *et al.*, 2016; Lu *et al.*, 2010; Zou *et al.*, 2010). I consider uniform distributions of exponent base 10 for all prior choices. Parameters k_V and q_V are restricted by making sure their ratio, $\frac{k_V}{q_V}$, lies within the dissociation constant range $3.40\mu M$-$1.1nM$ reported in Refs. (Basler *et al.*, 2000; Edwards *et al.*, 2016).

output. As mentioned earlier, the inference of the parameters indicates narrow posterior histograms, which illustrates that Bayesian inference is learning about these three parameters. However, it is essential to note the median number of molecules n_B is low, which might not be a biologically reasonable value. Medians presented in Figure. 4.8 and Figure. 4.10 have a comparable trend. Note that model 2, additionally, has a rather sharp jump at the start of the time course. This is most likely due to the inclusion of PACT as a secondary activator. Notably, the credible intervals of model 1 are narrower than those of model 2.

Parameter	Median	Mean	Credible interval
k_V	4.37×10^{-6}	2.04×10^{-2}	$[1.42 \times 10^{-7}, 5.58 \times 10^{-2}]$
q_V	1.03×10^{-5}	3.42×10^{-5}	$[1.35 \times 10^{-6}, 2.22 \times 10^{-4}]$
k_B	1.50	4.69×10^{2}	$[5.78 \times 10^{-4}, 5.62 \times 10^{-3}]$
q_B	1.29×10^{-2}	2.74×10^{-2}	$[6.24 \times 10^{-3}, 1.49 \times 10^{-1}]$
κ_V	6.70×10^{3}	374×10^{5}	$[2.00 \times 10^{1}, 4.08 \times 10^{6}]$
n_D	3.37×10^{2}	3.34×10^{3}	$[1, 2.92 \times 10^{4}]$
n_R	6.80×10^{1}	2.82×10^{4}	$[1, 3.67 \times 10^{5}]$
n_V	1.39×10^{4}	4.38×10^{4}	$[5.2 \times 10^{1}, 2.69 \times 10^{5}]$
n_B	2.26×10^{2}	1.91×10^{4}	$[1, 2.54 \times 10^{5}]$

Table 4.4: Summary statistics for each accepted parameter value sets from model 1. Mean, median and a 95% credible interval are summarised.
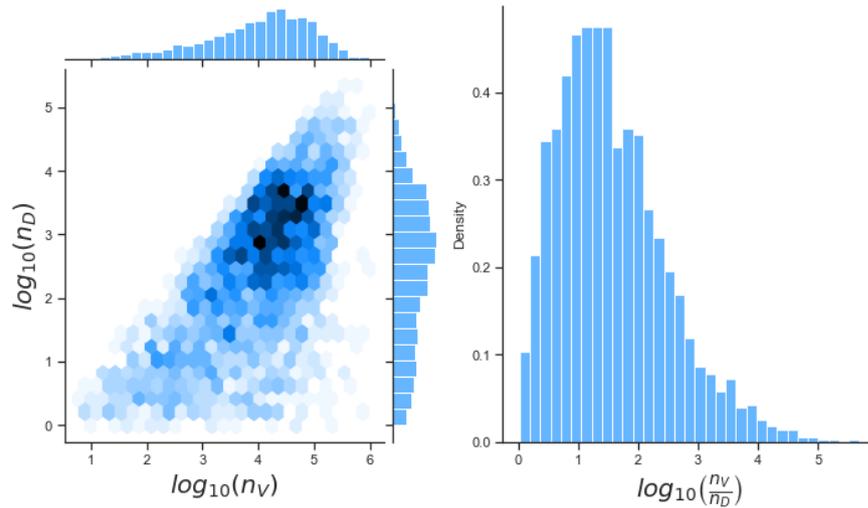
Figure 4.9: Left: Bivariate posterior histogram of $\log_{10}(n_D)$ and $\log_{10}(n_V)$ showing a positive correlation ($r = 0.65$). Right: posterior distribution for the ratio $\log_{10}(n_V/n_D)$.

Parameter	Median	Mean	Credible interval
k_V	5.19×10^{-5}	1.38	$[1.96 \times 10^{-7}, 1.52 \times 10^{1}]$
q_V	5.04×10^{-6}	6.36×10^{-2}	$[1.07 \times 10^{-6}, 4.81 \times 10^{-2}]$
k_P	1.05×10^{-5}	3.45×10^{-2}	$[1.60 \times 10^{-7}, 1.00 \times 10^{-1}]$
q_P	5.06×10^{-2}	3.03	$[1.02 \times 10^{-5}, 3.44 \times 10^{1}]$
k_M	6.08×10^{-3}	3.27	$[2.12 \times 10^{-7}, 3.89 \times 10^{1}]$
q_M	1.69×10^{-3}	6.83×10^{-1}	$[1.45 \times 10^{-6}, 4.949]$
k_B	1.55×10^{5}	4.75×10^{6}	$[1.134, 4.95 \times 10^{7}]$
q_B	1.41×10^{-1}	3.29	$[8.75 \times 10^{-3}, 27.86]$
κ_V	3.96×10^{3}	2.67×10^{5}	$[1.7 \times 10^{1}, 2.68 \times 10^{6}]$
n_D	1.14×10^{2}	2.64×10^{3}	$[2, 2.59 \times 10^{4}]$
n_R	9.2×10^{1}	1.36×10^{4}	$[1, 1.36 \times 10^{5}]$
n_V	1.76×10^{4}	6.74×10^{4}	$[1.98 \times 10^{2}, 4.57 \times 10^{5}]$
n_B	2	2.61×10^{3}	$[1, 2.97 \times 10^{2}]$
n_P	4.5×10^{1}	1.98×10^{3}	$[1, 1.72 \times 10^{4}]$

Table 4.5: Summary statistics for each accepted parameter value sets from model 2. Mean, median and a 95% credible interval are summarised.

(a) Model 2 posterior Histograms

(b) Model 2 ABC-SMC time course
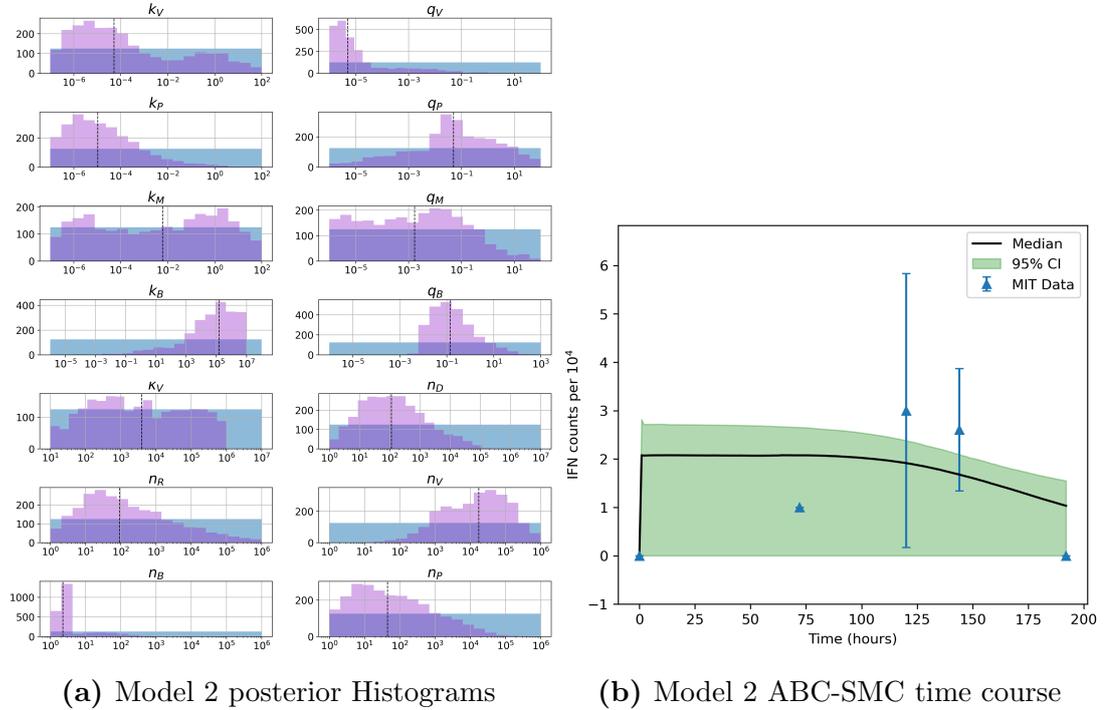
Figure 4.10: (a) Posterior histograms obtained from the ABC-SMC algorithm. Blue histograms indicate prior distributions, and purple histograms illustrate posteriors. (b) Model fit from accepted parameter sets obtained during the final iteration of the ABC-SMC algorithm. Blue triangles represent data presented in Table 4.2 plotted with its standard deviation. The black line illustrates the point-wise median value from the accepted parameter sets (shaded in green) with a 95% credible interval. These results represent 18 iterations with 2500 accepted parameter sets for model 2.

Figure 4.11 presents results from the Bayesian inference with model 3. The median course is similar to those in Figure. 4.8 and Figure. 4.10. Compared to model 2, the time course describes the data better. In contrast, the median of model 1 is similar but has a smaller credible interval. Upon examination of the posterior distributions, there is improved learning. Parameters for the number of viral RNA (n_D) and PKR molecules (n_A) have narrow posterior distributions compared to their prior ones. As mentioned in previous sections, these are important parameters identified by Sobol sensitivity analysis. Thus, such improved learning is a desired feature of model 3 compared to models 1 and

2. Many parameters have narrower posterior distributions compared to their prior ones, indicating overall learning for most parameter values. The binding and disassociation rates of VP35 and PACT have the broadest posterior distributions. Both rates could benefit from additional iteration steps of the ABC-SMC method. Strong correlations exist particularly between k_M and n_A ($r = -0.70$) and n_P ($r = -0.71$), which means one can only learn about their. Figure. 4.12 illustrates a positive correlation between q_M and n_V ($r = 0.60$), once again indicating learning about their ratio.

Parameter	Median	Mean	Credible interval
k_V	3.87×10^{-5}	3.00×10^{-2}	$[2.84 \times 10^{-7}, 1.40 \times 10^{-2}]$
q_V	1.92×10^{-3}	2.36×10^{-1}	$[2.43 \times 10^{-6}, 1.66]$
k_P	4.55×10^{-6}	7.74×10^{-2}	$[1.79 \times 10^{-7}, 3.15 \times 10^{-4}]$
q_P	1.26×10^{-5}	3.20×10^{-3}	$[1.22 \times 10^{-6}, 1.38 \times 10^{-3}]$
k_M	1.30×10^{-1}	3.30	$[2.61 \times 10^{-7}, 2.90 \times 10^{1}]$
q_M	1.14×10^{-4}	1.56	$[1.72 \times 10^{-6}, 1.81 \times 10^{1}]$
k_A	8.41×10^{-2}	2.30×10^{1}	$[6.54 \times 10^{-4}, 2.80 \times 10^{2}]$
q_A	2.32×10^{-3}	1.71×10^{1}	$[4.90 \times 10^{-6}, 2.07 \times 10^{2}]$
q_{AV}	6.31×10^{1}	9.32×10^{1}	$[1.25 \times 10^{1}, 3.82 \times 10^{2}]$
n_D	1.30×10^{1}	2.37×10^{3}	$[2, 5.98 \times 10^{2}]$
n_R	4.16×10^{3}	9.90×10^{3}	$[3.24 \times 10^{2}, 5.16 \times 10^{4}]$
n_V	1.70×10^{1}	3.46×10^{3}	$[1, 2.00 \times 10^{4}]$
n_A	1.20×10^{1}	1.61×10^{4}	$[2, 2.00 \times 10^{5}]$
n_P	3.74×10^{2}	2.19×10^{4}	$[1.40 \times 10^{1}, 1.76 \times 10^{5}]$

Table 4.6: Summary statistics for each accepted parameter value sets from model 3. Mean, median and a 95% credible interval are summarised.

(a) Model 3 posterior histograms

(b) Model 3 ABC-SMC time course
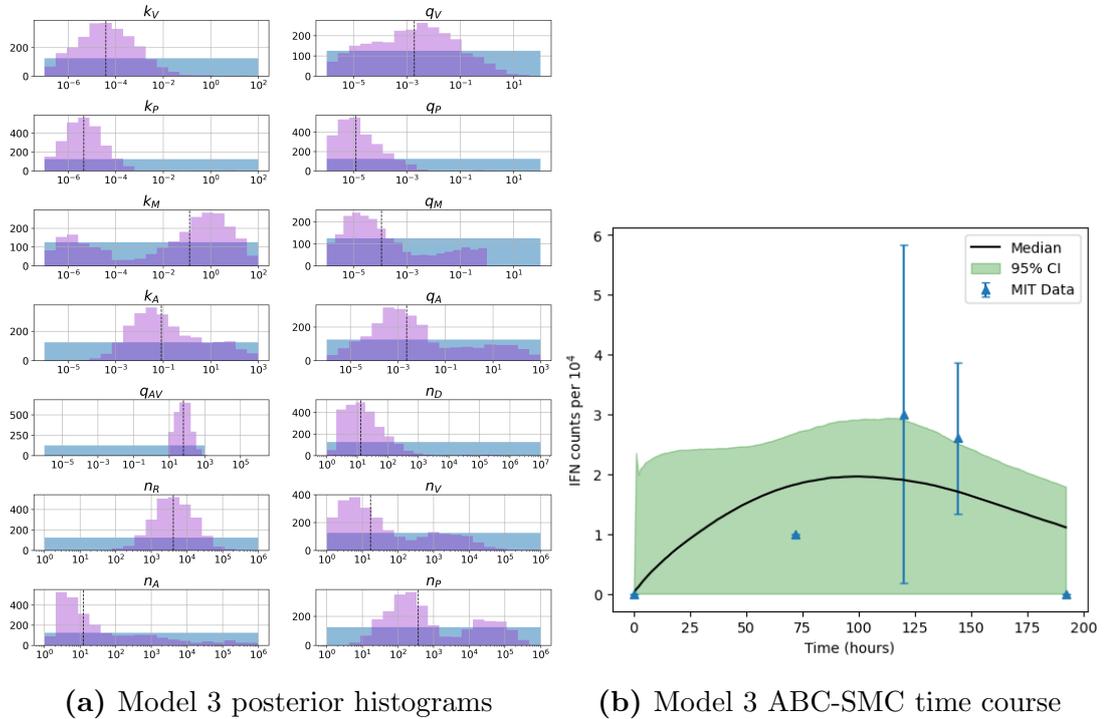
Figure 4.11: (a) Posterior histograms obtained from the ABC-SMC algorithm. Blue histograms indicate prior distributions, and purple histograms illustrate posteriors. (b) Model fit from accepted parameter sets obtained during the final iteration of the ABC-SMC algorithm. Blue triangles represent data presented in Table 4.2 plotted with its standard deviation. The black line illustrates the point-wise median value from accepted parameter sets (shaded in green) with a 95% credible interval. These results represent 18 iterations with 2500 accepted parameter sets for model 3.
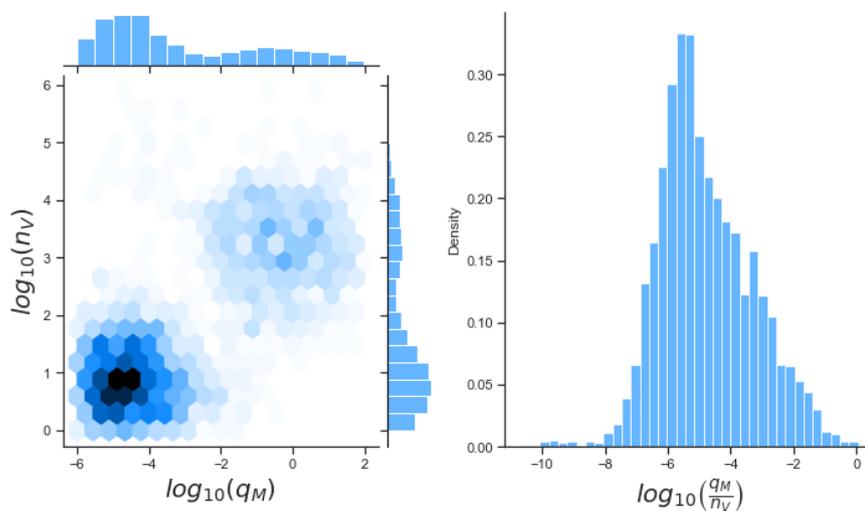
Figure 4.12: Left: Bivariate posterior histogram of $\log_{10}(q_M)$ and $\log_{10}(n_V)$ showing a positive correlation ($r = 0.60$). Right: Posterior distribution of the ratio $\log_{10}(q_M/n_V)$.

4.4.4 Model selection

The sensitivity of their associated parameters has been studied for the proposed three separate mathematical models, and Bayesian inference has been used to calibrate each model. Furthermore, their structural identifiability has also been assessed. These initial findings would indicate that either model 1 or 3 would be suitable to appropriately describe the data, since they have the best posterior histograms and overall time course. The use of ABC model selection and calculation of the Akaike information criterion (AIC) allows a way to quantify which model better describes the data. Thus, an indication of which mechanism of type I IFN inhibition is preferred (Anderson & Burnham, 2004). The second order AIC for small sample sizes is defined as

$$ AIC_c = -2log(\mathcal{L}(\boldsymbol{\theta})) + 2K_\theta \left(\frac{n_s}{n_s - K_\theta - 1} \right) , $$

with $\log(\mathcal{L}(\boldsymbol{\theta}))$ the log-likelihood given parameters $\boldsymbol{\theta}$, defined in Section 4.4.3, K_θ defined as the number of estimated parameters for a given model, and $n_s(= 5)$ the number of samples used to generate the data presented in Table 4.2. The AIC

value generated will be compared for each model, with a lower index being an indication of preferential model selection. I use a standard ABC rejection method as described in Algorithm 2.4.1, which, unlike the ABC-SMC, does not perform successive iterations. The ABC rejection method is run with $\varepsilon = 3$, and I count how many sampling instances are required to accept a total of 10^5 parameter sets. Table 4.7 summarises these results.

Models 1 and 2 have similar acceptance percentages: 5.35% of parameter sets accepted for model 1 and 5.12% for model 2. Model 3 has the largest percentage of accepted parameter sets at 9.03%, much higher than those by models 1 and 2. Note that model 3 has more parameters than models 1 and 2. Since we have a small number of data points and many parameters, I use the second-order AIC (Anderson & Burnham, 2004). This method gives AIC values of -19.25, -15.75 and -15.80 for models 1, 2 and 3, respectively. This shows that accounting for the number of parameters, model 1 is the best, followed by model 3, then 2. Taking into account the ABC rejection results, these results illustrate that model 2 characterises the data poorly and is less suitable than models 1 and 3. Now carrying out a pair-wise comparison between models 1 and 3. It shows a probability of 0.372 for choosing model 1 and 0.628 for model 3. Bayesian model selection indicates that model 3 better explains the limited data available.

	Number of samples	% accepted	Relative probability
Model 1	1,868,652	5.35	0.274
Model 2	1,952,835	5.12	0.263
Model 3	1,107,525	9.03	0.463

Table 4.7: Table with number of sample parameter sets required for 10^5 sets to be accepted. An Euclidean distance measure was used with $\varepsilon = 3$ as a threshold value for acceptance. The percentage of accepted parameter sets is also shown.

4.5 Numerical Simulations

The previous section showed that model 2 poorly represented the data. Furthermore, model selection revealed it was the worst model with a low acceptance

probability and the highest AIC value compared to models 1 and 3. Therefore, it is reasonable to exclude model 2 from further analysis, and I will focus on results generated using models 1 and 3. In what follows, median parameter values for each model found during Bayesian inference are used (Tables 4.4 and 4.6 for models 1 and 3, respectively). Furthermore, the analysis will be restricted to stochastic simulations since the matrices and vectors generated to use matrix analytical methods are too large to store in system memory with the current high-performance systems available. An example of how these matrix methods could be used in practice is described in Appendix A.2. In Section 4.1, it was highlighted that from these models understanding the late-time behaviour of the stochastic processes was of interest by investigating the steady-state distribution. The other stochastic descriptor of interest is the mean time to reach a threshold number N of activated TBK-1 (phosphorylated PKR for model 3). From the van Kampen expansion carried out at the start of this chapter for each model, correlations of stochastic variables can also be investigated.
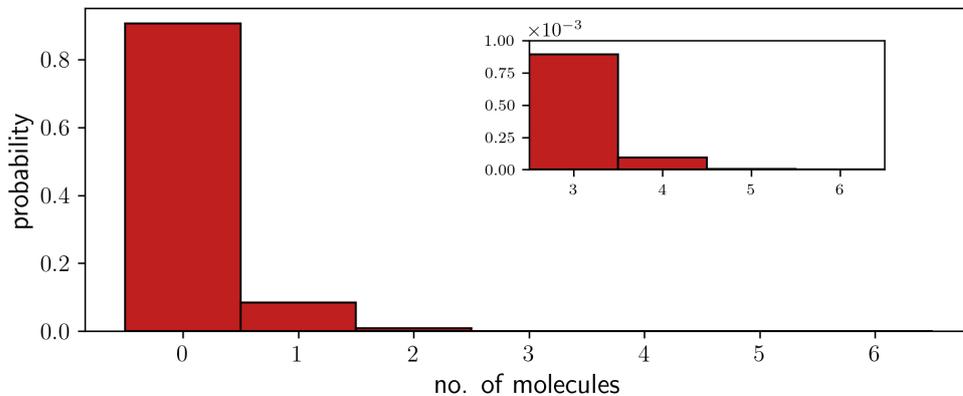
Figure 4.13: Probability distribution for process $\mathcal{X}_1$ for number of phosphorylated TBK-1 in steady-state from numerical simulations. The initial state for simulations was $\boldsymbol{n}(0) = (0, 0, 0)$.

Figure 4.13 illustrates the stationary probability distribution for stochastic process $\mathcal{X}_1$. This indicates that for late-time behaviour, there is a significant probability > 0.8 that there will be no activated TBK-1 present. This results from large amounts of viral protein VP35 competitively sequestering dsRNA with
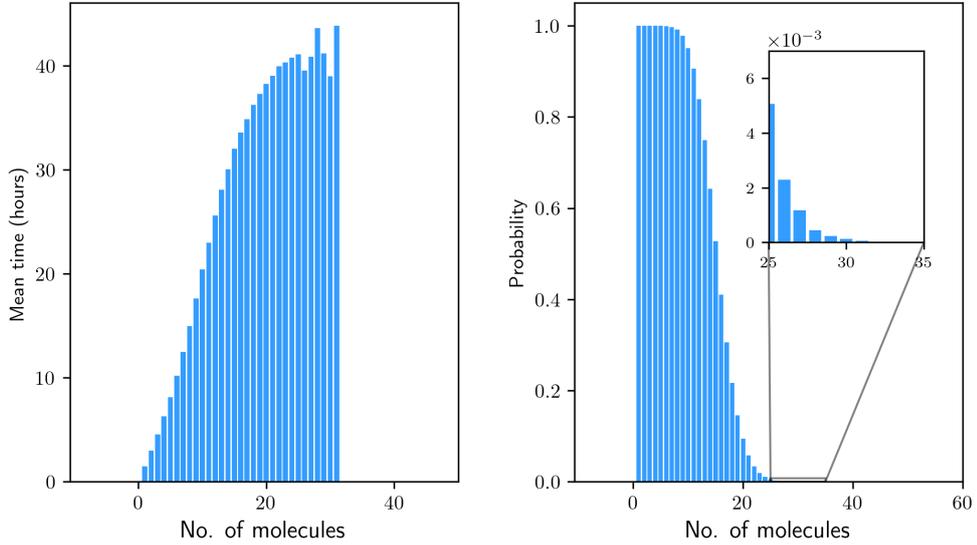
Figure 4.14: *Left:* Mean time to reach N phosphorylated TBK-1 for process $\mathfrak{X}_1$. *Right:* Probability of reaching N phosphorylated TBK-1 within 192 hours. The initial state for simulations was $\boldsymbol{n}(0) = (0, 0, 0)$.

RIG-1. As a result, fewer RIG-1:dsRNA complexes are formed, which reduces the induction of TBK-1. The distribution shifted to the left indicates a small probability of reaching a steady-state of 1 or two activated TBK-1 molecules < 0.1. Additionally, there is a small probability of 3-5 activated TBK-1 molecules, as seen in the zoomed box in Figure 4.13. Figure 4.14 on the left panel presents the mean time to reach a given number N of activated TBK-1 molecules (where an empty bar means no state was reached). The right panel shows the probability of reaching a given number of activated TBK-1. As seen from the left panel, the likelihood of achieving a "N" molecules decreases as the threshold number of molecules increases following an inverted logistic sigmoid curve until around 35 molecules, where the probability becomes effectively zero. The mean time follows a logistic S curve, with the time to reach a given state increasing until it reaches the maximum possible molecule level where the probability becomes zero.

Examining the stochastic correlations of Model 1 shown in Figure 4.15, it is interesting to note that for stochastic variables n_1 RIG-1:dsRNA and n_2 VP35:dsRNA there exists a weak negative correlation, showing that as one
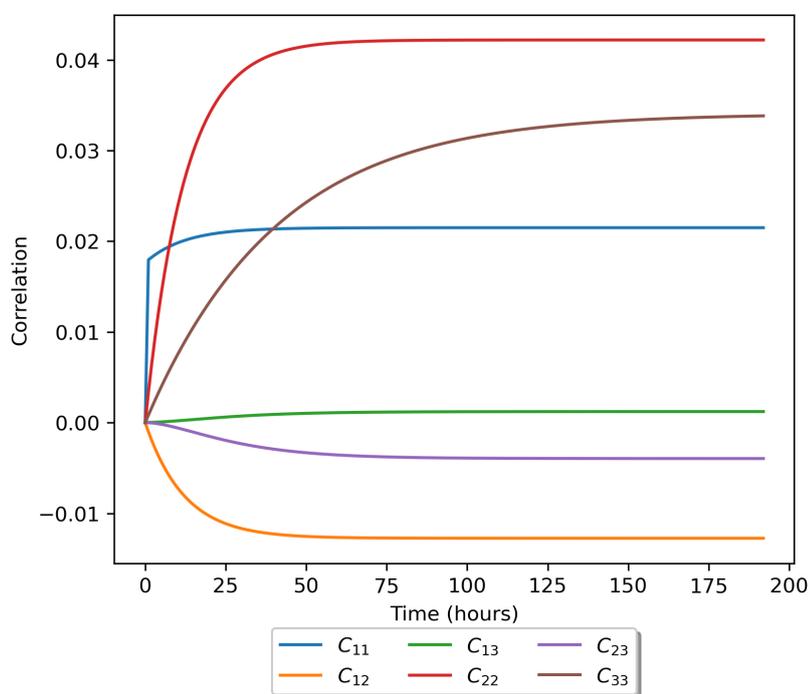
Figure 4.15: Correlations with respect to time for the variables of the stochastic process $\mathfrak{X}_1$ for model 1.
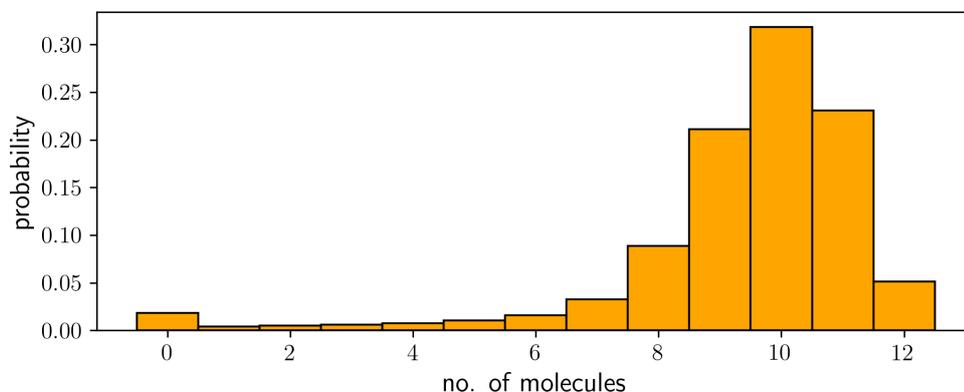
Figure 4.16: Probability distribution for the number of phosphorylated PKR in steady-state from numerical simulations for process $\mathfrak{X}_3$.

stochastic variable increases, the other will decreases. This makes sense due to the competitive nature of VP35 and RIG-1 that we have established in this model.
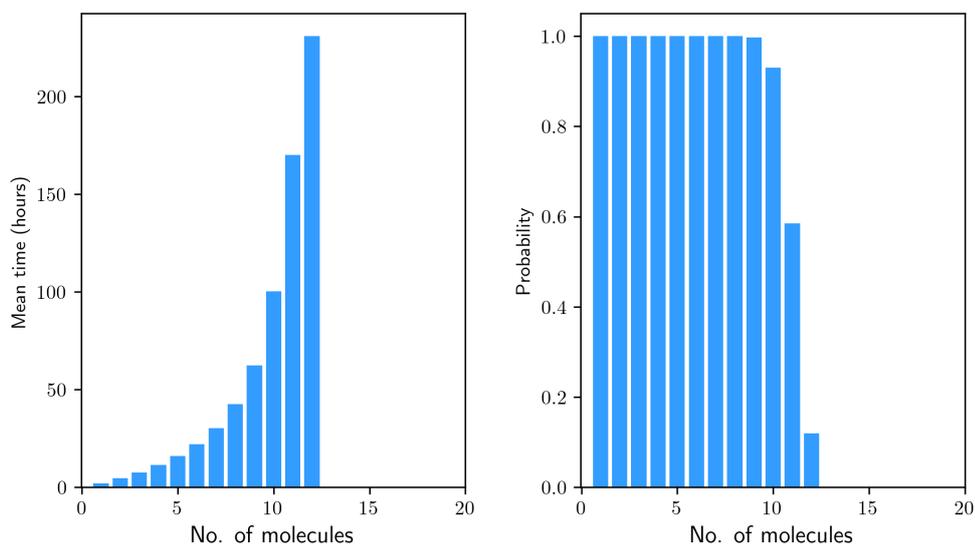
Figure 4.17: *Left:* Mean time to reach N phosphorylated PKR for process $\mathfrak{X}_3$. *Right:* Probability of reaching N phosphorylated PKR within 192 hours. The initial state for simulations was $\boldsymbol{n}(0) = (0, 0, 0, 0, 0)$.
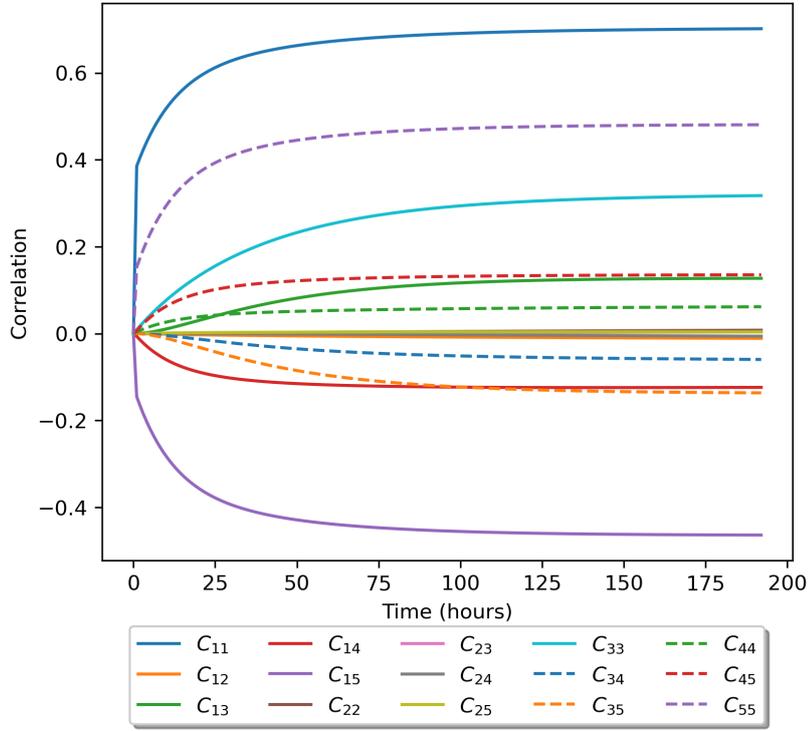
Figure 4.18: Correlations with respect to time for the variables of the stochastic process $\mathcal{X}_3$ for model 3.

From the median parameter values for model 3, the stationary probability distribution is shifted to the right with between 9-11 phosphorylated PKR molecules being present in steady-state. For late-time behaviour, this shows that the signalling via PKR can still be achieved even with being de-phosphorylated with a large probability. This is more consistent with what is witnessed *in vivo* experimentation since it is well documented not all IFN signalling is abolished. Figure 4.17 shows the probability of reaching N phosphorylated PKR on the right panel, while the left shows the mean time to reach this state. As can be seen states up to 8 molecules are guaranteed to be reached, with the amount of time to reach these exponentially increasing. For states 9-12, there is a large increase in time to reach the given state. Furthermore, the probability of reaching these states negatively correlates with the time to reach them. Figure 4.18 shows the negative correlations obtained from the van Kampen expansion. This figure illustrates the negative correlation between PKR:dsRNA and RIG-I:dsRNA complexes. This

145

indicates the negative effect of forming RIG-I:dsRNA complexes on PKR:dsRNA complexes. Furthermore, Figure 4.18 also shows a negative correlation between PACT:VP35 and PKR:dsRNA complexes. This is because PACT binding to VP35 allows more PKR to successfully bind to dsRNA without being down-regulated by VP35. Therefore, this highlights the benefits of PACT and PKR being in the same system.

4.6 Discussion

In this chapter, I have focused on viral strategies to antagonise type I IFN secretion pathways via PRRs. I have also proposed three potential mathematical models with upstream and downstream type I IFN inhibition mechanisms. These models have been formulated based on the current biological understanding of the interactions between intra-cellular proteins. In particular, I have investigated the role of RLRs, which detect viral RNAs and promote IFN responses. VAPs potentially antagonise a given RLR signalling pathway in its upstream portion, at the level of viral RNA recognition, by competing with RLRs for viral RNA binding or removing PAMP signatures recognised by RLRs. Viral proteins such as NS1 (Influenza A), VP35 (EBOV), and N (SARS-COV) all bind viral RNAs, thus inhibiting PRRs from binding and signalling (Basler *et al.*, 2000; Cui *et al.*, 2015; Donelan *et al.*, 2003; Edwards *et al.*, 2016; Zinzula & Tramontano, 2013). PAMPs such as NP (Lassa Fever) and NSP14 (SARS-COV) are removed, preventing RLR recognition. Downstream effects include the inhibition, mediated by VAPs, of RLR-induced antiviral proteins. For instance, VAPs may modify the binding sites of proteins, inhibit the formation of signalling complexes, or prevent translocation and phosphorylation (Zinzula & Tramontano, 2013).

Mathematical models have been previously proposed to model interferon inhibition by viruses (Rand *et al.*, 2012; Rinas, 2016; Zou *et al.*, 2010) or to describe inter-cellular interactions via IFN-α receptors (Schmid *et al.*, 2015). These models are virus-specific or require detailed knowledge of many protein-protein interactions along the signalling pathway under consideration. My aim with this chapter was to characterise key biological hypotheses quantitatively while avoiding, in principle, unnecessary complexity. Since clinical data sets from early

viral infections are typically limited, it is essential to have mathematical models in place that can be parameterised given this severe restriction. While the proposed models do not include all possible mechanisms of viral protein antagonism and inhibition of signalling pathways, which result in type I IFN induction, the three models presented are a good first start. Moreover, these models can be generalised to account for other mechanisms, proteins, or additional signalling pathways and separate viruses. I proposed three different models for the inhibition of type I interferon expression by VAPs. Each model considers a different biological mechanism or an alternate signalling pathway. Figures 4.1, 4.3 and 4.4 represent mechanisms which have been recently proposed in the literature (Donelan *et al.*, 2003; Edwards *et al.*, 2016; Randall & Goodbourn, 2008; Zou *et al.*, 2010). For each model I have assessed its sensitivity and parameter identifiability, as well as carried out model selection and parameter calibration. In particular, Sobol sensitivity analysis identified which parameters would need to be closely controlled for each model and its output. In each model, two parameters must be well characterised to avoid large variations in my model outputs. For models 1 and 2, these are the total number of TBK1 molecules, n_B, and its activation rate, k_B. In the case of model 3, the most important parameters are the total number of PKR molecules, n_A, and the total copy number of viral RNA molecules, n_D.

Unfortunately, very little is known about the values for the parameters considered in my models. Thus, I aimed to conduct Bayesian inference to narrow down these values. To this end, carrying out a structural identifiability analysis was also important. This analysis led to the following results: model 3 (considering the PKR pathway) is locally identifiable, but models 1 and 2 are not. I note that these results are in light of the limited data set I had at hand. Yet, this indicated that even with limited data, model 3 might be better, compared to the other two models, at allowing us to infer parameter values. This was further supported by model selection and parameter inference: the PKR signalling pathway has a higher percentage of acceptance as illustrated in Table 4.7 and narrower posterior distributions for most parameter values (see Figure. 4.11). Model 2 was deemed the worst model of the three: many parameters were non-identifiable, leading to the worst percentage of parameter set acceptance from model selection, and broad posterior distributions for its parameters. Model 1 cannot be rejected since it

had the lowest AIC coefficient. However, as previously mentioned, this model is structurally unidentifiable and leads to poor learning for most parameter values.

In this section, parameter estimation leads to poor estimates for the parameters in models 1 and 2. In general, while the ABC-SMC method allows one to start with large prior distributions and then narrow them, it reduces the potential to learn about parameter values in this implementation. The first way to improve this model fit would be to use an Euclidean alternative where the distances are divided by the standard deviation at each time point. This allows one to fit closer to the points with less variation than those with more considerable variation, which may help narrow posterior distributions. A second way to increase the estimated parameters' reliability would be to use a mixed-effects model. Since the experimental data used is from different rhesus monkeys, there is considerable variation on an animal-to-animal basis and cellular basis for each animal. Using a mixed-effects model would allow one to parameterise not only on a population level but also in a monkey-specific way. This would improve the reliability of the parameters gained from the model since one would not be averaging just over the population.

While the models I have presented can be generalised to other viruses, it should be remembered that additional biological phenomena have not been included in the models. EBOV infection, which was examined as a case study, has several specific features not characterised by the mechanisms included in these models. Plasmacytoid dendritic cells (pDCs) have been shown to be refractory to EBOV infection, whereas common dendritic cells are susceptible (Leung *et al.*, 2011). This could be due to the fact that pDCs express basal (or constitutive) levels of IRF7 before infection (Honda *et al.*, 2005), and, therefore, can be considered in an antiviral state. Thus, when considering the development of a mathematical model, it is vital to understand not only the virus but also the cellular tropism of the virus and the host (*i.e.,* invertebrate or vertebrate); that is, which cells are the target cells of the virus (Perelson, 2002). The main focus of this chapter was on type I IFNs as essential antiviral cytokines, yet immune responses require a complex and coordinated interaction of an extensive collection of cytokines and cells, which have not been investigated. Developing such complex mathematical models would require a more comprehensive data set. These models could be further expanded

to incorporate other intra-cellular or viral mechanisms, or additional signalling pathways. Additional data sets, *e.g.,* from quantitative proteomics, could be used to improve parameter inference for EBOV and other viruses of global concern and public health threats.

A final point to consider is the difference between *in vivo* and *in vitro* infection. These models have been parameterised with an *in vivo* data set. As discussed in the introduction of this chapter, there is a stark contrast in responses when comparing *in vivo* and *in vitro* infection. In particular, and for the *in vitro* case, type I IFN production is abrogated after three days post-infection. whereas in the case of *in vivo* infection type I IFN is present throughout the entire infective period (Bosworth *et al.*, 2017; Kotliar *et al.*, 2020). Hence it is critical to keep this in mind when developing a mathematical model. Along this line of thought, it is also important to note that differences in *in vivo* experimental models can lead to different innate immune responses. For instance, bats are a proposed reservoir for EBOV but are known to be asymptomatic for the disease. Experiments have indicated that bats have detectable viral RNA levels, but no detectable viremia (Paweska *et al.*, 2016). Yet, in the case of humans and non-human primates, the clinical presentation tends to be symptomatic and with measurable viremia (Feldmann & Geisbert, 16th November 2010; Geisbert *et al.*, December 2003). Bat dendritic cells have shown an enhanced capacity to initiate IFN-dependent responses upon filovirus infection in comparison with, for example, human cells (Olival *et al.*, 2020). Other studies have reported a difference in immune responses depending on the specific tissue analysed: EBOV viral RNA levels persist in male gonads even after a negative PCR test from blood samples (Mate *et al.*, 2015). These factors would need to be considered for further model development and experimental design for data used to calibrate these models.

4. MATHEMATICAL MODEL OF INTERFERON ANTAGONISM

Chapter 5

Mathematical models of viral infection

Mathematical models combined with computational modelling are handy tools when it comes to the study of viral infections. They offer relatively cheap and quick methods that can be used instead of using *in vivo* or *in vitro* experiments for predictive analysis. The role of mathematical models has never been more relevant than in the last several years during the 2019 SARS-CoV-2 pandemic. The predictive analysis allowed governments and health organisations to make informed choices about public health policy and investigate the potential outcomes if preventative measures were not implemented and followed by the general public (Metcalf *et al.*, 2020). There is a long-standing history of mathematical models being used for modelling viral and bacterial infections such as, HIV (Perelson, 1999), Ebola (Liao *et al.*, 2020) and Anthrax (Williams *et al.*, 2021) to name a select few. In this chapter, I will present two mathematical models: the first is for intra-cellular replication of SARS-CoV-2 in the presence of defective interfering particles. The second model was used to investigate the replication kinetics of Foot and Mouth Disease virus (FMDV) in vitro. The latter work was preliminary in support of a grant application.

5.1 Defective interfering particles as a potential therapeutic during SARS-CoV-2 infection

In December 2019, a new infectious agent was reported to the World Health Organisation (WHO) that would later be identified as a novel coronavirus (SARS-CoV-2) (Wu *et al.*, 2020). The WHO declared SARS-CoV-2 a "public health emergency of international concern" on the 30th January 2020 (Li *et al.*, 2020a), rapidly spreading to 113 countries. By the 11th March 2020, it had caused 118,319 infections and 4,292 deaths, and as a result, many countries began to initiate lockdowns and introduce mask and social distancing mandates. Consequently, the WHO declared SARS-CoV-2 a pandemic (Martellucci *et al.*, 2020; World Health Organization *et al.*, 2020) and as of 29th July 2022, about 572 million infections and over 6 million deaths have been recorded worldwide. Furthermore, the pandemic seriously impacted the global economy leaving many businesses and governments feeling the effects (Mohapatra *et al.*, 2020). During the early stages of the pandemic, treatment options were limited to chloroquine and remdesivir (Colson *et al.*, 2020; Morse *et al.*, 2020). However, several effective vaccines have since been developed that provide protection and reduce transmission, with many countries rolling out mass vaccination programs (Castells & Phillips, 2021). Although vaccines for SARS-CoV-2 now exist, new strains due to mutations have led to further concerns about vaccine effectiveness (Boehm *et al.*, 2021; Jangra *et al.*, 2021). This fact, together with waning immunity and the existence of individuals who are unable to be vaccinated or out-right refuse to do so, highlight the need for additional therapeutics and prophylactics (Fu *et al.*, 2020; Naqvi *et al.*, 2020) in addition to booster programmes (Du *et al.*, 2022).

One such potential therapy is viral interfering particles. During viral replication, mutants, lacking essential parts of the viral genome, are generated (Alnaji & Brooke, 2020; Fatehi *et al.*, 2021) which are unable to replicate in the absence of wild type (WT) virus. These are known as defective interfering particles (DIPs). DIPs can be exploited to make therapeutic interfering particles (TIPs), which inhibit the replication of WT virus by competing with WT gene segments for resources required during viral replication and assembly (Bdeir *et al.*, 2019; Chaturvedi *et al.*, 2021). TIPs/DIPs have been investigated for several viruses,

including HIV, Ebola, influenza and SARS-CoV-2 and have been found to cause a reduction in viral titres (Bdeir *et al.*, 2019; Chaturvedi *et al.*, 2021; Smither *et al.*, 2020). However, caveats exist in their production; for instance, how the DIP genome is formed to allow for faster replication than WT and which sections must be conserved to allow replication. DIPS are also virus-specific, and little is known about how mutations change replication dynamics (Fatehi *et al.*, 2021; Rouzine & Weinberger, 2013).

From a mathematical modelling perspective, a long-standing effort exists to describe transmission dynamics at the population level and within-host level (Grebennikov *et al.*, 2022) in the presence of DIPs. At the within-host level, DIPs as therapeutics have been studied by Perelson & Ke (2021) and Zhao & Xing (2021). However, little effort has been devoted to investigating the intra-cellular replication kinetics of DIPs in the presence of WT virus. Grebennikov *et al.* (2021) have recently proposed a model for SARS-CoV-2 intra-cellular replication dynamics. This model allowed for the quantification of the viral genomes and proteins during the replication cycle. Here, the aim is to extend this model to explore co-infection with DIPs and the effect of DIPs on the replication dynamics of the WT virus on a intracellular level.

In this section, I formulate a mathematical model of SARS-CoV-2 replication in a cell co-infected with DIPs. As in the paper presented by Grebennikov *et al.* (2021), I will follow a deterministic approach to calibrate model parameters. I use the sensitivity analysis described in Section 2.5.1 to study the impact parameters have on the release of both WT and DIP virus particles. I also introduce a stochastic description of this model to compare against the deterministic one. Finally, I will also investigate how initial doses of each virus affect virus particle production (WT and DIPs), to quantify DIP inhibition of WT replication and the reliance of DIPs on the WT replication machinery.

5.1.1 Mathematical model

In what follows, I use a system of ordinary differential equations to describe the biochemical reactions associated with the intra-cellular replication of SARS-CoV-2 in the presence of DIPs. The system of ordinary differential equations (ODEs) is
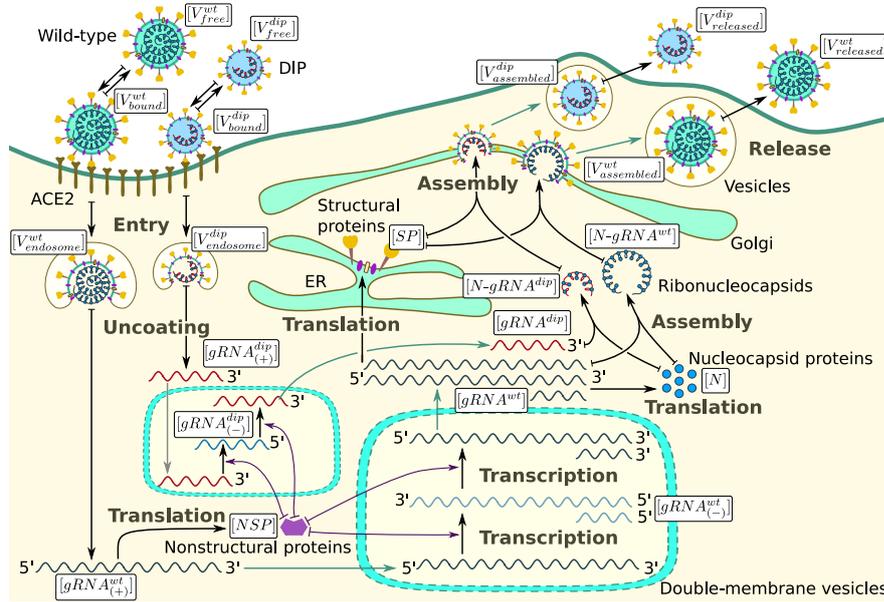
Figure 5.1: Biological scheme of the competitive replication of the infectious SARS-CoV-2 and defective interfering particles. Figure provided by Dmitry Grebennikov.

formulated under the assumption of mass-action kinetics and Michaelis-Menten approximations, based on the biological scheme presented in Figure 5.1. The equations describe cell entry, RNA release, RNA transcription and DIP parasitism of WT replication machinery. This is followed by translating new RNA molecules and competition for nucleocapsid and structural proteins. Finally, new WT virions and DIPs are assembled and released from the cell. Later I will describe the model as a stochastic process.

Cell entry and RNA release

Binding of the free infectious WT virion to the cellular trans-membrane protein ACE2 via the receptor binding domain of the spike protein allows entry and release of the viral RNA into the host cell. Subsequently, the virus will then bind to the cellular or endosomal membrane, which causes un-coating and release of viral genomic RNA. These processes are described by equations specifying the rates of change of free, receptor-bound, and fused virions, as well as the viral RNA

genome in the cytoplasm:

$$\frac{d[V^{wt}_{free}]}{dt} = -k_{bind}[V^{wt}_{free}] - d^{wt}_V[V^{wt}_{free}] + k_{diss}[V^{wt}_{bound}], \tag{5.1}$$

$$\frac{d[V^{wt}_{bound}]}{dt} = k_{bind}[V^{wt}_{free}] - (k_{fuse} + k_{diss} + d^{wt}_V)[V^{wt}_{bound}], \tag{5.2}$$

$$\frac{d[V^{wt}_{endosome}]}{dt} = k_{fuse}[V^{wt}_{bound}] - (k_{uncoat} + d^{wt}_{endosome})[V^{wt}_{endosome}], \tag{5.3}$$

$$\frac{d[gRNA^{wt}_{(+)}]}{dt} = k_{uncoat}[V^{wt}_{endosome}] - d^{wt}_{gRNA}[gRNA^{wt}_{(+)}]. \tag{5.4}$$

Here $[V^{wt}_{free}]$ is the number of extra-cellular free infectious virions, $[V^{wt}_{bound}]$ the number of virions bound to ACE2 and activated by transmembrane serine protease 2 (TMPRSS2), $[V^{wt}_{endosome}]$ is the number of virions in endosomes, and $[gRNA^{wt}_{(+)}]$ the number of single-stranded (ss)-positive sense genomic RNA. Free WT virus binds to ACE2 receptors with rate k_{bind} and can be cleared with rate d^{wt}_V. Bound WT virus can then fuse to the cellular or endosome membrane with rate k_{fuse} or disassociate from the receptor with rate k_{diss}. The clearance rate for bound WT virus remains as previously described for free WT virus since there have been no major structural changes to the virion. Upon fusion and forming of an endosome, the viral RNA can be un-coated with rate k_{uncoat}. Virions in the endosome can degrade with rate $d^{wt}_{endosome}$. Finally, the newly un-coated genomic RNA can be degraded with rate d^{wt}_{gRNA}. A similar set of equations is used to describe the cell entry and RNA release of non-infectious defective interfering particles:

$$\frac{d[V^{dip}_{free}]}{dt} = -k_{bind}[V^{dip}_{free}] - d^{dip}_V[V^{dip}_{free}] + k_{diss}[V^{dip}_{bound}], \tag{5.5}$$

$$\frac{d[V^{dip}_{bound}]}{dt} = k_{bind}[V^{dip}_{free}] - (k_{fuse} + k_{diss} + d^{dip}_V)[V^{dip}_{bound}], \tag{5.6}$$

$$\frac{d[V^{dip}_{endosome}]}{dt} = k_{fuse}[V^{dip}_{bound}] - (k_{uncoat} + d^{dip}_{endosome})[V^{dip}_{endosome}], \tag{5.7}$$

$$\frac{d[gRNA^{dip}_{(+)}]}{dt} = k_{uncoat}[V^{dip}_{endosome}] - d^{dip}_{gRNA}[gRNA^{dip}_{(+)}]. \tag{5.8}$$

Here $[V^{dip}_{free}]$ is the number of extra-cellular free DIPs, $[V^{dip}_{bound}]$ the number of DIPs bound to ACE2 and activated by TMPRSS2, $[V^{dip}_{endosome}]$ the number of DIPs in endosomes, and $[gRNA^{dip}_{(+)}]$ the number of ss-positive sense genomic RNA. DIPs

for SARS-CoV-2 would require a functional spike (S) protein to bind to ACE2 receptors and mediate cell entry successfully. Consequently, it was assumed that the rates for k_{bind}, k_{diss}, k_{fuse}, and k_{uncoat} are the same for both WT virus and DIPs. However, degradation rates related to cell entry will differ between WT and DIPs since the shorter genome of DIPs might imply a different degradation rate.

RNA transcription and DIP parasitism

The released WT viral genomic RNA undergoes translation into non-structural viral polyproteins, $[NSP]$, with the rate k_{transl} according to the length of open reading frame 1 (ORF1) f_{ORF1}, which operate to form the viral replication and transcription complex, , *i.e.,* the RNA-dependent RNA polymerase (RdRp). These NSPs can become degraded with rate d_{NSP}. The main function of the RdRp replication complex is to generate a negative-sense full-length genome and subgenomic RNAs. As DIPs lack the ability of self-replication, the conditional transcription of DIP RNAs results in competition with WT SARS-CoV-2 for replication proteins (Chaturvedi *et al.*, 2021). Using WT virus *trans* elements by DIPs reduces $[NSP]$ availability for transcribing WT viral RNA. This is encoded in Equation (5.9). From positive sense genomic RNA ($[gRNA^{wt}_{(+)}]$) negative genomic RNA strands ($[gRNA^{wt}_{(-)}]$) are generated proportionally to RdRp with rate $k^{wt}_{tr_{(-)}}$. These negative strains can degrade with rate $d^{wt}_{gRNA_{(-)}}$. The subsequent formation of negative RNA strands allows the formation of positive sense genomic and subgenomic strands required to form new virions. These are denoted by $[gRNA^{wt}]$, and are produced with rate $k^{wt}_{tr_{(+)}}$. New positive strands form ribonucleocapsids are recruited with the rate $k^{wt}_{complex}$ proportional to the maximal number of N proteins available. These strands have a degradation rate d^{wt}_{gRNA}. The abundance of non-structural proteins, $[NSP]$, the negative sense genomic and subgenomic RNAs, $[gRNA^{wt}_{(-)}]$, and positive sense genomic and subgenomic RNAs, $[gRNA^{wt}]$, associated with the replication of the infectious virions is described by the following

equations:

$$\frac{d[NSP]}{dt} = k_{transl}f_{ORF1}[gRNA^{wt}_{(+)}] - d_{NSP}[NSP]$$
$$- (k^{wt}_{trans_{(-)}}[gRNA^{wt}_{(+)}] + k^{wt}_{trans_{(+)}}[gRNA^{wt}_{(-)}]+ \qquad (5.9)$$
$$k^{dip}_{trans_{(-)}}[gRNA^{dip}_{(+)}] + k^{dip}_{trans_{(+)}}[gRNA^{dip}_{(-)}])[NSP],$$

$$\frac{d[gRNA^{wt}_{(-)}]}{dt} = k^{wt}_{tr_{(-)}}[gRNA^{wt}_{(+)}]\theta_{RdRp} - d^{wt}_{gRNA_{(-)}}[gRNA^{wt}_{(-)}], \qquad (5.10)$$

$$\frac{d[gRNA^{wt}]}{dt} = k^{wt}_{tr_{(+)}}[gRNA^{wt}_{(-)}]\theta_{RdRp} - (k^{wt}_{complex}\theta_{complex} + d^{wt}_{gRNA})[gRNA^{wt}], \qquad (5.11)$$

where

$$\theta_{RdRp} = \frac{[NSP]}{[NSP] + K_{NSP}}, \quad \theta_{complex} = \frac{[N]}{[N] + K_N}. \qquad (5.12)$$

Equation (5.9) reflects the fact that non-structural proteins are translated only from the viral genomic RNA of infectious WT virions. Transcription of negative-sense viral genomic and subgenomic RNAs described by Equations (5.10) and (5.11) is regulated by the positive-sense viral genomic RNA. The set of equations for transcription of negative sense and positive sense DIP subgenomic RNAs, *,i.e.,* $[gRNA^{dip}_{(-)}]$, $[gRNA^{dip}]$, are as follows:

$$\frac{d[gRNA^{dip}_{(-)}]}{dt} = k^{dip}_{tr_{(-)}}[gRNA^{dip}_{(+)}]\theta_{RdRp} - d^{dip}_{gRNA_{(-)}}[gRNA^{dip}_{(-)}], \qquad (5.13)$$

$$\frac{d[gRNA^{dip}]}{dt} = k^{dip}_{tr_{(+)}}[gRNA^{dip}_{(-)}]\theta_{RdRp} - (k^{dip}_{complex}\theta_{complex} + d^{dip}_{gRNA})[gRNA^{dip}]. \qquad (5.14)$$

The rates for $k^{dip}_{tr_{(-)}}$ and $k^{dip}_{tr_{(+)}}$ could all be assumed to be faster than those of WT due to the shorter length of DIP genomes. The degradation rates would depend on the overall stability of DIP genomes for positive and negative sense strands.

Translation and competition for nucleocapsid protein and other structural proteins

DIPs compete with WT virions for packaging proteins, *e.g.,* nucleocapsid N proteins ([N]) (Chaturvedi *et al.*, 2021). Structural S, envelope E, and membrane M

proteins are translated from positive sense subgenomic RNAs in the endoplasmic reticulum (ER) and are considered in the mathematical model as a single population, $[SP]$. Nucleocapsid proteins, on the other hand, are translated in cytosolic ribosomes from positive-sense RNAs. SP and N proteins are required to form virus-like particles, WT or DIPs. N proteins are translated with the rate k_{transl} over the length of the genome f_N from positive sense WT virus $gRNA^{wt}$. Similarly, SP are translated with the same rate over genome length f_{SP}. N proteins are lost to ribonucleocapsid ($N - gRNA^{wt}$) formation with rate $k_{complex}^{wt}$ requiring n_N^{wt} N proteins. Furthermore, DIPs undergo the same process, requiring n_N^{dip} N proteins. SP is only required for virion assembly with rate k_{assemb}^{wt} for WT virions requiring n_{SP}^{wt} structural proteins. n_{SP}^{dip} structural proteins are required for DIPs with rate k_{assemb}^{dip}. It can be assumed that $n_{SP}^{dip} \leq n_{SP}^{wt}$ and $n_N^{dip} \leq n_N^{wt}$ since the shorter DIP genome will require fewer N proteins for formation of the ribonucleocapsid ($N - gRNA^{dip}$) and construction of a complete DIP. N proteins and SP structural proteins degrade with rates d_N and d_{SP}, respectively. The set of equations to describe these dynamics are as follows:

$$\frac{d[N]}{dt} = k_{transl}f_N[gRNA^{wt}] - k_{complex}^{wt}n_N^{wt}\theta_{complex}[gRNA^{wt}]$$
$$- k_{complex}^{dip}n_N^{dip}\theta_{complex}[gRNA^{dip}] - d_N[N], \tag{5.15}$$

$$\frac{d[SP]}{dt} = k_{transl}f_{SP}[gRNA^{wt}] - k_{assemb}^{wt}n_{SP}^{wt}\theta_{assemb}^{wt}[N\text{-}gRNA^{wt}]$$
$$- k_{assemb}^{dip}n_{SP}^{dip}\theta_{assemb}^{dip}[N\text{-}gRNA^{dip}] - d_{SP}[SP], \tag{5.16}$$

where

$$\theta_{assemb}^{wt} = \frac{[SP]}{[SP] + K_{V_{rel}}^{wt}n_{SP}^{wt}}, \tag{5.17}$$

and

$$\theta_{assemb}^{dip} = \frac{[SP]}{[SP] + K_{V_{rel}}^{dip}n_{SP}^{dip}}, \tag{5.18}$$

Assembly and release of WT SARS-CoV-2 and DIPs

New virions are assembled in the endoplasmic reticulum-Golgi compartment, where N-RNA complexes become encapsulated. These assembled virions can then exit an infected cell by exocytosis via a lysosomal pathway, budding, or cell

death (Mendonça *et al.*, 2021; V'kovski *et al.*, 2021). No competition is associated with releasing new infectious virions and DIPs. Still, the rates θ_{assemb}^{wt} and θ_{assemb}^{dip} associated with viral assembly will be regulated by the availability of structural proteins, since DIPs will likely require fewer than WT virions. $d_{N\text{-}gRNA^{wt}}$ describes the degradation rate of WT ribonucleocaspids, $d_{assembled}^{wt}$ the degradation rate of newly assembled WT virions. Newly assembled WT virions are released with rate $k_{release}^{wt}$, and released WT virions can be cleared with the rate d_V^{wt}. A similar set of rates are defined for DIPS. The rates of change for the ribonucleocapsid $[N\text{-}gRNA]$, assembled $[V_{assembled}]$ and released $[V_{released}]$ infectious SARS-CoV-2 and DIPs are described below:

$$\frac{d[N\text{-}gRNA^{wt}]}{dt} = k_{complex}^{wt}\theta_{complex}[gRNA^{wt}]$$
$$- (k_{assemb}^{wt}\theta_{assemb}^{wt} + d_{N\text{-}gRNA}^{wt})[N\text{-}gRNA^{wt}], \qquad (5.19)$$

$$\frac{d[V_{assembled}^{wt}]}{dt} = k_{assemb}^{wt}\theta_{assemb}^{wt}[N\text{-}gRNA^{wt}]$$
$$- (k_{release}^{wt} + d_{assembled}^{wt})[V_{assembled}^{wt}], \qquad (5.20)$$

$$\frac{d[V_{released}^{wt}]}{dt} = k_{release}^{wt}[V_{assembled}^{wt}] - d_V^{wt}[V_{released}^{wt}], \qquad (5.21)$$

and

$$\frac{d[N\text{-}gRNA^{dip}]}{dt} = k_{complex}^{dip}\theta_{complex}[gRNA^{dip}]$$
$$- (k_{assemb}^{dip}\theta_{assemb}^{dip} + d_{N\text{-}gRNA}^{dip})[N\text{-}gRNA^{dip}], \qquad (5.22)$$

$$\frac{d[V_{assembled}^{dip}]}{dt} = k_{assemb}^{dip}\theta_{assemb}^{dip}[N\text{-}gRNA^{dip}]$$
$$- (k_{release}^{dip} + d_{assembled}^{dip})[V_{assembled}^{dip}], \qquad (5.23)$$

$$\frac{d[V_{released}^{dip}]}{dt} = k_{release}^{dip}[V_{assembled}^{dip}] - d_V^{dip}[V_{released}^{dip}]. \qquad (5.24)$$

In this study, I aim to analyse the model behaviour for different initial conditions, $[V_{free}^{wt}](0)$ and $[V_{free}^{dip}](0)$, and thus, understand the replication dynamics of WT virus particles in the presence of DIPs, and how the initial dose of WT and DIP particles regulates infection and production kinetics of WT virions. Here I study the amount of WT virus and DIPs released over time.

Variable	Definition
$[V^{wt}_{free}]$	number of free infectious (*i.e.,* wild-type) virions outside the cell membrane
$[V^{wt}_{bound}]$	number of infectious virions bound to ACE2 and activated by TMPRSS2
$[V^{wt}_{endosome}]$	number of infectious virions in endosomes
$[V^{dip}_{free}]$	number of free DIPs
$[V^{dip}_{bound}]$	number of DIPS bound to ACE2 and activated by TMPRSS2
$[V^{dip}_{endosome}]$	number of non-infectious virions in endosomes
$[gRNA^{wt}_{(+)}]$	single strand positive sense genomic RNA
$[gRNA^{dip}_{(+)}]$	single strand positive sense DIP genomic RNA
$[NSP]$	non-structural proteins
$[gRNA^{wt}_{(-)}]$	negative sense genomic and subgenomic RNAs of infectious virus
$[gRNA^{wt}]$	positive sense genomic and subgenomic RNAs of infectious virus
$[gRNA^{dip}_{(-)}]$	negative sense subgenomic RNAs of DIPs
$[gRNA^{dip}]$	positive sense genomic and subgenomic RNAs of DIPs
$[SP]$	total number of structural proteins $S + M + E$ per virion
$[N]$	N proteins per virion
$[N\text{-}gRNA^{wt}]$	ribonucleocapsid molecules for infectious virions
$[N\text{-}gRNA^{dip}]$	ribonucleocapsid molecules for DIPs
$[V^{wt}_{assembled}]$	assembled infectious virions in endosomes
$[V^{wt}_{released}]$	released infectious viruses
$[V^{dip}_{assembled}]$	assembled DIPs in endosomes
$[V^{dip}_{released}]$	released DIPs

Table 5.1: Dynamical variables of the mathematical model for the life cycle of SARS-CoV-2, with defective interfering particles.

Parameter	Description, Units	Value	Range
k_{bind}	rate of virion binding to ACE2 receptor, h^{-1}	12	(3.6, 12)
d_V^{wt}	clearance rate of WT extracellular virions, h^{-1}	0.12	(0.06, 3.5)
k_{diss}	dissociation rate constant of bound virions, h^{-1}	0.61	(0.32, 1.08)
k_{fuse}	fusion rate constant, h^{-1}	0.5	(0.33, 1)
k_{uncoat}	un-coating rate constant, h^{-1}	0.5	(0.33, 1)
$d_{endosome}^{wt}$	degradation rate of WT virions in endosomes h^{-1}	0.06	(0.0001, 0.12) ,
k_{transl}	translation rate, nt/mRNA h^{-1}	45,360	(40,000, 50,000)
$1/f_{ORF1}$	length of ORF1 of the RNA genome coding [NSP], nt	21,000	fixed
d_{NSP}	degradation rate of proteins in the cell, h^{-1}	0.069	(0.023, 0.69), tuned to (0.023, 0.1)
$k_{tr(-)}^{wt}$	transcription rate of WT negative sense genomic and subgenomic RNAs, copies/mRNA h^{-1}	3	(1, 100), tuned to (1, 20)
K_{NSP}	threshold number of [NSP] enhancing vRNA transcription, molecules	100	(10, 150)
d_{gRNA}^{wt}	degradation rate of WT positive sense RNAs in cell, h^{-1}	0.2	(0.069, 0.69), tuned to (0.069, 0.4)

Table 5.2: Estimates of the calibrated model parameters, for the model presented in Equations. (5.1)-(5.24), for the variables defined in Table 5.1, as reported by Grebennikov et al. (2021)

Parameter	Description, Units	Value	Range
$d_{gRNA(-)}^{ut}$	degradation rate of WT negative sense RNAs in double-membrane vesicles, h^{-1}	0.1	(0.05, 0.2)
$k_{tr(+)}^{ut}$	replication rate of positive sense WT RNAs, copies/mRNA/h	1000	(620, 1380)
$k_{complex}^{ut}$	rate of the WT nucleocapsid formation [$N - gRNA^{ut}$], h^{-1}	0.4	(0.02, 0.4)
K_N	threshold number of N proteins at which nucleocapsid formation slows down, molecules	5×10^6	$(3.5, 6.5) \times 10^6$
$1/f_N$	length of RNA genome coding N protein, nt	1200	fixed
$1/f_{SP}$	length of genome coding structural proteins S, E, M, nt	10^4	fixed
d_N	degradation rate of N protein, h^{-1}	0.023	(0.023, 0.069)
d_{SP}	mean degradation rate of the pool of E, S, M proteins, h^{-1}	0.044	(0.023, 0.36)
n_{SP}^{ut}	total number of structural proteins S, M, E per WT virion, molecules	2×10^3	(1125, 2230)
n_N^{ut}	number of N proteins per WT virion, molecules	456	fixed
$K_{V_{rel}}^{ut}$	threshold number of WT virions at which the virion assembly process slows down, virions	10^3	(10, 10^4)
k_{assemb}^{ut}	rate of WT virion assembling, h^{-1}	1	(0.01, 10)
d_{N-gRNA}^{ut}	degradation rate of WT ribonucleoprotein, h^{-1}	0.2	(0.069, 0.69)
$k_{release}^{ut}$	rate of WT virion release via exocytosis, h^{-1}	8	(8, 7,200)
$d_{assembled}^{ut}$	assembled WT virion degradation rate, h^{-1}	0.06	$(10^{-4}, 0.12)$

Table 5.3: Estimates of the calibrated model parameters continued. References for each parameter can be found in the appendices (Appendix A.3)

Parameter	Description, Units	Value	Range
d_V^{dip}	clearance rate of extracellular DIPs, h^{-1}	0.481	$10^{[-1.2,0.55]}$
$d_{endosome}^{dip}$	degradation rate of DIPs in endosomes, h^{-1}	3.29×10^{-3}	$10^{[-4,-0.93]}$,
$k_{tr_{(-)}}^{dip}$	transcription rate of DIP negative sense genomic and subgenomic RNAs, copies/mRNA, h^{-1}	34	$10^{[0,3]}$,
d_{gRNA}^{dip}	degradation rate of DIP positive sense RNAs in cell, h^{-1}	0.218	$10^{[-1.16,-0.16]}$,
$d_{gRNA_{(-)}}^{dip}$	degradation rate of DIP negative sense RNAs in double-membrane vesicles, h^{-1}	0.218	$10^{[-1.30,0]}$
$k_{tr_{(+)}}^{dip}$	replication rate of positive sense DIP RNAs, copies/mRNA/h	2,540	$10^{[2.79,4.14]}$
$k_{complex}^{dip}$	rate of the DIP nucleocapsid formation $[N - gRNA^{dip}]$, h^{-1}	0.14	$10^{[-1.69,0]}$
n_{SP}^{dip}	total number of structural proteins per DIP	112	$10^{[1,3.1]}$
n_N^{dip}	number of N proteins per DIP, molecules	53	$10^{[1.2,3.5]}$
$K_{V_{rel}}^{dip}$	threshold number of DIPs,	380	$10^{[1,4.31]}$
k_{assemb}^{dip}	rate of DIP assembling, h^{-1}	0.38	$10^{[-2,1.31]}$
d_{N-gRNA}^{dip}	degradation rate of DIP ribonucleoprotein, h^{-1}	0.268	$10^{[-1.16,0]}$
$k_{release}^{dip}$	rate of DIP release via exocytosis, h^{-1}	105	$10^{[0.9,3.15]}$
$d_{assembled}^{dip}$	assembled DIP degradation rate, h^{-1}	4.89×10^{-3}	$10^{[-4,-0.62]}$
$k_{trans_{(-)}}^{wt}$	rate of loss of NSPs by trans elements from negative sense WT RNA, h^{-1}	5.39×10^{-5}	$10^{[-5,-3.7]}$
$k_{trans_{(+)}}^{wt}$	rate of loss of NSPs by trans elements from positive sense WT RNA , h^{-1}	6.17×10^{-3}	$10^{[-2.22,-1.86]}$
$k_{trans_{(-)}}^{dip}$	rate of loss of NSPs by trans elements from negative sense DIP RNA, h^{-1}	4.72×10^{-5}	$10^{[-5.69,-3]}$
$k_{trans_{(+)}}^{dip}$	rate of loss of NSPs by trans elements from positive sense DIP RNA , h^{-1}	8.61×10^{-3}	$10^{[-2.9,-1.17]}$

Table 5.4: Median estimates of unknown model parameters values for the model presented in Equations. (5.1)-(5.24), with variables defined in Table 5.1, as well as relevant search ranges. References for each parameter can be found in the appendices (Appendix A.3).

Sensitivity analysis

To evaluate the model response to variation in parameter values, Sobol global sensitivity analysis was performed on four different model outputs using the method described in Section 2.5.1. I first considered the variability of WT genomic RNA, $[gRNA^{wt}]$, and DIP genomic RNA, $[gRNA^{dip}]$, as a result of changing parameter values within a set range summarised in Tables 5.2, 5.3 and 5.4. Secondly, I investigated how parameter variability affects the release kinetics of both WT $[V_{released}^{wt}]$ and DIP $[V_{released}^{dip}]$ particles by 48 hours post-infection. Understanding which parameters cause the most variability in this model would allow model calibration with careful consideration to minimise output variation.

Figure 5.2 illustrates the first and total-order sensitivities for WT genomic RNA, $[gRNA^{wt}]$, and DIP genomic RNA, $[gRNA^{dip}]$, as an output of the proposed model. For $[gRNA^{dip}]$ the parameter $k_{tr_{(-)}}^{dip}$ was identified to contribute the largest variation. $k_{tr_{(-)}}^{dip}$ is associated with the transcription of negative sense RNAs for DIPs, which is vital in the formation of new positive sense genomic and subgenomic RNAs. The rate $k_{tr_{(+)}}^{dip}$ was also identified as a parameter sensitive to change, since it is associated with the transcription of positive-sense RNAs. Consequently, $k_{tr_{(-)}}^{wt}$ was the second most important parameter in minimising variation in model output for $[gRNA^{wt}]$ (Figure 5.2), following the same reasoning as for DIP positive sense genomic RNA.

A parameter that was of great importance, and not only causes a large variation in model outputs of $[gRNA]$ for WT or DIPs, but also $[V_{released}^{\cdot}]$, was the threshold parameter of non-structural proteins, K_{NSP}. K_{NSP} causes the most variation for $[V_{released}^{dip}]$ and $[gRNA^{wt}]$ compared to any other parameter, and for $[V_{released}^{wt}]$ (Figure 5.3) and $[gRNA^{dip}]$ (Figure 5.2) it is the second most important parameter. K_{NSP} is associated with the transcription of both negative and positive sense genomic RNAs, and changes in the value of this parameter will modify the number of WT virions and DIPs released. $k_{tr_{(-)}}^{wt}$ was identified as an important parameter to minimise variation in the release of both WT $[V_{released}^{wt}]$ and DIPs $[V_{released}^{dip}]$. Consequently, transcription of negative sense WT genomic RNAs is vital in producing positive stranded gRNA that is then translated to form structural proteins S, M and E $[SP]$ and nucleocapsid proteins $[N]$ that are required to
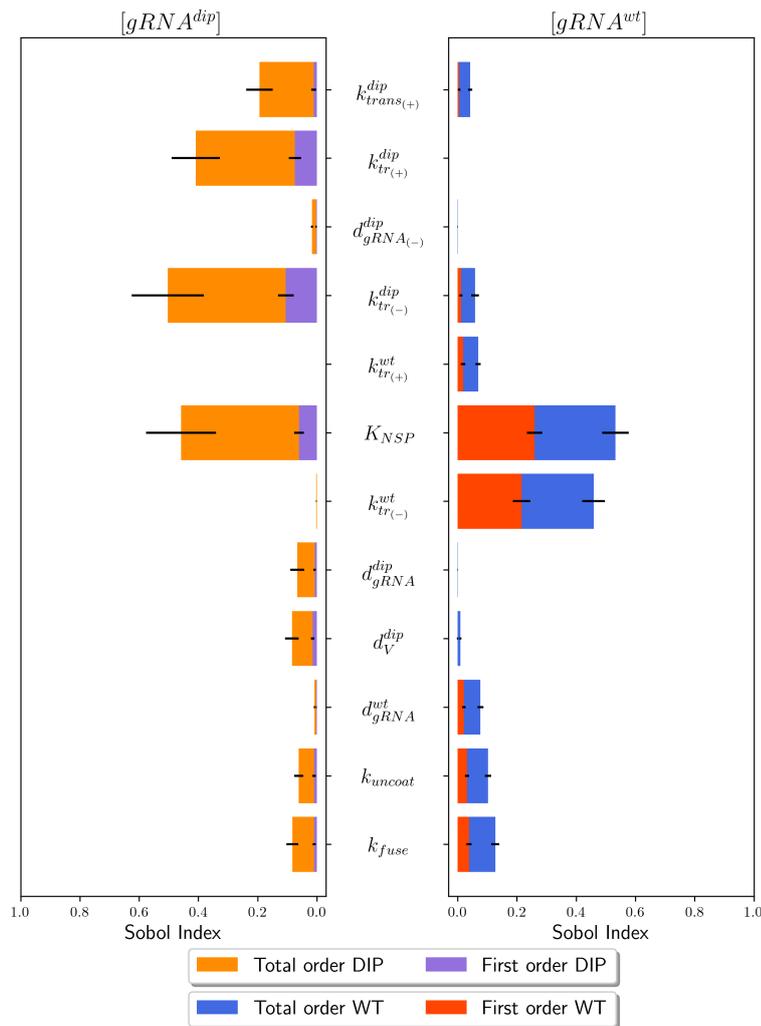
Figure 5.2: First and total-order sensitivities from 10^4 samples with 95% confidence interval (black line). (**Left:**) sensitivities for the variable $[gRNA^{dip}]$. (**Right:**) sensitivities for the variable $[gRNA^{wt}]$.

Table 5.5: Fold-log reductions for 24 and 48 hours post-infection as reported in Ref. Chaturvedi *et al.* (2021) for TIP2.

Time (hours)	Fold log reduction (2 d.p.)
24	1.20
48	1.14

form new particles. Parameters associated with WT virion or DIP assembly are also important to monitor to reduce variation in model outputs. Several parameters identified by Sobol sensitivity analysis have been previously estimated by Grebennikov *et al.* (2021) and are summarised in Table 5.2. Table 5.4 lists other parameters that require estimation.

Parameter calibration

Several parameters were introduced in this extension of the model proposed by Grebennikov *et al.* (2021), which have not been previously quantified. To estimate their values, I performed Bayesian parameter calibration. Since experimental data sets on co-infection with DIPs are limited, I aimed to achieve the fold reduction experimentally quantified by Chaturvedi *et al.* (2021). Chaturvedi *et al.* (2021) investigated two SARS-CoV-2 DIPs as TIPs. Both DIPs had shorter genomes, around 6%-10%, than the WT virus. Chaturvedi *et al.* performed a virus yield-reduction assay by transfecting Vero cells with TIP or control RNAs (one μg/million cells) 24 hours before infection with SARS-CoV-2 at a MOI=0.05, and harvesting culture supernatants for titration at various time-points (24, 48, or 72 hours post-infection). More details on the manufacture of these DIPs can be found in Chaturvedi *et al.* (2021). They discovered that these particles lead to a $1.5 - 1.2 \log$ fold reduction in virus produced compared to control samples and highlighted the potential of DIPs as therapeutic. I compared the fold reduction generated by therapeutic interfering particle two (TIP2) (Chaturvedi *et al.*, 2021) for 24 and 48 hours to the fold reduction from the mathematical model of $[V_{released}^{wt}]$ against the original model proposed by Grebennikov *et al.* These fold reductions are summarised in Table 5.5. The ABC rejection algorithm as presented in Section 2.4.1 was used with 10^6 sample sets. Since a choice of ε is hard to determine,
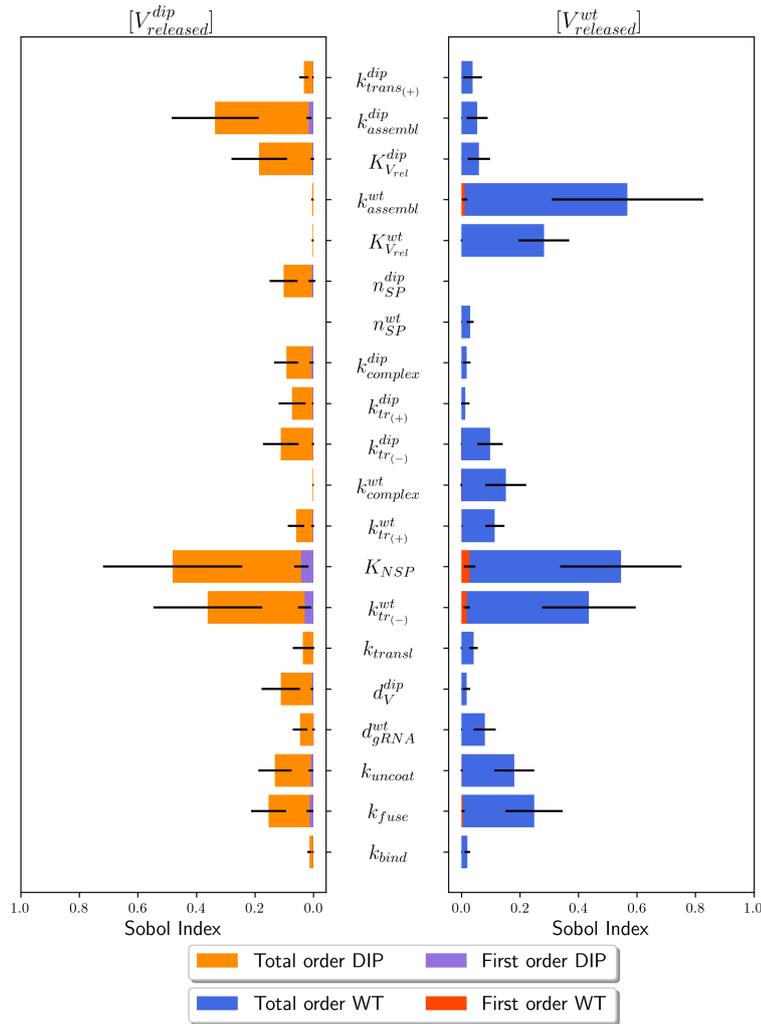
Figure 5.3: First and total-order sensitivities from 10^4 samples with 95% confidence interval (black line). **(Left:)** sensitivities for the variable $[V_{released}^{dip}]$. **(Right:)** sensitivities for the variable $[V_{released}^{wt}]$.
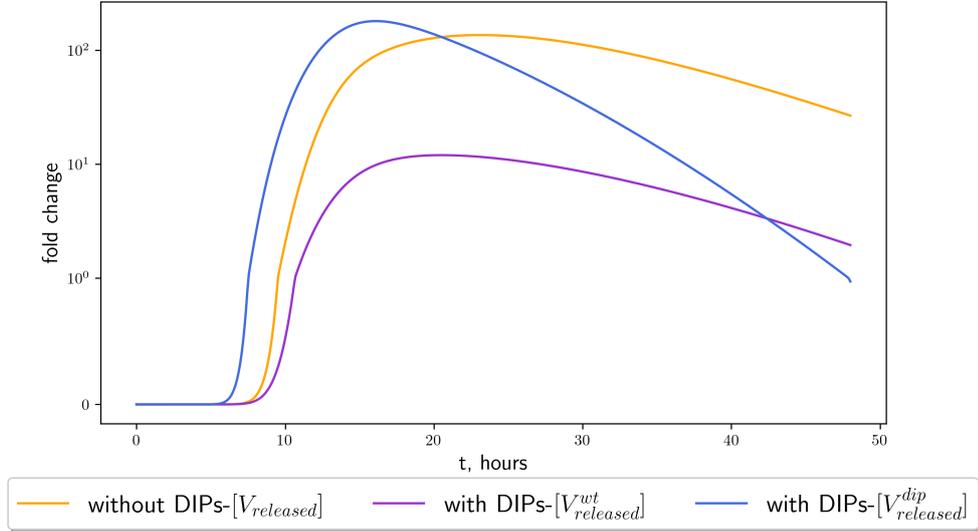
Figure 5.4: Virus particle release kinetics predicted by the model with initial conditions $[V_{free}^{wt}](0) = 10$ and $[V_{free}^{dip}](0) = 10$ within 48 hours after model calibration using ABC rejection and data from Chaturvedi *et al.* (2021). Median parameter values summarised in Table 5.4 were used for previously unknown parameter values. **(Yellow line:)** shows the reference solution to a model where DIPs are not considered in the replication dynamics. **(Purple line:)** illustrates the production of WT virions $[V_{released}^{wt}]$ with DIPs **(blue line)** $[V_{released}^{dip}]$.

I took the 0.1% of parameter sets instead, minimising the Euclidean distance. Samples were taken in the exponent of the search ranges stated in Table 5.4. As a result, this sample size would provide a large coverage of parameter space. Figure 5.4 illustrates the model output where median parameter values are used from the accepted 0.1% sample sets. From these median values, I obtained a fold change of 1.08 (2 d.p.) at 24 hours post-infection and 1.14 (2 d.p.) at 48 hours post-infection, compared to the reference solution originally determined by Grebennikov *et al.* (2021). Posterior histograms in Figure 5.5 showed that with the data set and the mathematical model, Bayesian inference has led to poor learning for all but one newly introduced parameter. Posterior distributions are still wide, with $k_{trans_{(+)}}^{wt}$ being the only one with narrow posteriors hence being structurally identifiable. This was due to a lack of longitudinal data to compare modelled DIP replication dynamics. Improving the parameter estimates would

require additional data that, is not currently available. The question remains how additional experimental outputs would affect the structural identifiability of the parameters. To investigate this, one can use the SIAN toolbox and input additional observed outputs to see how identifiability changes.

The following results are generated by using SIAN. Suppose that observed experimental data exist for not only $[V_{released}^{dip}]$ and $[V_{released}^{wt}]$, but also $[V_{endosome}^{dip}]$ and $[V_{assembled}^{dip}]$ assessing the structural identifiability now allows one to locally identify fifteen of the newly introduced parameters, only three of the parameters are unidentifiable these being $k_{tr(-)}^{dip}$, $k_{tr(+)}^{dip}$ and $k_{trans(+)}^{dip}$. This is also the outcome if experimental data quantified the structural proteins $[SP]$ and nucleocapsid proteins $[N]$. Local identifiability implies that several sets of values work to represent the data. However, it would be beneficial to know how much experimental data would be required to obtain some or all parameters as globally identifiable. If it were possible to experimentally quantify $[V_{released}^{dip}]$, $[V_{released}^{wt}]$, $[V_{endosome}^{dip}]$, $[V_{assembled}^{dip}]$, $[SP]$, $[N]$ and $[NSP]$ and use all these experimental outputs to calibrate this model only three parameters remain unidentifiable. In contrast, all other parameters become globally identifiable with unique solutions. $k_{tr(-)}^{dip}$, $k_{tr(+)}^{dip}$ and $k_{trans(+)}^{dip}$ all remain unidentifiable and would still have wide posteriors after calibration. To make every parameter globally identifiable, one would also need to experimentally quantify $[gRNA_{(-)}^{dip}]$ for negative sense DIP genomic RNA. In total, there would need to be eight total experimental outputs to make this model globally identifiable on the 18 newly introduced parameters giving a unique solution for each parameter. Figure 5.6 illustrates the time evolution for each variable in Table 5.1 given the median values found via ABC rejection. From the upper panels of Figure 5.6, it can be seen that the entry kinetics of the WT virus into the cell are similar to those of the reference solution. DIPs, however, enter the cell at a faster rate than WT virions. It is important to remember that it was assumed there are sufficient ACE2 receptors, and thus, there is no competition between WT and DIP for receptor binding. The number of non-structural proteins is greatly reduced (Figure 5.6 middle left panel), peaking at 7 hours with ≈ 20 molecules as opposed to the reference solution, which peaks at roughly 13 hours with ≈ 40 molecules. The production of $[gRNA_{(-)}^{wt}]$ halves and peaks earlier in the time course, with a greater number of DIP negative sense genomic RNA than WT. Consequently, there was
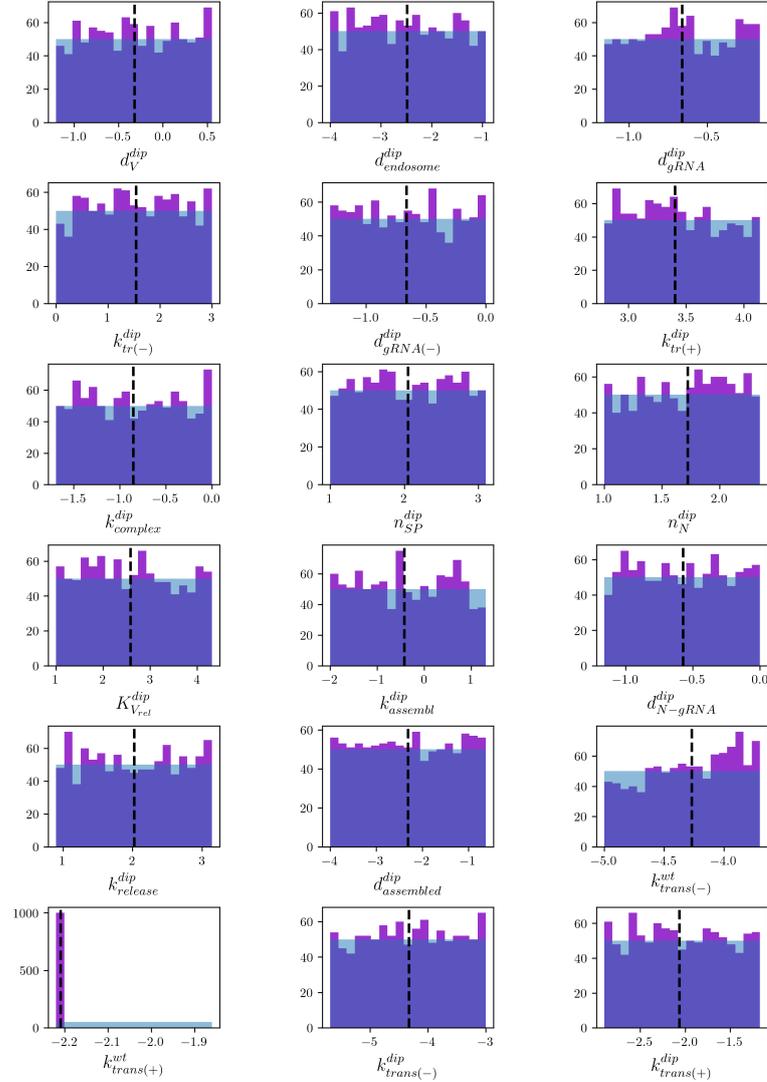
Figure 5.5: Posterior histograms of the top 0.1% of sampled parameter sets as a result of 10^6 samples. Table 5.4 lists the search ranges for the above posterior histograms. **(Purple histogram:)** Posterior histogram of accepted parameter sets, **(blue histogram:)** histogram of prior beliefs and **(black dashed line:)** the median parameter value listed in Table 5.4 that are used to generate Figures 5.4-5.12.

an approximate fold reduction of positive sense genomic RNA, ribonucleocapsid proteins, assembled and released WT virus particles.

Stochastic Markov chain based model

The deterministic model (5.1)-(5.24) can be translated into a stochastic framework in terms of a continuous-time Markov chain (CTMC) $\mathcal{X}$. The stochastic model allows one to obtain the probability distributions rather than the mean field estimates for the variables, and to compute the probabilities of productive cell infection at low MOI. It is convenient to estimate the model parameters for the system of ODEs, and then leverage these estimates in the Markov chain to perform stochastic simulations. The Markov chain corresponding to equations (5.1)-(5.24) is presented in Table 5.6. It includes the state transition events and the propensities q_i for the i-th process. Each population will be considered a random variable such that one can investigate stochastic descriptors. Notice that the processes of ribonucleocapsid formation ($i = 33, 34$) and virion assembly ($i = 37, 38$) are formulated as single events with simultaneous changes of three variables. In these processes, the protein abundances are decreased by the corresponding number of protein molecules n_p needed to form a complex or assemble a pre-virion particle (*i.e.*, by $n_N^{wt}, n_N^{dip}, n_{SP}^{wt}$ or n_{SP}^{dip}). Alternatively, one can formulate the MC with three separate processes for each assembly event, in which the protein molecules are decreased by only one molecule with the propensity multiplied by n_p (see Sazonov *et al.* (2022) for an example of the extended MC formulation). I will use Gillespie simulations described in Section 2.3.1 to evaluate this stochastic model. From these simulations, stochastic descriptors for the expected total number $\mathbb{E}[V_{released}^{(\cdot)}(24)]$ and median total number of WT virions and DIPs released at 24 hours can be examined for varying initial doses of DIPs and WT virus. Furthermore, the probability of a productive infection at 24 hours can also be investigated. The probability of a productive infection is defined as $\mathbb{P}(V_{released}^{wt}(24) > 0)$. These are investigated in the following section.
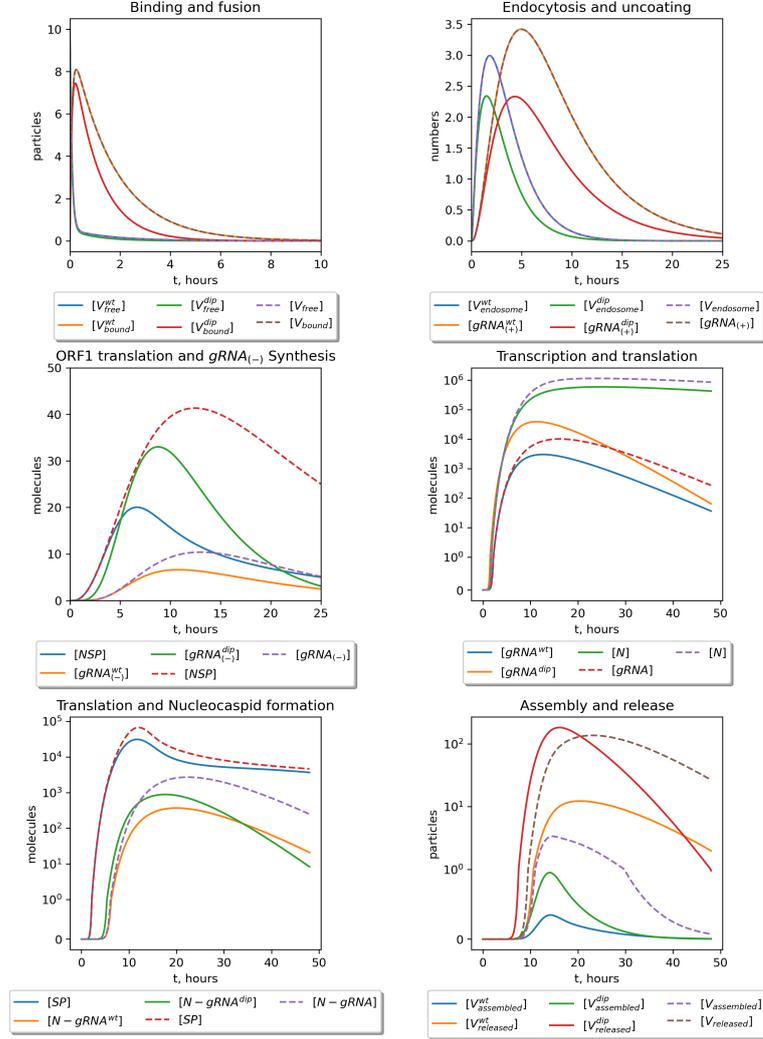
Figure 5.6: Time-dependent state variables of the mathematical model for the life cycle of SARS-CoV2-2 including the wild-type virions and defective interfering particles with initial conditions $[V_{free}^{wt}](0) = 10$ and $[V_{free}^{dip}](0) = 10$ over a 48-hour time course. **(Upper left:)** free WT or DIP virions bind and fuse to the cell ACE2 receptors, **(upper right:)** virions entering endosomes and the un-coating of viral positive sense genomic RNA. **(Middle left:)** transcription and translation to form a negative sense genome and ORF1 to form non structural proteins (NSPs), which is then followed by **(middle right:)** the production of new positive sense genomic RNAs and translation of N protein. **(Bottom left:)** translation of structural proteins and formation of ribonucleocaspid molecules, which lead to **(bottom right:)** the assembly and release of new virions, both WT and DIP. **(Dashed lines:)** represent the reference model solution proposed by Grebennikov *et al.* (2021).

Table 5.6: The Markov chain: individual transitions and their associated rates.

i	Transition	Rate, q_i
	Entry and RNA release (WT):	
1	$[V^{wt}_{free}] \to [V^{wt}_{free}] - 1,\ [V^{wt}_{bound}] \to [V^{wt}_{bound}] + 1$	$k_{bind}[V^{wt}_{free}]$
2	$[V^{wt}_{free}] \to [V^{wt}_{free}] - 1$	$d^{wt}_V[V^{wt}_{free}]$
3	$[V^{wt}_{free}] \to [V^{wt}_{free}] + 1,\ [V^{wt}_{bound}] \to [V^{wt}_{bound}] - 1$	$k_{diss}[V^{wt}_{bound}]$
4	$[V^{wt}_{bound}] \to [V^{wt}_{bound}] - 1,\ [V^{wt}_{endosome}] \to [V^{wt}_{endosome}] + 1$	$k_{fuse}[V^{wt}_{bound}]$
5	$[V^{wt}_{bound}] \to [V^{wt}_{bound}] - 1$	$d^{wt}_V[V^{wt}_{bound}]$
6	$[V^{wt}_{endosome}] \to [V^{wt}_{endosome}] - 1,\ [gRNA^{wt}_{(+)}] \to [gRNA^{wt}_{(+)}] + 1$	$k_{uncoat}[V^{wt}_{endosome}]$
7	$[V^{wt}_{endosome}] \to [V^{wt}_{endosome}] - 1$	$d^{wt}_{endosome}[V^{wt}_{endosome}]$
8	$[gRNA^{wt}_{(+)}] \to [gRNA^{wt}_{(+)}] - 1$	$d^{wt}_{gRNA}[gRNA^{wt}_{(+)}]$
	Entry and RNA release (DIPs):	
9	$[V^{dip}_{free}] \to [V^{dip}_{free}] - 1,\ [V^{dip}_{bound}] \to [V^{dip}_{bound}] + 1$	$k_{bind}[V^{dip}_{free}]$
10	$[V^{dip}_{free}] \to [V^{dip}_{free}] - 1$	$d^{dip}_V[V^{dip}_{free}]$
11	$[V^{dip}_{free}] \to [V^{dip}_{free}] + 1,\ [V^{dip}_{bound}] \to [V^{dip}_{bound}] - 1$	$k_{diss}[V^{dip}_{bound}]$
12	$[V^{dip}_{bound}] \to [V^{dip}_{bound}] - 1,\ [V^{dip}_{endosome}] \to [V^{dip}_{endosome}] + 1$	$k_{fuse}[V^{dip}_{bound}]$
13	$[V^{dip}_{bound}] \to [V^{dip}_{bound}] - 1$	$d^{dip}_V[V^{dip}_{bound}]$
14	$[V^{dip}_{endosome}] \to [V^{dip}_{endosome}] - 1,\ [gRNA^{dip}_{(+)}] \to [gRNA^{dip}_{(+)}] + 1$	$k_{uncoat}[V^{dip}_{endosome}]$
15	$[V^{dip}_{endosome}] \to [V^{dip}_{endosome}] - 1$	$d^{dip}_{endosome}[V^{dip}_{endosome}]$
16	$[gRNA^{dip}_{(+)}] \to [gRNA^{dip}_{(+)}] - 1$	$d^{dip}_{gRNA}[gRNA^{dip}_{(+)}]$
	ORF1 translation and competitive viral RNA replication:	
17	$[NSP] \to [NSP] + 1$	$k_{transl}f_{ORF1}[gRNA^{wt}_{(+)}]$
18	$[NSP] \to [NSP] - 1$	$d_{NSP}[NSP]$
19	$[NSP] \to [NSP] - 1$	$k^{wt}_{trans_{(-)}}[gRNA^{wt}_{(+)}][NSP]$
20	$[NSP] \to [NSP] - 1$	$k^{wt}_{trans_{(+)}}[gRNA^{wt}_{(-)}][NSP]$
21	$[NSP] \to [NSP] - 1$	$k^{dip}_{trans_{(-)}}[gRNA^{dip}_{(+)}][NSP]$
22	$[NSP] \to [NSP] - 1$	$k^{dip}_{trans_{(+)}}[gRNA^{dip}_{(-)}][NSP]$
23	$[gRNA^{wt}_{(-)}] \to [gRNA^{wt}_{(-)}] + 1$	$k^{wt}_{tr_{(-)}}\theta_{RdRp}[gRNA^{wt}_{(+)}]$
24	$[gRNA^{wt}_{(-)}] \to [gRNA^{wt}_{(-)}] - 1$	$d^{wt}_{gRNA_{(-)}}[gRNA^{wt}_{(-)}]$
25	$[gRNA^{wt}] \to [gRNA^{wt}] + 1$	$k^{wt}_{tr_{(+)}}\theta_{RdRp}[gRNA^{wt}_{(-)}]$
26	$[gRNA^{wt}] \to [gRNA^{wt}] - 1$	$d^{wt}_{gRNA}[gRNA^{wt}]$
27	$[gRNA^{dip}_{(-)}] \to [gRNA^{dip}_{(-)}] + 1$	$k^{dip}_{tr_{(-)}}\theta_{RdRp}[gRNA^{dip}_{(+)}]$
28	$[gRNA^{dip}_{(-)}] \to [gRNA^{dip}_{(-)}] - 1$	$d^{dip}_{gRNA_{(-)}}[gRNA^{dip}_{(-)}]$
29	$[gRNA^{dip}] \to [gRNA^{dip}] + 1$	$k^{dip}_{tr_{(+)}}\theta_{RdRp}[gRNA^{dip}_{(-)}]$
30	$[gRNA^{dip}] \to [gRNA^{dip}] - 1$	$d^{dip}_{gRNA}[gRNA^{dip}]$
	Translation and ribonucleocapsid formation:	
31	$[N] \to [N] + 1$	$k_{transl}f_N[gRNA^{wt}]$
32	$[N] \to [N] - 1$	$d_N[N]$
33	$[gRNA^{wt}] \to [gRNA^{wt}] - 1,\ [N] \to [N] - n^{wt}_N,$ $[N\text{-}gRNA^{wt}] \to [N\text{-}gRNA^{wt}] + 1$	$k^{wt}_{complex}\theta_{complex}[gRNA^{wt}]$
34	$[gRNA^{dip}] \to [gRNA^{dip}] - 1,\ [N] \to [N] - n^{dip}_N,$ $[N\text{-}gRNA^{dip}] \to [N\text{-}gRNA^{dip}] + 1$	$k^{dip}_{complex}\theta_{complex}[gRNA^{dip}]$
35	$[SP] \to [SP] + 1$	$k_{transl}f_{SP}[gRNA^{wt}]$
36	$[SP] \to [SP] - 1$	$d_{SP}[SP]$
	Assembly and release:	
37	$[N\text{-}gRNA^{wt}] \to [N\text{-}gRNA^{wt}] - 1,\ [SP] \to [SP] - n^{wt}_{SP},$ $[V^{wt}_{assembled}] \to [V^{wt}_{assembled}] + 1$	$k^{wt}_{assemb}\theta^{wt}_{assemb}[N\text{-}gRNA^{wt}]$
38	$[N\text{-}gRNA^{dip}] \to [N\text{-}gRNA^{dip}] - 1,\ [SP] \to [SP] - n^{dip}_{SP},$ $[V^{dip}_{assembled}] \to [V^{dip}_{assembled}] + 1$	$k^{dip}_{assemb}\theta^{dip}_{assemb}[N\text{-}gRNA^{dip}]$
39	$[N\text{-}gRNA^{wt}] \to [N\text{-}gRNA^{wt}] - 1$	$d^{wt}_{N\text{-}gRNA}[N\text{-}gRNA^{wt}]$
40	$[N\text{-}gRNA^{dip}] \to [N\text{-}gRNA^{dip}] - 1$	$d^{dip}_{N\text{-}gRNA}[N\text{-}gRNA^{dip}]$
41	$[V^{wt}_{assembled}] \to [V^{wt}_{assembled}] - 1,\ [V^{wt}_{released}] \to [V^{wt}_{released}] - 1$	$k^{wt}_{release}[V^{wt}_{assembled}]$
42	$[V^{wt}_{assembled}] \to [V^{wt}_{assembled}] - 1$	$d^{wt}_{assembled}[V^{wt}_{assembled}]$
43	$[V^{wt}_{released}] \to [V^{wt}_{released}] - 1$	$d^{wt}_V[V^{wt}_{released}]$
44	$[V^{dip}_{assembled}] \to [V^{dip}_{assembled}] - 1,\ [V^{dip}_{released}] \to [V^{dip}_{released}] - 1$	$k^{dip}_{release}[V^{dip}_{assembled}]$
45	$[V^{dip}_{assembled}] \to [V^{dip}_{assembled}] - 1$	$d^{dip}_{assembled}[V^{dip}_{assembled}]$
46	$[V^{dip}_{released}] \to [V^{dip}_{released}] - 1$	$d^{dip}_V[V^{dip}_{released}]$

5. MATHEMATICAL MODELS OF VIRAL INFECTION

Dose response analysis

In Figure 5.4 the release kinetics depending on initial doses $[V_{free}^{wt}](0) = [V_{free}^{dip}](0) = 10$ was examined for the deterministic model. However, one can expect that initial infection doses might vary from cell to cell. Therefore, examining the release kinetics of WT virions under different initial conditions seemed prudent. Figure 5.7 illustrates the total of WT virions (**left**) and DIPs (**right**) released with initial conditions $[V_{free}^{wt}](0) = $ 1-20 and $[V_{free}^{dip}](0) = $ 0-20 over a 24 hour time period. As can be seen from Figure 5.7, a low dose of DIPs (MOI= 1) with a high dose of WT virus (MOI= 20) results in $\approx$21% reduction of the WT particles released during DIP co-infection. Furthermore, as the initial number of WT virions was decreased with DIPs fixed at a MOI= 1, a continued decrease in WT virus released during co-infection is observed compared to the single infection case. As the dose of DIPs was increased, the total of WT virions released rapidly decreased. By a MOI=10 for WT and MOI= 4 for DIPs, WT particles only account for approximately 30% of particles released. These deterministic results were consistent with median estimates from the stochastic model presented in Figure 5.8. In contrast, the mean estimates in Figure 5.9 showed marginally higher release in WT virus and lower release of DIPs.

Figure 5.10 shows the result from stochastic simulations for varying initial doses of wild type virus ($[V_{free}^{wt}](0) = $ 1-10) and DIPs ($[V_{free}^{dip}](0) = $ 0-10). This figure illustrates that the efficacy of DIPs is dose-dependent, with a higher dosage leading to a reduction in WT virus released regardless of the WT initial dose. This is the same trend seen in Figures 5.8 and Figure 5.9, with mean predictions higher than median predictions. Figure 5.11 illustrates the probability of a productive infection. A productive infection is defined as the WT population releasing virus before going extinct. The likelihood of a productive infection tends to be one as the initial dosage of the WT virus hits 20 virus particles, likely due to increased available resources for replication. However, the probability is affected by the initial dosage of DIPs, which is reduced as the dose increases. The increased number of DIPs will sequester more replication machinery, leading to fewer copies of WT virions. Additionally, productive infection is almost guaranteed for high
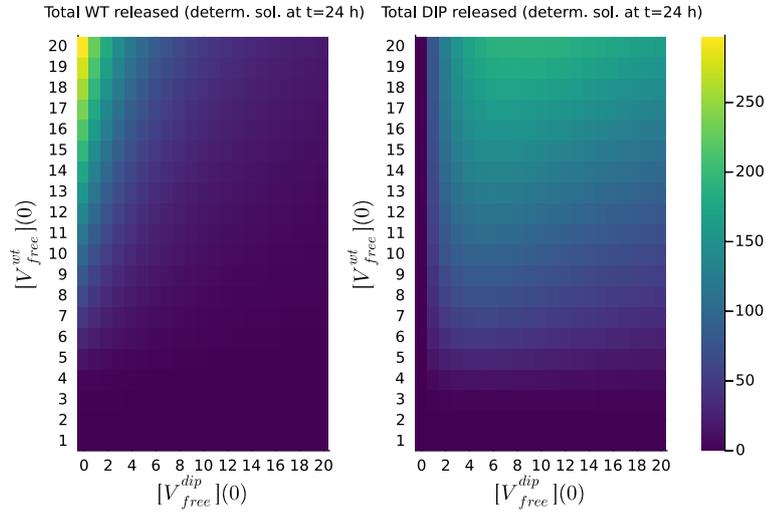
Figure 5.7: Effects of varying initial dose on viral particle release in the deterministic model. **Left:** Total WT virions released over the 24 hours post-infection for varying initial conditions of free WT virions $[V_{free}^{wt}](0) = 1\text{-}20$ and free DIPs $[V_{free}^{dip}](0) = 0\text{-}20$. **Right:** Total DIPs released for varying initial doses.
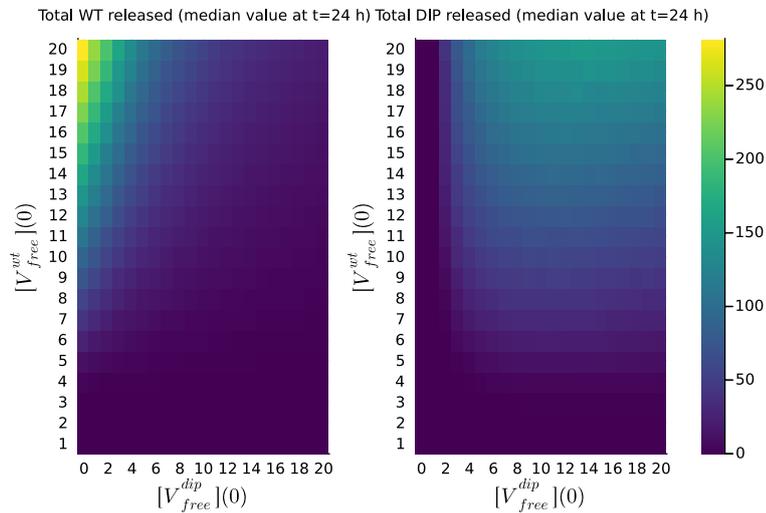
Figure 5.8: Effects of varying initial dose on viral particle release in the stochastic model. **Left:** Median total WT virions released over the 24 hours post-infection for, varying initial conditions of free WT virions $[V_{free}^{wt}](0) = 1\text{-}20$ and free DIPs $[V_{free}^{dip}](0) = 0\text{-}20$. **Right:** Total DIPs released for varying initial doses.
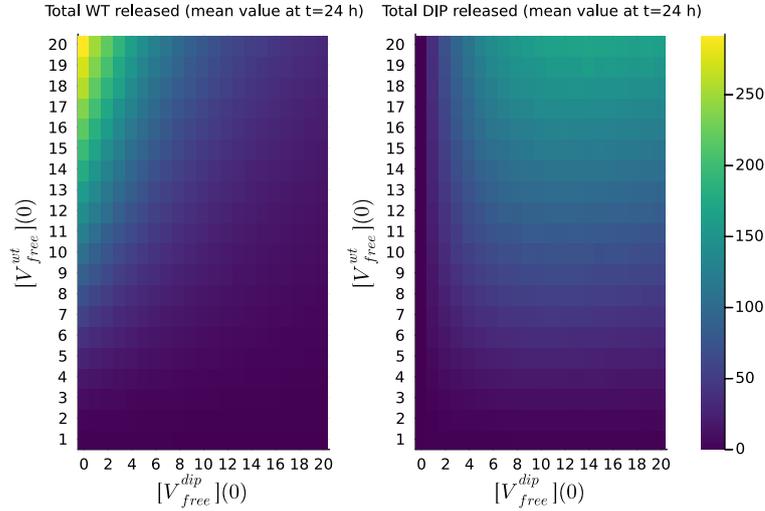
Figure 5.9: Figure: Effects of varying initial dose on viral particle release in the stochastic model. **Left:** Expected total WT virions released over the 24 hours post-infection, for varying initial conditions of free WT virions $[V_{free}^{wt}](0) = 0\text{-}20$ and free DIPs $[V_{free}^{dip}](0) = 0\text{-}20$ from the stochastic model. **Right:** Total DIP particles released for varying initial doses.
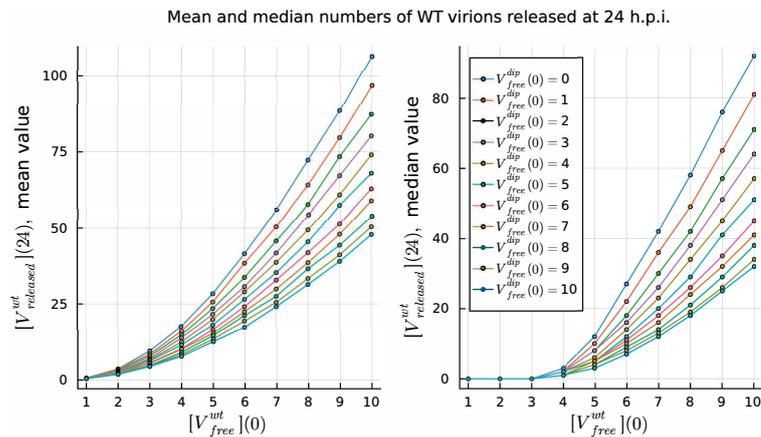
Figure 5.10: Stochastic model outputs for varying initial conditions. The mean and median values for the distribution of WT virions released at 24 hours post-infection, for varying initial conditions of free WT virions $[V_{free}^{wt}](0) = 1\text{-}10$ and free DIPs $[V_{free}^{dip}](0) = 0\text{-}10$.
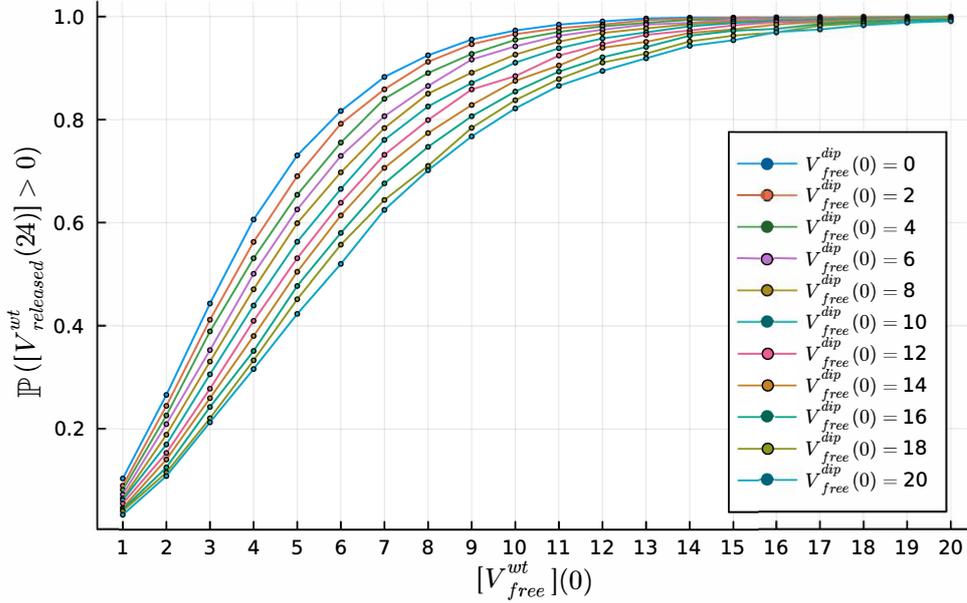
Figure 5.11: The probability of a productive infection $\mathbb{P}(V_{released}^{wt}(24) > 0)$ for varying initial doses of both WT $[V_{free}^{wt}](0) = 1\text{-}20$, and DIPs $[V_{free}^{dip}](0) = 0\text{-}20$ at 24 hours post-infection.

dosages of the WT virus. Still, as shown from Figure 5.8, the overall number of WT and potentially infectious particles is reduced even if an infection is guaranteed.

Figure 5.12 shows virus particle release kinetics predicted by the deterministic model with fixed initial conditions for WT virions $[V_{free}^{wt}](0) = 10$ and varying initial conditions for DIPs $[V_{free}^{dip}](0) = 1\text{-}100$. DIP release peaks at a MOI= 6 and then decreases as the dose increases. This is due to a lack of WT virus proteins that allow replication of DIPS, and as such, DIPs reach a replication threshold. An increase in dose continues to affect the release of WT virions, so that for an $MOI = 40$ total WT virion production is < 30 virions released over the 24 hours considered. This highlights the ability of DIPs to compete (with an advantage) for replication resources with WT virions. Consequently, if the dose is high enough, DIPs sequester so many intra-cellular resources that WT production is significantly reduced.
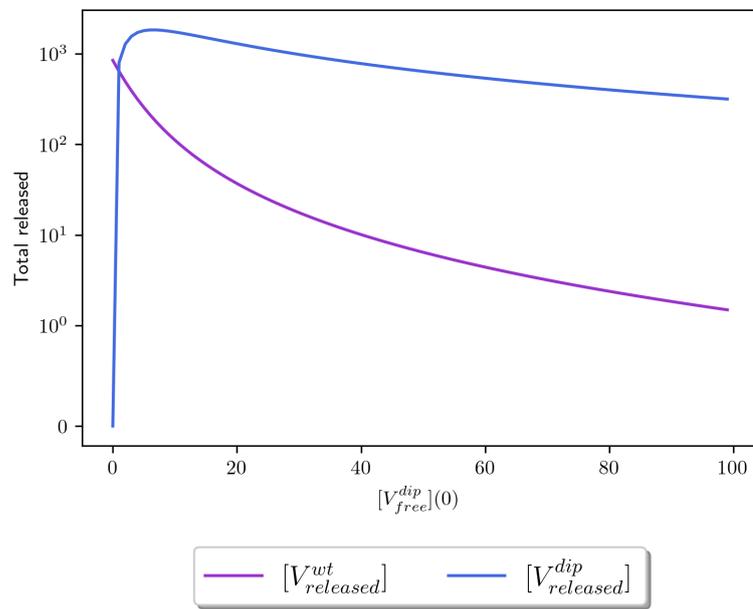
Figure 5.12: Total WT released after 24 hours for increased initial doses of DIPs. Virus particle release kinetics predicted by the model with fixed initial conditions for WT virions $[V^{wt}_{free}](0) = 10$, and varying initial conditions for DIPs $[V^{dip}_{free}](0) = 1\text{-}10^2$. **(Purple line:)** illustrates the release of WT virions $[V^{wt}_{released}]$ with DIP release **(blue line)** $[V^{dip}_{released}]$.

5.1 Defective interfering particles as a potential therapeutic during SARS-CoV-2 infection

Discussion

In this section, using a mathematical model, I investigated the intra-cellular replication kinetics of the WT virus in the presence of DIPs. To this end, I extended the model proposed by Grebennikov *et al.* (2021), which focused on the intra-cellular replication kinetics of SARS-CoV-2, to include co-infection with defective interfering particles, given their therapeutic potential (Rand *et al.*, 2021; Roux *et al.*, 1991). In particular, I investigated the ability of DIPs to reduce WT viral load by competing for resources required to replicate and encapsulate the viral genome to form new virions. Since DIP genomes lack key fragments, they need a "helper" virus, which encodes non-structural and structural proteins, for their replication. There is evidence of DIPs leading to a reduction in virus titres for several viruses, including influenza A, dengue fever and SARS-CoV-2 (Bdeir *et al.*, 2019; Chaturvedi *et al.*, 2021; Li *et al.*, 2021). With the emergence of new SARS-CoV-2 strains, the effectiveness of DIPs (derived from a particular viral strain) against novel ones remains to be investigated.

I aimed to assess the hypothesis that DIPs lead to a reduction not only in WT virions released but also, negatively impact the transcription of positive-sense genomic RNAs. Additionally, I investigated the effects of initial infection dose (WT and DIP) in releasing new WT virions and DIPs. Since experimental data sets are minimal, it is essential to note that the parameter values obtained in this section are based on the data set from Chaturvedi *et al.* (2021), and are not globally identifiable; hence an infinite set of parameter values could be considered. Identifying unique parameters would require additional experimental data. It was shown that four experimentally observed outputs for $[V_{released}^{dip}]$, $[V_{released}^{wt}]$, $[V_{endosome}^{dip}]$, $[V_{assembled}^{dip}]$ were sufficient to make fifteen of the eighteen new parameters locally identifiable. Hence, there would be a finite set of parameter values that would satisfy the data. Furthermore, it would require eight experimental outputs $[V_{released}^{dip}]$, $[V_{released}^{wt}]$, $[V_{endosome}^{dip}]$, $[V_{assembled}^{dip}]$, $[SP]$, $[N]$, $[NSP]$ and $[gRNA_{(-)}^{dip}]$ to make the parameters globally identifiable and as a result unique.

The extension of the model presented by Grebennikov *et al.* (2021) required new parameters to account for the kinetics of DIPs. Therefore, it was necessary to investigate the sensitivities of all model parameters. In particular, I made use of

179

5. MATHEMATICAL MODELS OF VIRAL INFECTION

Sobol sensitivity analysis to understand how variation in parameter values affects four different model outputs: $[gRNA^{wt}]$, $[gRNA^{dip}]$, $[V^{wt}_{released}]$, and $[V^{dip}_{released}]$. I found several parameters that affect all four model outputs: K_{NSP}, the threshold number of non-structural proteins, $k^{wt/dip}_{tr(-)}$, transcription rates of negative sense genomic RNA for WT virus and DIPs, respectively, and $k^{wt/dip}_{tr(+)}$, the transcription rates for positive sense genomic RNA. The rates associated with cell entry, k_{fuse} and k_{uncoat}, also lead to variation in model outputs. Finally, suppose I examine the sensitivity of the outputs WT and DIP release. In that case, I find that their associated assembly rates, $k^{wt}_{assembl}$ and $k^{dip}_{assembl}$, are the most sensitive parameters.

DIPs have the potential as therapeutics; thus, it is essential to explore how initial infection doses of WT and DIP alter the release of WT virus, to inform a treatment plan. I show that even a low MOI= 1 of DIPs can cause a reduction of approximately 50% in released WT virus compared to an infection in the absence of DIPs. Figure 5.7 illustrates how increasing the dose of DIPs leads to a reduction in the fraction of released WT virions, independently of the initial WT infection dose. These trends are consistent with the results from the stochastic model also shown within this paper (Figure 5.8). The doses of both WT virus and DIPs also affected the probability of a productive infection which decreased with increased DIPs but is almost guaranteed for high doses of WT virus. I also investigated, the effect of the initial MOI of DIPs given a fixed amount of WT virus (MOI=10) on particle release. My results show that while DIP release peaks at an initial DIP dose of MOI=5, the release of WT virions decreases in a dose-dependent manner. Furthermore, by an initial DIP dose of MOI=40 WT virion release is effectively inhibited.

The deterministic and stochastic models I presented are a good first approximation to the kinetics of WT and DIP co-infection. Yet, there are several biological processes which have not been considered. First and foremost, I omitted the anti-viral response of the cell. While I need not consider the adaptive immune response since the time interval is 48 hours, the innate immune response would play a pivotal role (Dempsey *et al.*, 2003; tenOever, 2016). As discussed in Chapter 4, viral RNA can be detected, leading to the production of type IFNs, which can inhibit viral replication and eventual viral clearance. Furthermore, innate immune responses have been shown to be induced by DIPs binding to PRRs,

providing additional stimuli and magnifying the anti-viral cellular response (Rand *et al.*, 2021). Consequently, it would be ideal to extend the proposed model to consider the role of an innate immune response. Another limitation of the model is that I do not distinguish between infectious and non-infectious particles for WT virions. This would be important to understand the potential infectivity of the virus particles released. I also failed to characterise the natural generation of DIPs during the WT replication cycle (which is inherently characterised by mutations). This process would contribute to releasing other defective interfering particles, potentially reducing the number of WT virions released. However, fully calibrating such a model would require data which is not currently available.

To conclude, I believe the proposed model shows the potential benefits of DIPs as a therapeutic tool to reduce WT virus production. I also have shown that even low doses of these particles can positively limit WT virus production and reduce the probability of a successful infection. This reduction continues, in a dose-dependent manner, to reduce virus production significantly. Future work will focus on incorporating immune responses and the natural production of DIPs into the mathematical model presented here. Still, it will require further carefully curated data to assist in parameter estimation, which was a limitation of this model and investigation. Alternatively, certain aspects of the model could be simplified to reduce the model complexity and number of parameters that need to be calibrated.

5.2 In vitro replication of Foot and Mouth Disease Virus

Foot-and-mouth disease virus (FMDV) of genus *Aphthovirus*, family *Picornaviridae* causes disease in livestock, especially cattle, pigs, sheep, and several other species (Grubman & Baxt, 2004). It is endemic in many countries and can result in large outbreaks; examples include the 1997 Taiwan outbreak (Yang *et al.*, 1999) and the United Kingdom in 2001 (Gibbens *et al.*, 2001; Thompson *et al.*, 2002). This has led to global concern due to the loss of cattle and financial implications that impacted the tourism and agricultural sectors. The 2001 UK outbreak resulted in a £3.1 billion loss to the British economy and the destruction of 4 million animals to control measures (Thompson *et al.*, 2002).

The dynamics of disease spread have been thoroughly investigated using mathematical modelling, including investigation of transmission routes such as farm-farm and within-farm. Additionally, models have examined the effects of control strategies on the overall spread of disease (Hayama *et al.*, 2013; Mushayabasa *et al.*, 2011). In contrast, very little has been done to model the within-host dynamics of FMDV infection. Howey *et al.* (2012) formulated a deterministic model that investigated within-host dynamics along with an antibody and IFN response, utilising *in vivo* data sets for parameter calibration and analysis. However, there is little investigation into the *in vitro* modelling of FMDV, so it seems fitting to investigate it.

Here I wish to investigate a simple model to capture *in vitro* replication dynamics of FMDV similar to those previously studied for Influenza virus and EBOV (Gonzàlez-Parra *et al.*, 2018; Liao *et al.*, 2020). I will perform parameter calibration using experimental data provided by Eva Perez from *The Pirbright Institute* along with analysing parameter identifiability and sensitivity (Wainwright *et al.*, 2014). Furthermore, I will investigate strain variability of parameters using data from infections with South African Territories (SAT) strains 1-3. Additionally, I examine the basic reproduction number (Yang, 2014) and time to infection (Gonzàlez-Parra *et al.*, 2018). Due to the lack of stochastic models for FMDV in the literature, I will propose a stochastic model to reflect the random nature of the infection process. This stochastic model will use parameters found

using the deterministic fitting. From this stochastic model, I will investigate the mean time to extinction depending on different initial infective doses. Furthermore, the probability of the virus population going extinct in the experimental window of 72 hours will also be investigated. First, I provide a brief overview of the experimental protocol and data curated by Eva Perez from *The Pirbright Institute*.

5.2.1 FMDV in vitro experimental data

The following describes the experimental protocol used to obtain the experimental data available. For growth curves, 1×10^5 IBRS2 cells were seeded in 24 well plates. Cells were infected with South African territories (SAT) 1, SAT2 or SAT3 at different MOIs for 1 hour at 37°C (MOIs are orientative as viruses were not titrated by plaque assay). Cells were washed thrice in PBS, and 1 ml of completed media was added to the wells. Plates were stored at -80°C at different times postinfection (0, 4, 8, 24, 28, 32 and 72 hours). For titrating the supernatants, plates were defrosted, the supernatant was transferred to a tube, centrifuged at 13000 rpm for 5 minutes to pellet the cell debris, and the supernatants were transferred to a new tube. Supernatants were then titrated by $TCID_{50}$ in IBRS2 cells (cells infected for 30 hours and the max viral titre tested in the plate is 6.2 $\log_{10} TCID_{50}$). The results of these experiments are summarised in Table 5.7.

Mock yield assays were carried out as follows: $100\mu L$ aliquots of SAT1, SAT2 and SAT3 were incubated at 37°C and stored at -80°C at different time points post-incubation (0, 2, 4, 6, 8, 24 and 28 hours). Samples were titrated by $TCID_{50}$ in IBRS2 cells after 48h postinfection and three-fold dilution across the plate, (max titres of 7.3 $\log_{10} TCID_{50}$/ml). The results of these mock yield assays are summarised in Table 5.8.

5.2.2 Deterministic model

Given the experimental data collected by Eva Perez, I wish to propose a mathematical model, described by a system of ordinary differential equations (ODEs), to characterise viral replication *in vitro* for FMDV. Equations (5.25)-(5.28) have been used previously in both EBOV and influenza infections (Gonzàlez-Parra *et al.*, 2018; Liao *et al.*, 2020). However, compared to the models used by Gonzàlez-Parra

	SAT1			
MOI	1.6×10^{-1}	1.6×10^{-2}	1.6×10^{-3}	1.6×10^{-4}
Time (hours)				
0	0	0	0	0
4	3.41	2.47	0	0
8	3.88	2.24	1.77	0
24	4.59	6.2	6.2	6
28	4.12	5.06	4.59	4.59
32	4.12	5.06	6	5.53
72	0			0

	SAT2			
MOI	4×10^{-1}	4×10^{-2}	4×10^{-3}	4×10^{-4}
Time (hours)				
0	0	0	0	0
4	3.41	2.47	0	0
8	3.88	3.41	2.24	0
24	4.82	5.76	5.06	4.35
28	4.35	5.06	5.29	4.12
32	3.18	4.35	5.06	5.06
72	0			0

	SAT3			
MOI	10^{-3}	10^{-4}	10^{-5}	10^{-6}
Time (hours)				
0	0	0	0	0
4	0	0	0	0
8	0	0	0	0
24	5.06	3.65	3.41	2.24
28	5.06	4.35	3.65	2.47
32	5.53	4.59	4.59	4.35
72	0			0

Table 5.7: $\log_{10} \text{TCID}_{50}/ml$ measurements for FMDV growth curves for each strain of South African territories (SAT) 1, 2 and 3. Each strain was tested with four initial MOIs, with measurements taken over a 72-hour experimental window.

Time (hours)	SAT1: MOI 0.16	SAT2: MOI 0.4	SAT3: MOI 0.001
0	5.12	5.59	3
2	3.94	4.41	0
4	3.47	4.41	0
6	2.77	3.47	0
8	2.77	2.77	0
24	0	0	0
28	0	0	0

Table 5.8: Mock yield assay results for each strain of SAT FMDV. Each mock yield assay was performed over 28 hours with a single MOI for each strain. Measurements are in $\log_{10} \mathrm{TCID}_{50}/ml$.

et al. (2018) and Liao *et al.* (2020), I will not consider an Erlang distribution of the eclipse and infective phases. I will assume these stages follow an exponential distribution instead. A model schematic for Equations (5.25)-(5.28) is illustrated in Figure 5.13.
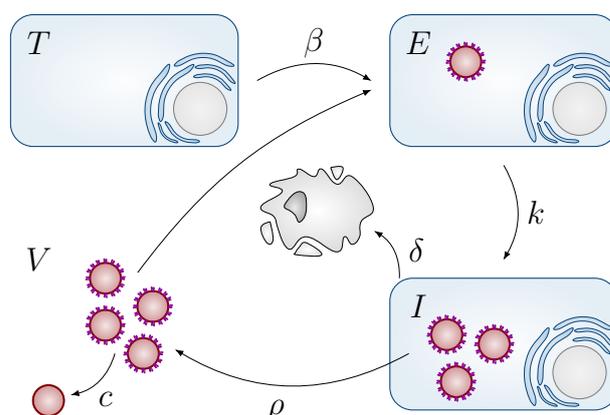
Figure 5.13: Model schematic of viral replication used to describe viral replication with target cells (T), eclipse phase (E), infected cells (I) and free virus (V).

$$\frac{dT}{dt} = -\beta TV \; , \tag{5.25}$$

$$\frac{dE}{dt} = \beta TV - kE \; , \tag{5.26}$$

$$\frac{dI}{dt} = kE - \delta I \; , \tag{5.27}$$

$$\frac{dV}{dt} = \rho I - cV \; . \tag{5.28}$$

This mathematical model considers target cells (T), which become infected with rate β ($\frac{mL}{\text{TCID}_{50}\cdot h}$) in the presence of infectious virus (V). These newly infected cells enter a eclipse phase (E), where with rate k (h^{-1}) they become productively infected cells (I). Infected cells produce new infectious virus with rate ρ ($\frac{\text{TCID}_{50}}{cell\cdot h}$), and die with rate δ (h^{-1}). Infectious virus decays with rate c (h^{-1}) becoming non-infectious. I assume that cells will only die after they have entered the infected compartment. Furthermore, assume that infected cells will only produce infectious particles until death. Since the experiments did not quantify non-infectious and infectious virus, it would not possible to quantify a model that had both of these virus types. This mathematical model aims to capture the infectious virus quantified by experimental TCID$_{50}$ values. I will use this model to characterise infection of three distinct viral strains of FMDV: South African territories (SAT) 1, 2 and 3. Each SAT strain was tested at four different multiplicities of infection (MOI), which I aim to characterise with the model. Experimental data will allow one to perform parameter calibration as described in the next section. The interpretation of each parameter is summarised in Table 5.9 along with units.

5.2.3 Parameter calibration

Before performing parameter calibration, assessing the structural identifiability and sensitivities was essential. Analysing these will inform me which parameters can be determined from the currently available data and which must be carefully chosen from literature. I will use Sobol global sensitivity analysis to assess the importance of parameters in the model output, which will be infectious virions released (V). I include the initial virus dose V_0 in this analysis. Figures 5.14a

(a) First-order Sobol indices.

(b) Total-order Sobol indices.
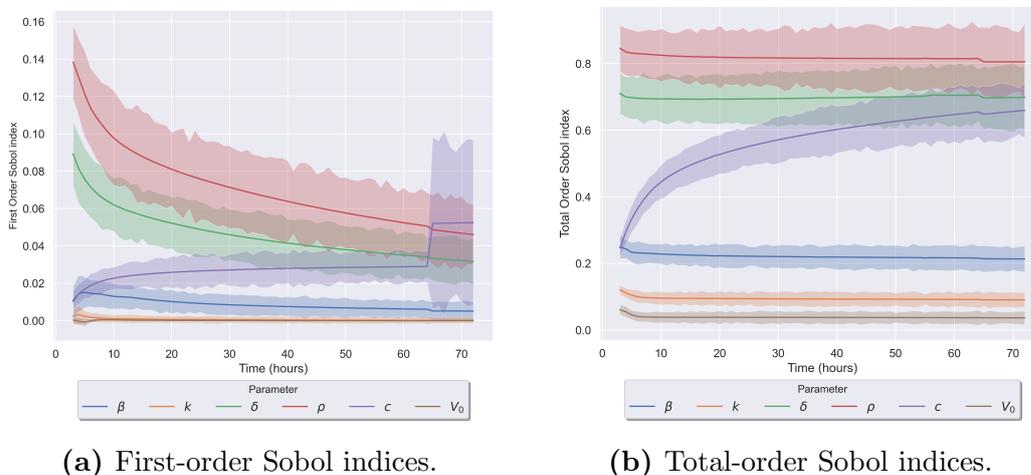
Figure 5.14: First and Total-order Sobol indices with respect to time. A time course of 72 hours was chosen to reflect the experimental design. I include the initial dose of virus V_0 in my analysis to see how this affects the model output. Model output considered is V for infectious virions released.

and 5.14b show the first and total-order Sobol indices with respect to the time course. These figures show that δ, ρ and c are the most important parameters since they cause the largest variation in the model output. This is what intuition would suggest since ρ is the virus's production rate, determining the amount of virus released from a cell. While, c is the rate at which the virus is cleared. δ is important since this is the death rate of infected cells. The longer an infected cell is alive, the more virus can be replicated before the cell dies. The infection rate β was the remaining parameter affecting model outputs, while the parameter k and initial condition V_0 carry little consequence on the model output. Consequently, low first-order Sobol indices indicate I should examine higher-order sensitivities. Figure 5.15 illustrates interactions between parameters with ρ and δ along with ρ and c having the largest interactions. If c and δ are small compared to ρ, more virions can be produced, so their interaction with ρ is important.

To assess the structural identifiability, SIAN (structural identifiability analyser) was used. This online tool determined that β is a globally identifiable parameter given that the virus released over time is the only experimental output quantified. c, δ and k are locally identifiable and ρ is an unidentifiable parameter (Ilmer *et al.*,
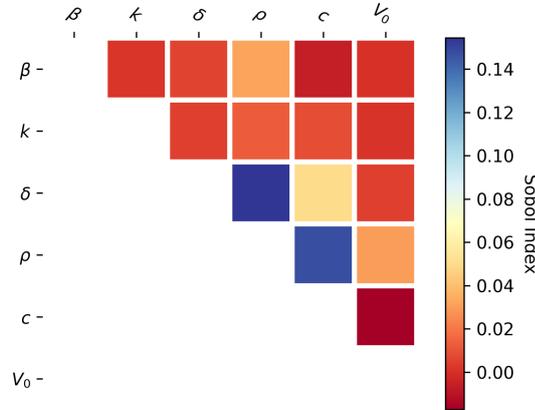
Figure 5.15: Second-order Sobol interactions. These quantify the effect varying one parameters has on other parameters. It can be seen that δ and ρ have the largest interactions followed closely by ρ and c.

2021) given the single experimental output. Global identifiability of β implies that I can approximate the true value given the data, while for c, δ and k multiple combinations of these parameter values will fit the current data. Furthermore, for ρ, I do not expect to learn anything from parameter calibration. Additional data would be required to make all parameters globally identifiable and improve the model fit. For example, if the number of infected cells was tracked and provided as an experimental output, it would be possible to identify global solutions of β, c and ρ. Furthermore, δ and k would be locally identifiable with only two solutions per parameter.

Infectivity decays over time exponentially (Gonzàlez-Parra *et al.*, 2018; Liao *et al.*, 2020) and as such, in the absence of target cells can be described by $V(t) = V_0 \exp^{-ct}$ where V_0 is the initial concentration of virus ($\frac{\text{TCID}_{50}}{ml}$). Since mock yield experimental data was available (as summarised in Table 5.8), it can approximate the value of c in the ODEs. Hence if one assumes the data follows an exponential decay, it can be transformed into a straight line by taking the log base 10. Therefore it can be described by,

$$\log_{10}(V(t)) = \log_{10}(V_0) - ct. \tag{5.29}$$

As such, linear regression can be used to determine the values of c since this is now an equation of a straight line. Linear regression algorithms in R were
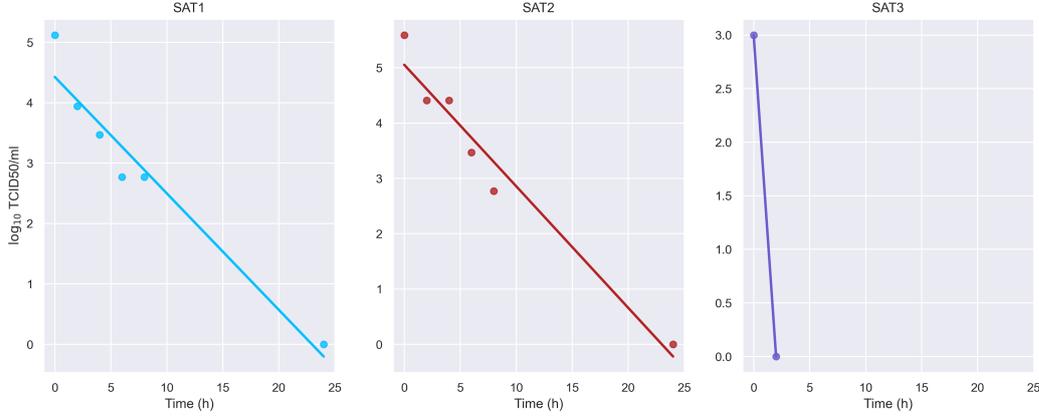
Figure 5.16: Regression lines fitted using a linear model function in R. The parameter c is defined as $0.19h^{-1}$ (p-value 1.2×10^{-3}), $0.22h^{-1}$ (p-value 6.5×10^{-4}) and $0.66h^{-1}$ for SAT1, SAT2 and SAT3 strains respectively. P-values were calculated to a 95% confidence level.

used to calculate these values. I determine that c is defined as $0.19h^{-1}$ (p-value 1.2×10^{-3}), $0.22h^{-1}$ (p-value 6.5×10^{-4}) and $0.66h^{-1}$ for SAT1, SAT2 and SAT3 strains respectively. These best fits from the regression are illustrated in Figure 5.16. Since values for the viral decay rate c of each strain have been estimated using mock yield assay data and linear regression, this left four further parameters to estimate. From identifiability analysis, I did not expect to learn about the true value of ρ, and it is possible to find multiple suitable values for δ and k. β should be estimated from what data is currently available. Approximate Bayesian computation (ABC) will be used to obtain parameter estimates as laid out in Section 2.4.1. For each strain, there were four multiplicity of infections (MOIs) with a single time-course for each (no replicates). Due to the lack of replicates, using a Euclidean distance measure seemed logical. In this parameter calibration, the distance measure is defined such that,

$$d(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) = \sqrt{\sum_{t \in \mathcal{T}} (x(t) - y(t))^2} \,,$$

where $\mathcal{T}$ is the set of time points in the data set, $x(t)$ denotes the output from the mathematical model at time t for parameters $\boldsymbol{\theta}$, and $y(t)$ represents experimental data at time t. Here I will compare the free virus V to the experimental data.

Parameter	interpretation	search range	Units
β	rate of infection	$[10^{-7}, 10^{-3}]$	$\frac{mL}{\text{TCID}_{50}\cdot h}$
k	(length of eclipse phase)$^{-1}$	$[10^0, 10^2]$	h^{-1}
δ	(length of infective phase)$^{-1}$	$[10^0, 10^2]$	h^{-1}
ρ	viral production rate	$[10^0, 10^4]$	$\frac{\text{TCID}_{50}}{cell\cdot h}$
c	rate of viral decay		h^{-1}

Table 5.9: Table of parameters with biological interpretation and search ranges used in approximate Bayesian computation. Values for c are fixed using linear regression. c is defined as $0.19h^{-1}$ (p-value 1.2×10^{-3}), $0.22h^{-1}$ (p-value 6.5×10^{-4}) and $0.66h^{-1}$ for SAT1, SAT2 and SAT3 respectively.

Since the experimental method has an initial infection time of one hour and then cells are washed, I incorporated this into the model. To account for this I will run the model from $t=0$ to $t=1$ with initial conditions $T(0)=10^5$, $E(0)=I(0)=0$, with the initial virus calculated as,

$$V_0 = T_0 \times \text{MOI},$$

where the MOIs are defined in Table 5.10. I then use the result from an initial hour to run this model from $t=1$ until $t=72$ where the new initial conditions were defined as $T(1)=T^*$, $E(1)=E^*$, $I(1)=I^*$ and $V(1)=0$. Here T^*, E^*, and I^* are the results from the initial hour. For each strain, I will run the model to generate 10^7 parameter sets where the top 1% are accepted such that the total Euclidean distance for all four provided MOIs is minimised.

The results of the ABC algorithm are summarised in Table 5.11 and a model fit with posterior distributions in Figure 5.17. As seen for all strains of SAT, the analysis learns a lot about parameters β as illustrated in narrower posteriors compared to prior distributions. Consequently, SAT1 and SAT2 have narrower

Strain	MOIs
SAT1	$[1.6 \times 10^{-1}, 1.6 \times 10^{-2}, 1.6 \times 10^{-3}, 1.6 \times 10^{-4}]$
SAT2	$[4 \times 10^{-1}, 4 \times 10^{-2}, 4 \times 10^{-3}, 4 \times 10^{-4}]$
SAT3	$[1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}, 1 \times 10^{-6}]$

Table 5.10: List of MOIs for each FMDV virus strain.

posteriors than SAT3. This is likely due to poor SAT3 mock yield data providing a poor estimate of the viral decay rate c. For parameters k and δ, I have learned very little about these parameter values with posterior distributions similar to the priors. As highlighted previously through structural identifiability, multiple possible parameter values will suit the available data. There is some learning of the parameter ρ, but overall the posterior distributions are still wide. From this, it can be seen that the median rate of infections β is lower for SAT1 ($10^{-6.14}$ $mL/TCID_{50} \cdot h$) than SAT2 ($10^{-5.74}$) and SAT3 ($10^{-5.46}$), and k and δ are comparable for all three strains. Furthermore, in agreement with Perez-Martin *et al.* (2022) the production rate ρ for SAT1 ($10^{1.99}$ $TCID_{50}/cell \cdot h$) is higher than that of SAT2 ($10^{1.50}$) and SAT3 ($10^{1.52}$). For SAT1 and SAT2 there are negative correlations between β and ρ ($r = -0.411$ and $r = -0.414$ respectively). Additionally, there are strong positive correlations between δ and ρ ($r = 0.871$ and $r = 0.889$ for SAT1 and SAT2 respectively). These are illustrated in Figures 5.18 and 5.19 for SAT1 and SAT2 respectively. These correlations indicate that the analysis provides knowledge of the ratios β/ρ and δ/ρ. This is also illustrated in Figures 5.18 and 5.19.

Time to infection and R_0 number

The time to infection can be calculated using the equation

$$T_{\text{inf}} = \sqrt{\frac{2}{\rho\beta}}.$$

(a) SAT1

(b) SAT2

(c) SAT3
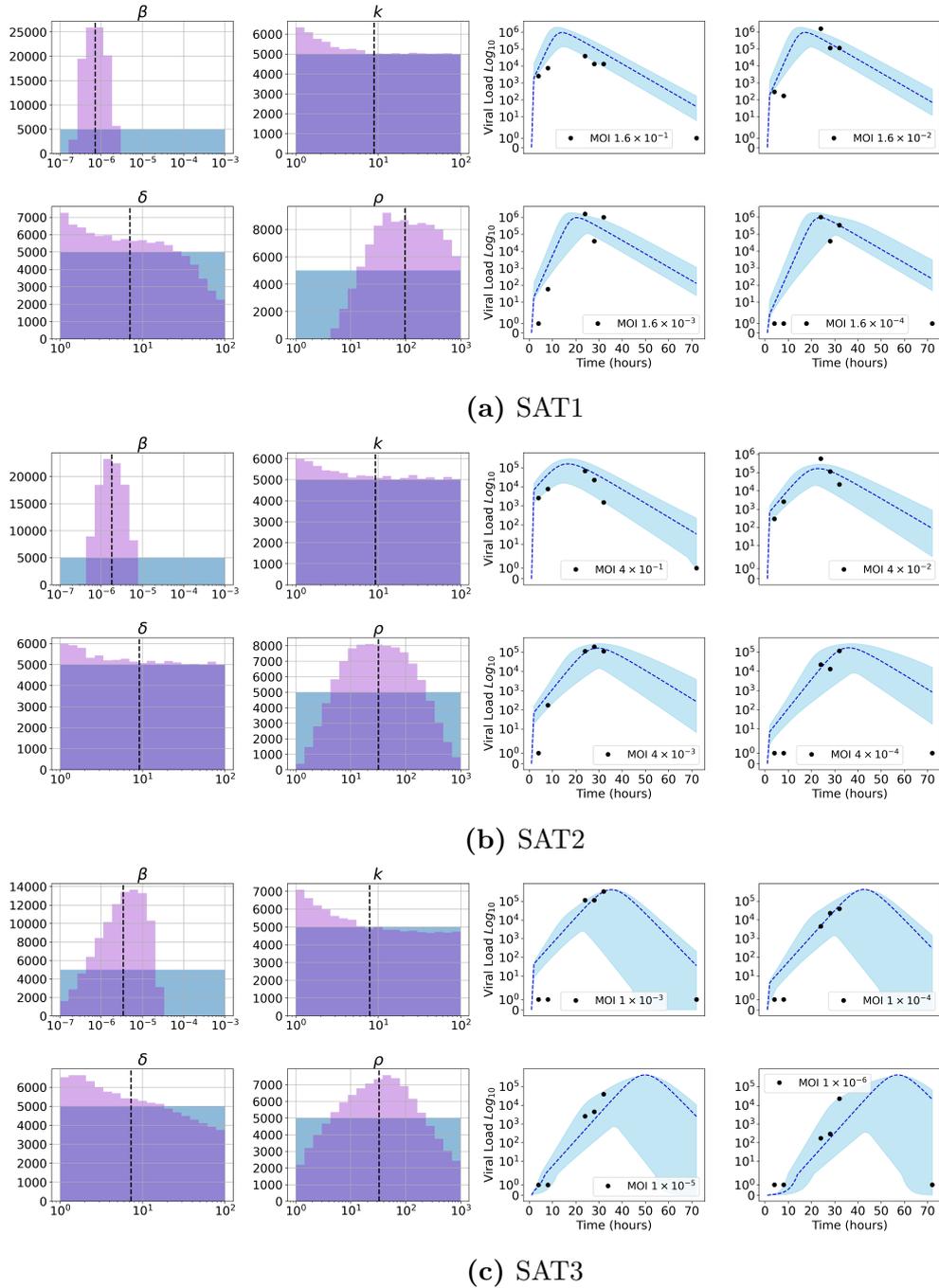
Figure 5.17: Posterior probability distributions of parameters for the model presented in Equations (5.25)-(5.28) in purple from 10^5 accepted parameter sets. blue indicates prior distributions before ABC. Median model fit from 10^5 accepted parameters sets. The Blue dashed line represents a point-wise median model fit, with light blue 95% credible intervals. Black dots represent experimental data for each MOI.
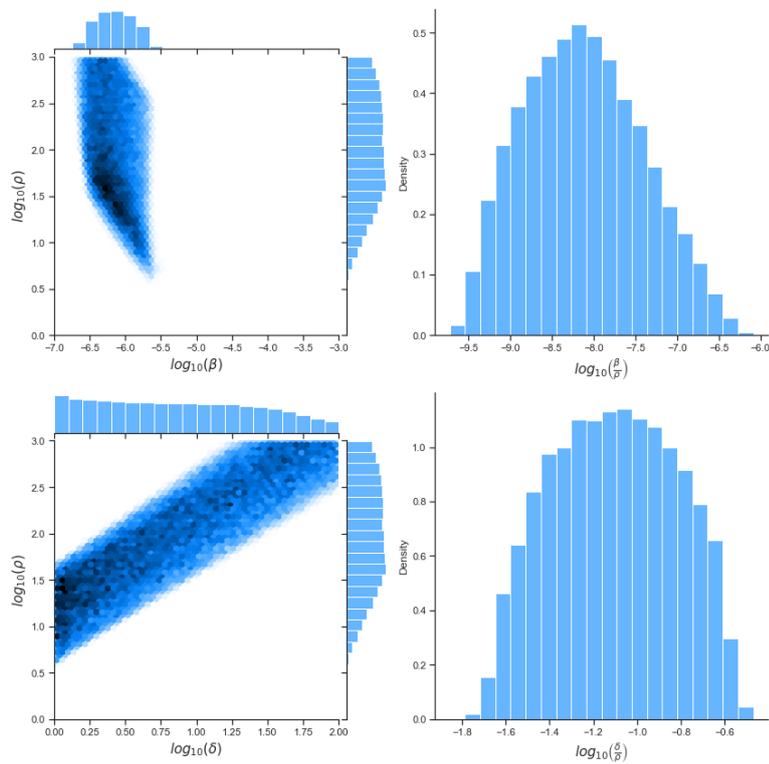
Figure 5.18: Correlated parameters β and ρ ($r = -0.411$) top left figure with a density histogram of the ratio in the top right figure. Highly correlated parameters δ and ρ ($r = 0.871$) bottom left with a density histogram of the ratio in the bottom right. All parameters from fitting SAT1 experimental data.
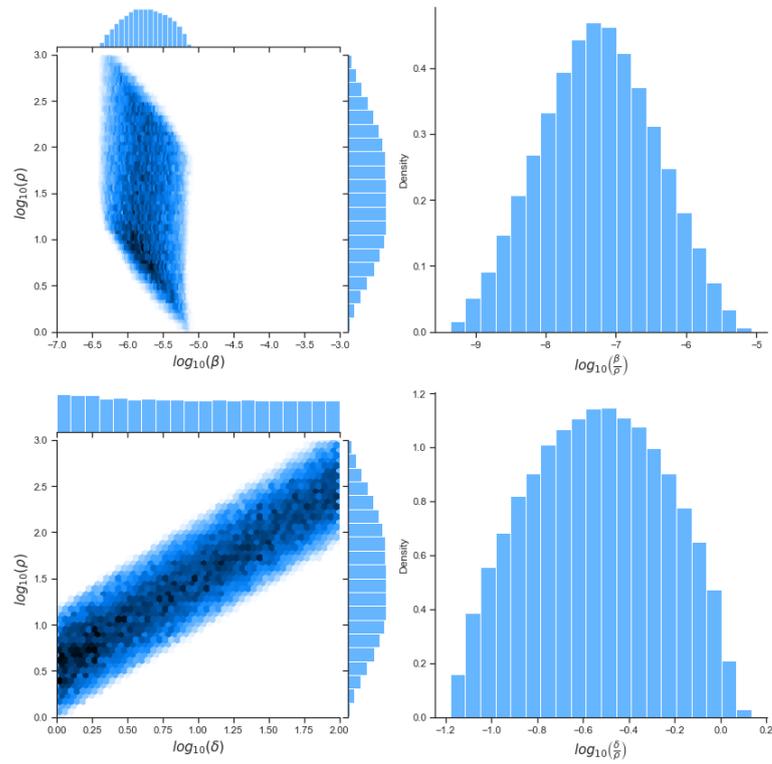
Figure 5.19: Correlated parameters β and ρ ($r = -0.414$) top left figure with a density histogram of the ratio in the top right figure. Highly correlated parameters δ and ρ ($r = 0.889$) bottom left with a density histogram of the ratio in the bottom right. All parameters from fitting SAT2 experimental data.

Parameter	SAT1	SAT2	SAT3
β	$10^{-6.14[-6.58,-5.69]}$	$10^{-5.74[-6.27,-5.22]}$	$10^{-5.46[-6.71,-4.66]}$
k	$10^{0.95[0.04,1.95]}$	$10^{0.97[0.04,1.95]}$	$10^{0.90[0.03,1.94]}$
δ	$10^{0.85[0.03,1.88]}$	$10^{0.96[0.04,1.95]}$	$10^{0.86[0.04,1.93]}$
ρ	$10^{1.99[0.91,2.93]}$	$10^{1.50[0.36,2.69]}$	$10^{1.52[0.17,2.84]}$
c	0.19	0.22	0.66

Table 5.11: Estimated parameters of FMDV infection in vitro for SAT1, SAT2 and SAT3. The data show a median with 95% credible region.

The infecting time is when a single new cell is infected as described by Gonzàlez-Parra *et al.* (2018). I find the infection time T_{inf} is defined as 169.28 (h), 186.42 (h) and 133.26 (h) for SAT1, SAT2 and SAT3 respectively. The basic reproduction number R_0 can be calculated using next-generation matrix methods as described by Diekmann *et al.* (2010). From Equations (5.25)-(5.28) I have two infectious compartments in the model, these are

$$\frac{dE}{dt} = \beta TV - kE \ ,$$
$$\frac{dI}{dt} = kE - \delta I \ . \tag{5.30}$$

In a disease-free equilibrium I have that $dV/dt = 0$ so I find V can be expressed as,

$$V = \frac{\rho}{c}I.$$

Consider the equations in (5.30) written in the form $dx_i/dt = F_i(x) - V_i(x)$ for $i = 1, 2, \ldots, m$. In this arrangement, $F_i(x)$ is the rate of appearance of new infections in compartment i, and $V_i(x)$ is the rate of other transitions between compartment i and other infectious compartments (Diekmann *et al.*, 2010; Van den

Driessche, 2017). Therefore one can define two matrices F and V which are defined as,

$$F = \left[\frac{\partial F_i(x_0)}{\partial x_j} \right] , \qquad V = \left[\frac{\partial V_i(x_0)}{\partial x_j} \right] , \qquad \text{for } 1 \le i, j \le m. \qquad (5.31)$$

From this the basic reproduction number is defined,

$$\mathcal{R}_0 = \boldsymbol{\rho}(FV^{-1})$$

where $\boldsymbol{\rho}$ is the spectral radius. Consequently, performing this analysis with the equations (5.30) the matrices F and V are defined to be,

$$F = \begin{pmatrix} 0 & \frac{\beta\rho T_0}{c} \\ 0 & 0 \end{pmatrix} , \qquad V = \begin{pmatrix} k & 0 \\ -k & \delta \end{pmatrix}.$$

Thus, to determine the basic reproduction value, the spectral radius of,

$$FV^{-1} = \begin{pmatrix} \frac{\beta\rho T_0}{c\delta} & \frac{\beta\rho T_0}{c\delta} \\ 0 & 0 \end{pmatrix}. \qquad (5.32)$$

It can be shown that the disease free equilibrium would be defined as $X_0 = (T_0, 0, 0, 0)$. Hence the non-zero basic reproduction number is to be defined as,

$$\mathcal{R}_0 = \frac{\rho\beta T_0}{c\delta}.$$

From median value prediction of the parameters listed in Table 5.11 and given the initial number of cells in the experiment was $T_0 = 10^5$, I find the basic reproduction number for SAT1, SAT2 and SAT3 to be 5.196, 2.85 and 2.54 respectively. This shows for SAT1 and SAT2, infection is always guaranteed, with SAT1 leading to more than double the secondary infections than SAT2. Following the trend in the data, SAT3 leads to fewer secondary infections than the other two circulating strains in agreement with Perez-Martin *et al.* (2022).

Stochastic Model

Now consider a stochastic model of the system previously studied, with a fixed initial population of cells T_0. As previously introduced, T is the random variable for the number of uninfected cells, E eclipse phase cells, I infected cells and released virus V. These will be modelled using a stochastic approach: as a continuous time Markov process (CTMC) $\mathcal{X} = \{\boldsymbol{X}(t) = (T(t), E(t), I(t), V(t)) : t \geq 0\}$. Here, $T(t)$, $E(t)$, $I(t)$ and $V(t)$ are the number of each population from T, E, I and V at time t. Lets denote $(T(t), E(t), I(t), V(t)) = (n_T, n_E, n_I, n_V) = \mathbf{n}$. Here, the condition is imposed that $n_T + n_E + n_I \leq T_0$ as such I can define the state space $\mathcal{S}$ of $\mathcal{X}$ for $t \geq 0$ as $\mathcal{S} = \{(n_T, n_E, n_I, n_V) \in (\mathbb{N} \cup \{0\})^4 : n_T + n_E + n_I \leq T_0\}$. The transition rates can be defined as,

$$
q_{(\mathbf{n}, \mathbf{n}')} = \begin{cases}
\beta n_T n_V & \text{if } \mathbf{n}' = (n_T - 1, n_E + 1, n_I, n_V - 1), \\
k n_E, & \text{if } \mathbf{n}' = (n_T, n_E - 1, n_I + 1, n_V), \\
\delta n_I, & \text{if } \mathbf{n}' = (n_T, n_E, n_I - 1, n_V), \\
\rho n_I, & \text{if } \mathbf{n}' = (n_T, n_E, n_I, n_V + 1), \\
c n_V, & \text{if } \mathbf{n}' = (n_T, n_E, n_I, n_V - 1), \\
0, & \text{otherwise.}
\end{cases}
$$

This stochastic formulation can be compared against the deterministic formulation and also allows the study of stochastic descriptors using Gillespie simulations. In the experiments presented, several different initial doses were considered. From the plots presented in Figure 5.17, the viral population would be expected to go extinct if the time course were extended. It would therefore be interesting to investigate two stochastic descriptors relating to the extinction of the viral population. First is the probability for the viral population to go extinct within the 72-hour experimental time frame, as several of the populations go extinct according to the data. Since there is no influx of new susceptible cells, the viral population will always go extinct given enough time. Therefore the probability can be defined as,

$$
\mathbb{P}(V(t) = 0 | V(0) = iT(0)), \qquad t \leq 72,
$$

where i will be a initial MOI in the range $[10^{-5}, 1]$, and each strain of FMDV will be evaluated. The second stochastic descriptor will be the expected time

for the population to reach extinction for a given initial dose of the virus. It is interesting to examine how long the viral population can persist *in vitro* with a limited number of susceptible target cells to infect. Let T^0 be the time for the virus population to become extinct. Then a descriptor can be defined as,

$$\mathbb{E}[T^0|V(0) = iT(0)].$$

Matrix analytic methods such as those introduced in Section 4.1.1 could be used here. However, due to these experiments' large initial population sizes, this method is not computationally possible with the current high-performance systems. This results from the matrices associated with the method getting too large to store and solve efficiently. Instead, stochastic simulation algorithms will be used as introduced in Section 2.3.1.
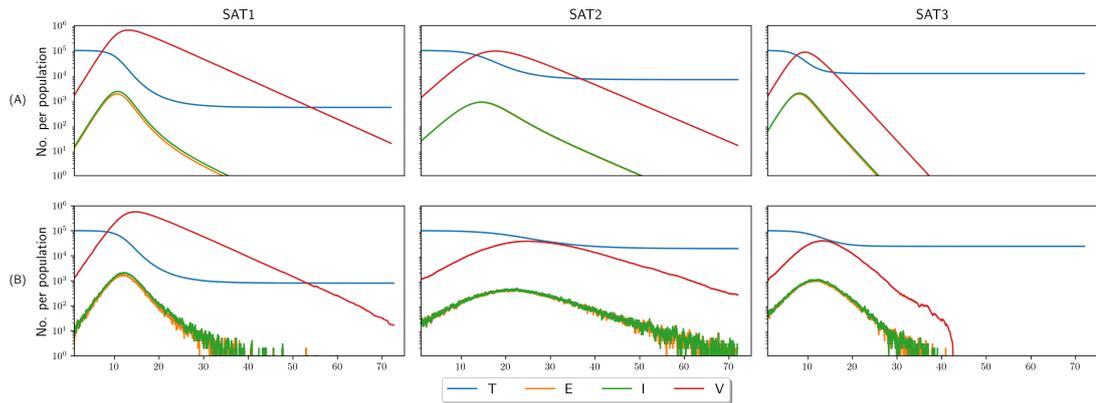
Figure 5.20: (A) Deterministic model output using median parameter values for each strain summarised in Table 5.11, using initial MOI= 0.01. (B) Single realisation of the Gillespie stochastic simulation algorithm for each strain using median parameter values.

Figure 5.20(A) shows illustrative examples of the deterministic model output using median parameter values from Table 5.11 for an initial MOI of 0.01. Figure 5.20(B) shows a single realisation of the Gillespie stochastic simulation algorithm described in Section 2.3.1 for the same set of parameter values and initial MOI. To investigate the mean time to extinction, 10^4 realisations of the Gillespie algorithm were run for a range of MOIs in $[10^{-5}, 1]$. The results are
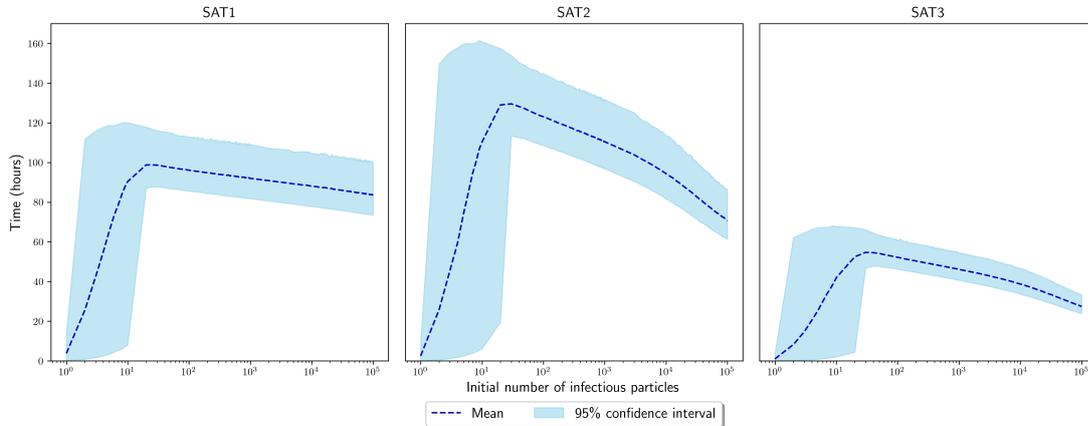
Figure 5.21: Mean time to extinction of the viral population with MOIs in the range $[10^{-5}, 1]$ for each viral strain using median parameters values. A 95% confidence region is plotted for each. Means for each MOI result from 10^4 realisations from the Gillespie algorithm.

illustrated in Figure 5.21 with a 95% confidence interval. SAT1 and SAT2 present similar behaviour when considering the initial MOI and time to extinction. Between $[0, 10]$ initial infectious viral particles, the time to infection rapidly increases for both, with SAT1 increasing from ≈ 3.7 (C.I $[0.09, 13.8]$) hours to ≈ 90 (C.I $[8.5, 120]$) hours while SAT2 goes from ≈ 2.5 (C.I $[0.07, 9.2]$) hours to ≈ 110 (C.I $[6.1, 160]$) hours. SAT1 achieves a maximum mean time to extinction of 98.7 hours for initial infectious particles between $V_O \in [20, 30]$. From here, the time to extinction begins to decrease; however, this is gradual. For SAT1, the stochastic simulations also disagree with the experimental outputs. Simulations suggest the viral population will never go extinct for higher infectious MOIs within the 72-hour experimental window, as observed. This is further supported by Figure 5.22, which illustrates the probability of going extinct for each viral strain for a range of initial MOIs. As can be seen from Figure 5.22 above MOIs of $10^{-3.31}$, the viral population is guaranteed to survive longer than 72 hours. SAT2 reaches a maximum mean time until extinction at an initial infectious dose of 30 viral particles, after this, the mean time begins to decrease. Unlike SAT1, as the initial MOI surpasses 10^{-1}, the probability of the viral population going extinct in the 72 hours increases as demonstrated by Figure 5.22. This is in agreement

with the experimental observations for the higher doses. It also indicates that for SAT2, a high infectious dose does not guarantee a long-term persistent infection. This is likely because SAT2 has a lower production rate ρ, which leads to fewer viruses. Additionally, there is a higher death rate for infected cells and a higher decay rate for virions than SAT1. SAT2-infected cells, therefore, have less time to produce new viruses and less time for new viruses to infect target cells. In the initial infection, the higher decay rate also means viruses decay before infecting a target cell. As a result, not every cell becomes infected, and there is a chance for the infection to become extinct before every cell is infected. For strains SAT1 and SAT2, Figures 5.21 and 5.22 highlight the potential of small infectious doses in the range of MOIs $[10^{-5}, 10^{-4}]$ to cause a successful and persistent infection. SAT3 follows a similar trend to SAT1 and SAT2 with the maximum mean time of 54.6 hours to extinction between $[20:30]$ initial infectious particles. The mean time then decreases similarly to SAT1 and SAT2. However, from Figure 5.22, since the maximum mean time is 54.6, the probability that SAT3 goes extinct in the experimental window is guaranteed with probability 1. However, it is worth remembering due to the mock yield assay for SAT3 leading to a high decay rate, this is to be expected and is not necessarily reflective of the actual dynamics.

Discussion

In the preceding work, I did a preliminary investigation into modelling *in vitro* replication dynamics for Foot and Mouth disease virus. Since little mathematical modelling was done in this scenario, it seemed appropriate to investigate the replication dynamics with this preliminary work forming part of a proposal submitted by Dr. Anna Jolles in collaboration with Eva Pérez. Following the ideas used by Gonzàlez-Parra *et al.* (2018) and Liao *et al.* (2020), a deterministic mathematical model was introduced to characterise replication. The sensitivity analysis performed on the model showed that the three most important parameters that cause variance in virus release are: the rate of decay c, the death rate of infected cells δ and the viral production rate ρ. Parameters c and ρ are directly related to the ODE for virus particles, so this makes sense as to why they are important. δ is also important since if cells die at a higher rate, then the virus
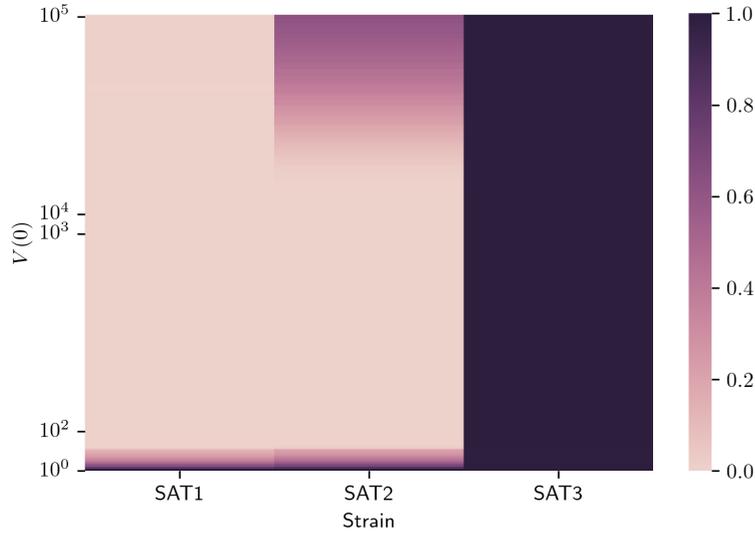
Figure 5.22: Probability of extinction within the 72-hour experimental time window for each viral strain. These probabilities are generated using 10^4 realisations of the Gillespie algorithm for each MOI in the $[10^{-5}, 1]$ range.

will have to reproduce at a higher rate to compensate and be able to continue to infect cells. This is further supported by the second-order interactions of ρ with δ and c.

Another point to investigate was what can be determined about parameter values given limited data, mock yield assays and growth curves, even with a simple model. Since mock yield assays were available, it allowed the determination of the decay rate c for each viral strain, while the others would need to be determined through parameter fitting. Structural identifiability analysis showed that only β would be globally identified for the available growth curves, with δ and k only being locally identifiable while ρ would be unidentifiable. If additional experimental data were available, such as the number of infected cells during the experiment, β and ρ would be globally identifiable, with δ and k only having two unique solutions. The available data for parameterisation achieved good model fits with relatively narrow credible intervals for SAT1 and SAT2 as shown in Figure 5.17, but SAT3 had much larger intervals. Figure 5.17 also supports the results of identifiability analysis for β and ρ. It is shown in Table 5.11 that SAT1 has the lowest rate of infectivity. At the same time, SAT2 and SAT3 are similar. However, SAT1 has

a higher viral production rate; this agrees with Perez-Martin *et al.* (2022) who saw genome in serum earlier for SAT1 compared to SAT2 and SAT3 *in vivo*. The parameterisation of this model for each strain also allowed determination of the *in vitro* basic reproduction number, which showed for SAT1, SAT2 and SAT3 that since $\mathcal{R}_0 > 1$ a productive infection is always possible.

From this deterministic model, a stochastic approach was developed to investigate the effect of initial MOI doses on the mean time to extinction of the viral population and the probability of it going extinct in the experimental window. This was interesting because the experiments for each MOI indicated the population would go extinct within 72 hours. Stochastic simulations agreed that this is the case for SAT3, which is guaranteed with probability 1 as indicated in Figures 5.22 and 5.21. However, for SAT1 and SAT2, stochastic simulations did not agree with experimental growth curves. For SAT1, extinction was only guaranteed for low infectious MOIs and was not for MOIs greater than 10^{-5} within the experimental window, in disagreement with the experimental results. SAT2 follows a similar trend as SAT1 for low MOIs, but unlike SAT1, for MOIs larger than 10^{-1}, the probability of extinction begins to increase; however, not with probability 1. This could be due to the median parameter values obtained through the fitting, or it could be due to the formulation of the model not fully encapsulating the mechanism that led to a reduction in viral titres seen in the experiments.

Even though the work in this chapter is a preliminary investigation, it is important to discuss several limitations and potential improvements that could be made to both the deterministic and stochastic approaches. First, unlike Gonzàlez-Parra *et al.* (2018), and Liao *et al.* (2020), Erlang distributed compartments for cells in the eclipse and infected compartments have not been used. This could result in different behaviour in the overall replication kinetics with the potential for the eclipse phase to be drawn out and delay the production of new virions. For infected cells, this could lead to more opportunities for virus production. However, with limited data, it would be difficult to identify how many intermediate stages would be needed to characterise this data correctly. Another point of contention is that the virus has not been analysed as infectious and non-infectious particles. Since during viral replication, non-infectious interfering

and non-interfering particles are generated (Brooke, 2014), it is important to distinguish the difference in a model since if it was not included, it leads to an overestimation of the infectious virus generated. Another factor and a running theme of this thesis that has not been considered is innate immune responses. In this short time frame of 72 hours, the innate immune system will play a significant role in controlling the replication of FMDV. Therefore immune dynamics should be incorporated into the model to make it a more accurate representation.

One major drawback of the model presented is the characterisation of released infectious virions. While virions egress from the cell, accumulation of virus within the cell and damage to the plasma membrane from egress eventually results in cell lysis (Sarry *et al.*, 2022). This catastrophe will result in a "burst" of new infectious viruses into the extracellular medium. The deterministic and stochastic models presented do not consider this mechanism of cell death, an essential part of the virus replication dynamics. One approach for the deterministic model would include an extra term relating to virus-induced cell lysis and a burst term N for the number of viruses released. For a stochastic approach, one could use the methods employed by Williams *et al.* (2021) and Carruthers *et al.* (2020) to develop a birth and death process with catastrophe. These methods would allow one to calculate a distribution of rupture sizes for a cell and the mean time to rupture. This can be incorporated into a within-host model with dose responses already examined for the three strains. While this work is only preliminary, I believe it presents an excellent start to investigating *in vitro* replication dynamics. Exploring and expanding the mathematical model to incorporate additional biological details is worthwhile.

Appendix A

Appendix

A.1   St Jude children's Research Hospital SARS-CoV-2 trace study

Figures A.1 and A.2 are both referenced in Chapter 3 but placed here due to their low statistical significance results. Figure A.1 illustrates the hCCoV antibody response for all four endemic strains depending on race. Figure A.2 illustrates antibody responses to SARS-CoV-2 proteins depending on race.
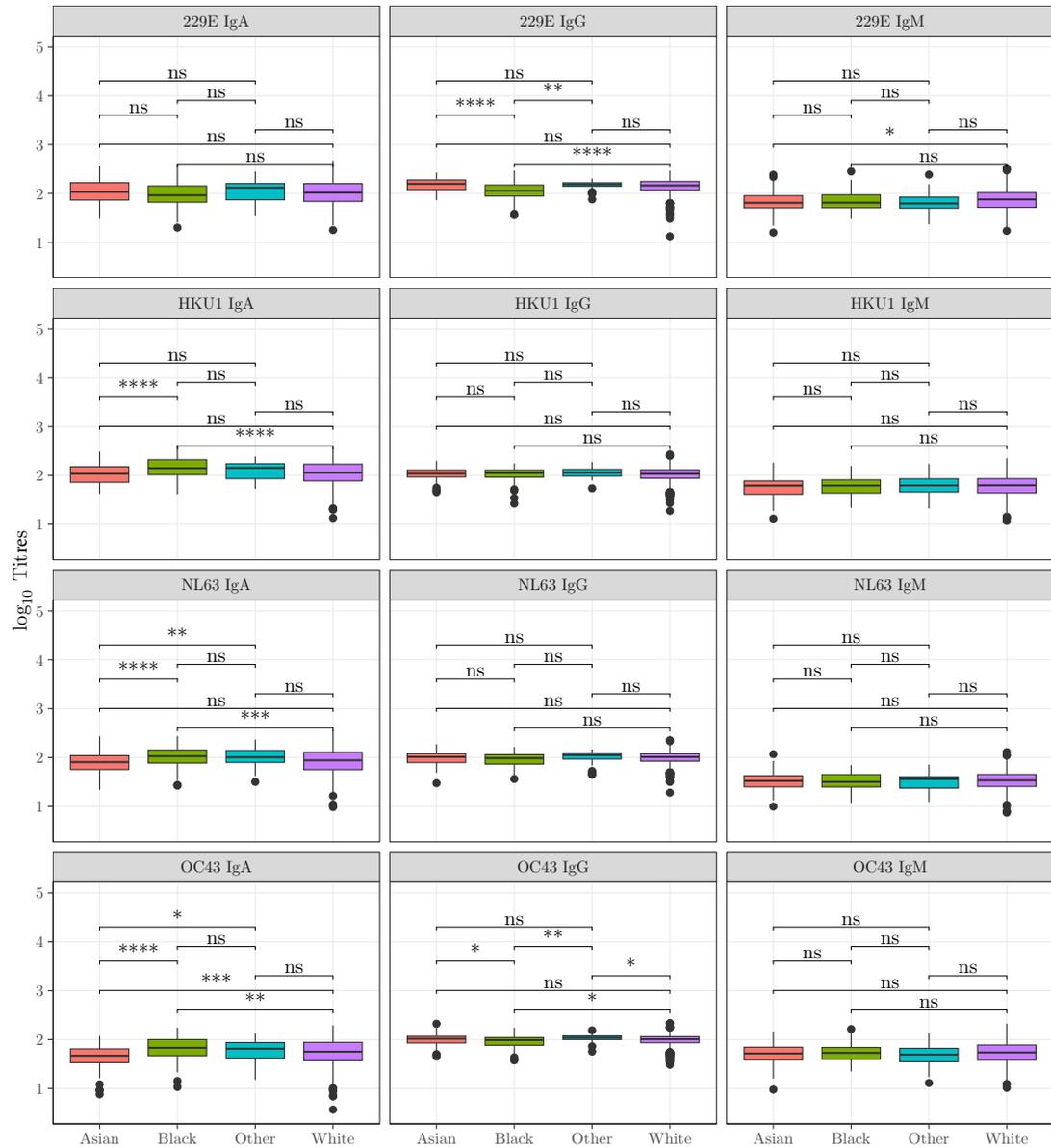
Figure A.1: Box plots of hCCoV antibody isotype response compared between races. Statistical significance was determined by the Wilcoxon–Mann–Whitney test with Bonferroni adjustment (ns, not significant; *$p < 0.05$ **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$).
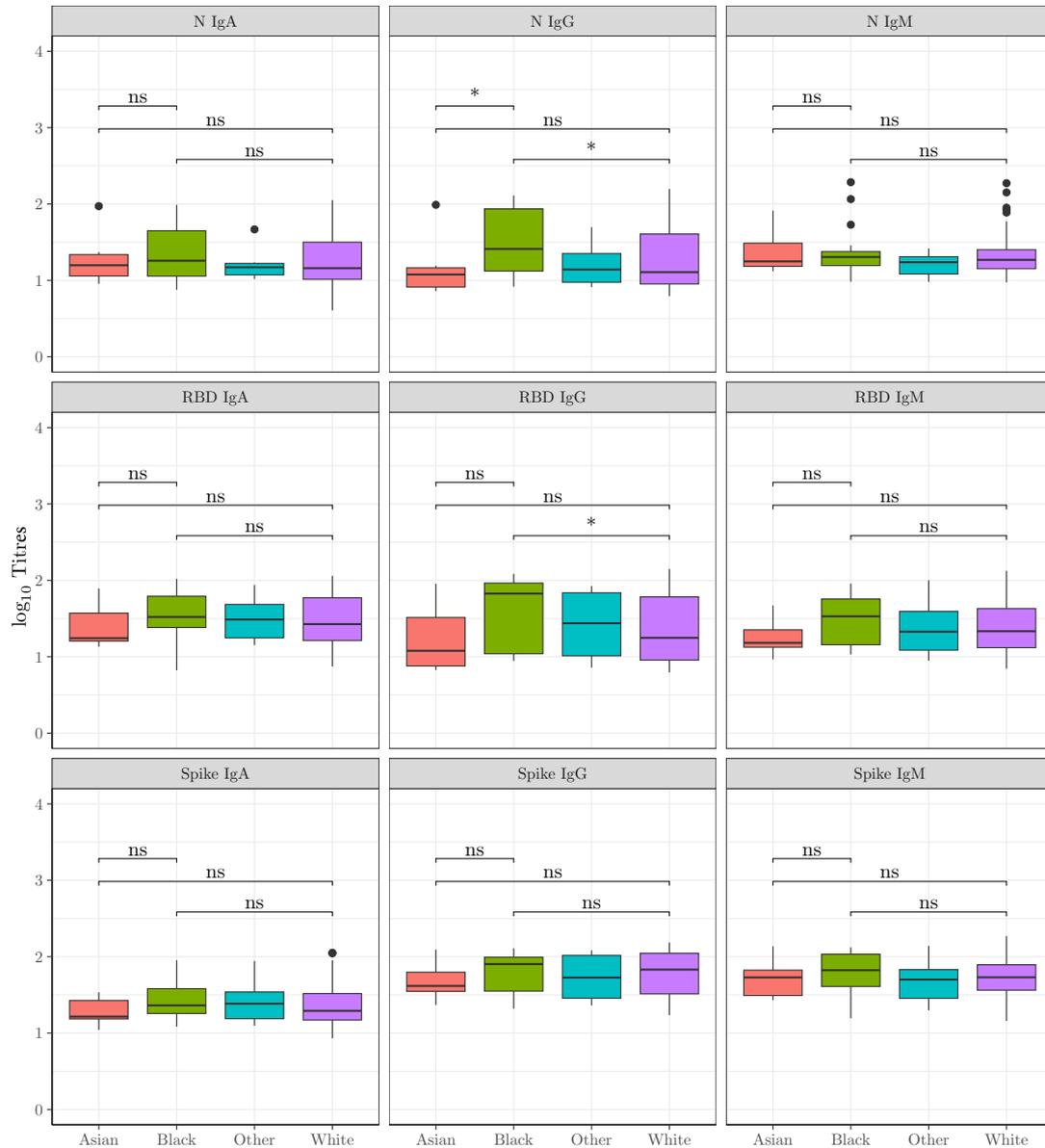
Figure A.2: Box plots of SARS-CoV-2 antibody isotype response compared between races for infected individuals with no vaccination ($n = 159$). Statistical significance was determined by the Wilcoxon–Mann–Whitney test with Bonferroni adjustment (ns, not significant; *$p < 0.05$ **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$).

A.2 Matrix analytic methods in practice

In Chapter 4, matrix expressions were introduced to calculate the steady state distribution and expected time to reach a threshold level of phosphorylated TBK-1 n_3 for model 1. Due to the initial number of proteins found during parameter estimation, matrix analytical approaches were impossible. The matrix sizes become too large to store in the system memory. As a result, in Chapter 4, stochastic simulation techniques were used instead. Here I present an example of using the matrix analytical approach to obtain the steady-state distribution and the algorithm to calculate them. The algorithm to calculate the steady state is presented in Algorithm A.1. The results of this calculation are presented in Figure A.3 where the initial number of proteins is set $n_D = 30$, $n_R = 10$, $n_V = 10$, $n_B = 5$ and all other parameters are the medians presented in Chapter 4. I also provide the algorithm for the Laplace-Stieltjes transforms $\boldsymbol{g}^N(z)$ and the l-th order moments $\boldsymbol{m}^{N,(l)}$ of time to reach a given state $T_{(n_1,n_2,n_3)}(N_3)$ (Algorithm A.2).

Algorithm A.1: Steady state distribution $\boldsymbol{\pi}$

$\boldsymbol{H}_0 = \boldsymbol{A}_{00}$
for $k = 1, \ldots, n_B - 1$ **do**
  $\quad \boldsymbol{H}_k = \boldsymbol{A}_{kk} - \boldsymbol{A}_{kk-1}\boldsymbol{H}_{k-1}^{-1}\boldsymbol{A}_{k-1k}$
end
$\boldsymbol{\pi}_{n_B}^* = \boldsymbol{1}$
for $k = n_B - 1, \ldots, 0$ **do**
  $\quad \boldsymbol{\pi}_k^* = -\boldsymbol{\pi}_{k+1}^*\boldsymbol{H}_k^{-1}$
end
for $k = 0, \ldots, n_B$ **do**
  $\quad \boldsymbol{\pi}_k = \frac{1}{\sum_{j=0}^{n_B} \boldsymbol{\pi}_j^* \boldsymbol{e}_j(r)}\boldsymbol{\pi}_k^*$
end

Figure A.3: Example of steady-state distributions using matrix analytical methods where initial proteins are set to $n_D = 30$, $n_R = 10$, $n_V = 10$, $n_B = 5$ and all other parameters are the medians presented in Chapter 4.

Algorithm A.2: Algorithm for the Laplace-Stieltjes transforms $\boldsymbol{g}^N(z)$ and the l-th order moments $\boldsymbol{m}^{N,(l)}$

Part 1
$\boldsymbol{H}_0^N(z) = \boldsymbol{I}_j(0) - \boldsymbol{A}_{00}(z)$
for $k = 1, \ldots, N-1$ **do**
$\quad \boldsymbol{H}_k^N(z) = \boldsymbol{I}_{j(k)} - \boldsymbol{A}_{kk}(z) - \boldsymbol{A}_{kk-1}(z)\boldsymbol{H}_{k-1}^N(z)^{-1}\boldsymbol{A}_{k-1,k}(z)$
$\quad \boldsymbol{g}_{N-1}^{N,(0)} = \boldsymbol{H}_{N-1}^N(z)^{-1}\boldsymbol{c}_{N-1}(z)$
end
$\boldsymbol{m}_{N-1}^N(z)^{-1} = \boldsymbol{g}_{N-1}^N(0)$
for $k = N-2, \ldots, 1, 0$ **do**
$\quad \boldsymbol{g}_k^N(z) = \boldsymbol{H}_k^N(z)^{-1}\boldsymbol{A}_{kk+1}(z)\boldsymbol{g}_{k+1}^N(z)$
$\quad \boldsymbol{m}_k^{N,(0)} = \boldsymbol{g}_k^N(0)$
end
Part 2
$\boldsymbol{m}_{N-1}^{N,(0)} = \boldsymbol{g}_k^N(0)$
for $N-2, \ldots, 1, 0$ **do**
$\quad \boldsymbol{m}_k^{N,(0)} = \boldsymbol{g}_k^N(0)$
end
for $p = 1, \ldots, l$ **do**
$\quad \boldsymbol{P}_0^{N,(p)} = \sum_{k=1}^p \binom{p}{k}(-1)^k(\boldsymbol{A}_{00}^{(k)}(0)\boldsymbol{m}_0^{N,(p-k)} + \boldsymbol{A}_{01}^{(k)})\boldsymbol{m}_1^{N,(p-k)}\}$
end
for $j = 1, \ldots, N-1$ **do**
$\quad \boldsymbol{P}_j^{N,(P)} =$
$\quad \boldsymbol{A}_{j,j-1}(0)\boldsymbol{H}_{j-1}^N(0)^{-1}\boldsymbol{P}_{j-1}^{N,(p)} + \sum_{k=1}^p \binom{p}{k}(-1)^k(\boldsymbol{A}_{j,-1}^{(k)}\boldsymbol{m}_{j-1}^{N,(p-k)} +$
$\quad \boldsymbol{A}_{jj}^{(k)}(0)\boldsymbol{m}_j^{N,(p-k)} + (1 - \delta_{j,N-1}\boldsymbol{A}_{jj+1}^k(0)\boldsymbol{m}_{j+1}^{N,(p-k)})$
$\quad \boldsymbol{m}_{N-1}^{N,(p)} = \boldsymbol{H}_{N-1}^N(0)^{-1}(\boldsymbol{P}_{N-1}^{N,(p)} + (-1)^p\boldsymbol{c}_{N-1}^p(0))$
end
for $j = N-2, \ldots, 1, 0$ **do**
$\quad \boldsymbol{m}_j^{N,(p)} = \boldsymbol{H}_j^N(0)^{-1}(\boldsymbol{P}_j^{N,(p)} + \boldsymbol{A}_{jj+1}(0)\boldsymbol{m}_{j+1}^{N,(p)})$
end

A.3   SARS-CoV-2 DIPs parameter references

k_{bind} - Ozono *et al.* (2021); Walls *et al.* (2020),

d_V^{wt} - Baccam *et al.* (2006); Baggen *et al.* (2021); Bocharov & Romanyukha (1994),

k_{diss} - Ozono *et al.* (2021); Walls *et al.* (2020),

k_{fuse} - Zhu *et al.* (2020b),

k_{uncoat} - Zhu *et al.* (2020b),

$d_{endosome}^{wt}$ - Baggen *et al.* (2021); Heldt *et al.* (2015),

k_{transl} -Buccitelli & Selbach (2020); Irigoyen *et al.* (2016),

$1/f_{ORF1}$ - Kim *et al.* (2020),

d_{NSP} - Buccitelli & Selbach (2020); Gasteiger *et al.* (2003),

$k_{tr(-)}^{wt}$ - Buccitelli & Selbach (2020),

d_{gRNA}^{wt} - Buccitelli & Selbach (2020); Nelson *et al.* (2008),

$k_{tr(+)}^{wt}$ - Adelman *et al.* (2002),

$k_{complex}^{wt}$ - Chen *et al.* (2013); Klein *et al.* (2020); Spencer & Hiscox (2006); Spencer *et al.* (2008); Zinzula *et al.* (2021),

K_N - Bar-On *et al.* (2020); Cubuk *et al.* (2021); Jack *et al.* (2021),

$1/f_N$ - Viehweger *et al.* (2019),

$1/f_{SP}$ - Viehweger *et al.* (2019),

d_N - Gasteiger *et al.* (2003),

d_{SP} - Gasteiger *et al.* (2003),

n_{SP}^{wt} - Klein *et al.* (2020); Neuman *et al.* (2011); Yao *et al.* (2020),

n_N^{wt} - Klein *et al.* (2020),

K_{Vrel}^{wt} - Bar-On *et al.* (2020); Gordon *et al.* (2020),

k_{assemb}^{wt} - Heldt *et al.* (2015); Shcherbatova *et al.* (2020),

d_{N-gRNA}^{wt} - Buccitelli & Selbach (2020); Nelson *et al.* (2008),

$k_{release}^{wt}$ - Mooney *et al.* (2014); Shcherbatova *et al.* (2020),

$d_{assembled}^{wt}$ - Baggen *et al.* (2021),

d_V^{dip} - Baccam *et al.* (2006); Baggen *et al.* (2021); Bocharov & Romanyukha (1994),

$d_{endosome}^{dip}$ - Baggen *et al.* (2021); Heldt *et al.* (2015),

$k_{tr(-)}^{dip}$ - Buccitelli & Selbach (2020),

d_{gRNA}^{dip} - Buccitelli & Selbach (2020); Nelson *et al.* (2008),

$k_{tr(+)}^{dip}$ - Adelman *et al.* (2002),

$k_{complex}^{dip}$ -Chen *et al.* (2013); Klein *et al.* (2020); Spencer & Hiscox (2006); Spencer *et al.* (2008); Zinzula *et al.* (2021),

n_{SP}^{dip} -Klein *et al.* (2020); Neuman *et al.* (2011); Yao *et al.* (2020),

K_{Vrel}^{dip} - Bar-On *et al.* (2020); Gordon *et al.* (2020),

k_{assemb}^{dip} -Heldt *et al.* (2015); Shcherbatova *et al.* (2020),

d_{N-gRNA}^{dip} - Buccitelli & Selbach (2020); Nelson *et al.* (2008),

$k_{release}^{dip}$ -Mooney *et al.* (2014); Shcherbatova *et al.* (2020),

$d_{assembled}^{dip}$ - Baggen *et al.* (2021).

References

ABDELRAHMAN, Z., LI, M. & WANG, X. (2020). Comparative review of SARS-CoV-2, SARS-CoV, MERS-CoV, and Influenza a respiratory viruses. *Frontiers in immunology*, 2309. 4

ACHARYA, D., LIU, G. & GACK, M.U. (2020). Dysregulation of type I interferon responses in COVID-19. *Nature Reviews Immunology*, **20**, 397–398. 99

ADAKEN, C., SCOTT, J.T., SHARMA, R., GOPAL, R., DICKS, S., NIAZI, S., IJAZ, S., EDWARDS, T., SMITH, C.C., COLE, C.P. *et al.* (2021). Ebola virus antibody decay–stimulation in a high proportion of survivors. *Nature*, **590**, 468–472. 56, 58

ADELMAN, K., LA PORTA, A., SANTANGELO, T.J., LIS, J.T., ROBERTS, J.W. & WANG, M.D. (2002). Single molecule analysis of RNA polymerase elongation reveals uniform kinetic behavior. *Proceedings of the National Academy of Sciences*, **99**, 13538–13543. 211

AGRESTI, A. (2012). *Categorical data analysis*, vol. 792. John Wiley & Sons. 64

AGUILAR-BRETONES, M., WESTERHUIS, B.M., RAADSEN, M.P., DE BRUIN, E., CHANDLER, F.D., OKBA, N.M., HAAGMANS, B.L., LANGERAK, T., ENDEMAN, H., VAN DEN AKKER, J.P. *et al.* (2021). Seasonal coronavirus–specific b cells with limited SARS-CoV-2 cross-reactivity dominate the IgG response in severe COVID-19. *The Journal of clinical investigation*, **131**. 81

AHLQUIST, P., NOUEIRY, A.O., LEE, W.M., KUSHNER, D.B. & DYE, B.T. (2003). Host factors in positive-strand RNA virus genome replication. *Journal of virology*, **77**, 8181–8186. 2

REFERENCES

ALCAMI, A. & KOSZINOWSKI, U.H. (2000). Viral mechanisms of immune evasion. *Trends in microbiology*, **8**, 410–418. 2

ALLEN, L.J. (2007). An introduction to mathematical biology. *Upper Saddle River, New Jersey*. 101

ALLEN, L.J. (2010). *An Introduction to Stochastic Processes with Applications to Biology*. Chapman and Hall/CRC. 14, 22, 24, 115

ALNAJI, F.G. & BROOKE, C.B. (2020). Influenza virus DI particles: Defective interfering or delightfully interesting? *PLoS pathogens*, **16**, e1008436. 152

ANDERSON, D. & BURNHAM, K. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*, **63**, 10. 139, 140

APPANNANAVAR, S.B. & MISHRA, B. (2011). An update on Crimean Congo hemorrhagic fever. *Journal of global infectious diseases*, **3**, 285. 1

AUDSLEY, M.D. & MOSELEY, G.W. (2013). Paramyxovirus evasion of innate immunity: Diverse strategies for common targets. *World journal of virology*, **2**, 57. 97, 98, 100, 103

AYDILLO, T., ROMBAUTS, A., STADLBAUER, D., ASLAM, S., ABELENDA-ALONSO, G., ESCALERA, A., AMANAT, F., JIANG, K., KRAMMER, F., CARRATALA, J. *et al.* (2021). Immunological imprinting of the antibody response in COVID-19 patients. *Nature communications*, **12**, 3781. 70, 73, 91

BACCAM, P., BEAUCHEMIN, C., MACKEN, C.A., HAYDEN, F.G. & PERELSON, A.S. (2006). Kinetics of Influenza a virus infection in humans. *Journal of virology*, **80**, 7590–7599. 210, 211

BAGGEN, J., PERSOONS, L., VANSTREELS, E., JANSEN, S., VAN LOOVEREN, D., BOECKX, B., GEUDENS, V., DE MAN, J., JOCHMANS, D., WAUTERS, J. *et al.* (2021). Genome-wide crispr screening identifies TMEM106B as a proviral host factor for SARS-CoV-2. *Nature Genetics*, **53**, 435–444. 210, 211, 212

BAIZE, S., LEROY, E.M., GEORGES-COURBOT, M.C., CAPRON, M., LANSOUD-SOUKATE, J., DEBRÉ, P., FISHER-HOCH, S.P., McCORMICK, J.B. & GEORGES, A.J. (1999). Defective humoral responses and extensive intravascular apoptosis are associated with fatal outcome in Ebola virus-infected patients. *Nature medicine*, **5**, 423–426. 43, 64

BANGA NDZOUBOUKOU, J.L., ZHANG, Y.d., LEI, Q., LIN, X.s., YAO, Z.j., FU, H., YUAN, L.y. & FAN, X.l. (2021). Human IgM and IgG responses to an inactivated SARS-CoV-2 vaccine. *Current medical science*, **41**, 1081–1086. 85

BAR-ON, Y.M., FLAMHOLZ, A., PHILLIPS, R. & MILO, R. (2020). Science forum: SARS-CoV-2 (COVID-19) by the numbers. *elife*, **9**, e57309. 211

BASLER, C.F., WANG, X., MÜHLBERGER, E., VOLCHKOV, V., PARAGAS, J., KLENK, H.D., GARCÍA-SASTRE, A. & PALESE, P. (2000). The Ebola virus VP35 protein functions as a type I IFN antagonist. *Proceedings of the National Academy of Sciences*, **97**, 12289–12294. xxx, 125, 130, 133, 146

BDEIR, N., ARORA, P., GÄRTNER, S., HOFFMANN, M., REICHL, U., PÖHLMANN, S. & WINKLER, M. (2019). A system for production of defective interfering particles in the absence of infectious Influenza A virus. *PLoS One*, **14**, e0212757. 152, 153, 179

BELL, B.P. (2016). Overview, control strategies, and lessons learned in the CDC response to the 2014–2016 Ebola epidemic. *MMWR supplements*, **65**. 3, 40

BELLAN, S.E., PULLIAM, J.R., DUSHOFF, J. & MEYERS, L.A. (2014). Ebola control: effect of asymptomatic infection and acquired immunity. *The Lancet*, **384**, 1499–1500. 51

BERKE, I.C., LI, Y. & MODIS, Y. (2013). Structural basis of innate immune recognition of viral RNA. *Cellular microbiology*, **15**, 386–394. 2

BOCHAROV, G. & ROMANYUKHA, A. (1994). Mathematical model of antiviral immune response III. Influenza A virus infection. *Journal of Theoretical Biology*, **167**, 323–360. 210, 211

REFERENCES

Boehm, E., Kronig, I., Neher, R.A., Eckerle, I., Vetter, P., Kaiser, L. *et al.* (2021). Novel SARS-CoV-2 variants: the pandemics within the pandemic. *Clinical Microbiology and Infection*, **27**, 1109–1117. 152

Bonilla, F.A. & Oettgen, H.C. (2010). Adaptive immunity. *Journal of Allergy and Clinical Immunology*, **125**, S33–S40. 3

Boon, A., de Mutsert, G., Graus, Y., Fouchier, R., Sintnicolaas, K., Osterhaus, A. & Rimmelzwaan, G. (2002). The magnitude and specificity of Influenza A virus-specific cytotoxic T-lymphocyte responses in humans is related to HLA-A and-B phenotype. *Journal of virology*, **76**, 582–590. 59

Bosworth, A., Dowall, S.D., Garcia-Dorival, I., Rickett, N.Y., Bruce, C.B., Matthews, D.A., Fang, Y., Aljabr, W., Kenny, J., Nelson, C. *et al.* (2017). A comparison of host gene expression signatures associated with infection in vitro by the Makona and Ecran (Mayinga) variants of Ebola virus. *Scientific reports*, **7**, 43144. xxx, 98, 133, 149

Brolin Ribacke, K.J., Saulnier, D.D., Eriksson, A. & Von Schreeb, J. (2016). Effects of the West Africa Ebola virus disease on health-care utilization–a systematic review. *Frontiers in public health*, **4**, 222. 1, 39

Brooke, C.B. (2014). Biological activities of'noninfectious' Influenza A virus particles. *Future virology*, **9**, 41–51. 203

Buccitelli, C. & Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, **21**, 630–644. 211, 212

Campos, G.S., Bandeira, A.C. & Sardi, S.I. (2015). Zika virus outbreak, Bahia, Brazil. *Emerging infectious diseases*, **21**, 1885. 1

Carias, C., O'Hagan, J.J., Gambhir, M., Kahn, E.B., Swerdlow, D.L. & Meltzer, M.I. (2019). Forecasting the 2014 West African Ebola outbreak. *Epidemiologic reviews*, **41**, 34–50. 39

CARROLL, M.W., MATTHEWS, D.A., HISCOX, J.A., ELMORE, M.J., POL-
LAKIS, G., RAMBAUT, A., HEWSON, R., GARCÍA-DORIVAL, I., BORE, J.A.,
KOUNDOUNO, R. *et al.* (2015). Temporal and spatial analysis of the 2014–2015
Ebola virus outbreak in West Africa. *Nature*, **524**, 97. 43, 46, 51, 66

CARRUTHERS, J., LYTHE, G., LÓPEZ-GARCÍA, M., GILLARD, J., LAWS,
T.R., LUKASZEWSKI, R. & MOLINA-PARÍS, C. (2020). Stochastic dynamics
of Francisella tularensis infection and replication. *PLOS Computational Biology*,
16, e1007752. 203

CASTELLS, M.C. & PHILLIPS, E.J. (2021). Maintaining safety with SARS-CoV-2
vaccines. *New England Journal of Medicine*, **384**, 643–649. 152

CASTRO, M. & DE BOER, R.J. (2020). Testing structural identifiability by a
simple scaling method. *PLOS Computational Biology*, **16**, e1008248. 37, 129

CDC, S.C.. (2022). variant classifications and definitions [internet]. CDC. 93

CENTERS FOR DISEASE CONTROL AND PREVENTION (October 2018). Ebola
virus disease EVD. https://www.cdc.gov/vhf/Ebola/pdf/Ebola-factsheet-P.pdf.
6

CHAN, K., CHENG, V., WOO, P., LAU, S., POON, L., GUAN, Y., SETO,
W., YUEN, K. & PEIRIS, J. (2005). Serological responses in patients with
severe acute respiratory syndrome coronavirus infection and cross-reactivity with
human coronaviruses 229E, OC43, and NL63. *Clinical and Vaccine Immunology*,
12, 1317–1321. 85

CHAN, Y.K. & GACK, M.U. (2016). Viral evasion of intracellular DNA and
RNA sensing. *Nature Reviews Microbiology*, **14**, 360–373. 99

CHAPLIN, D.D. (2010). Overview of the immune response. *Journal of allergy and
clinical immunology*, **125**, S3–S23. 1

CHATFIELD, C. (2018). *Introduction to multivariate analysis*. Routledge. 17, 19,
20, 21, 44

REFERENCES

CHATURVEDI, S., VASEN, G., PABLO, M., CHEN, X., BEUTLER, N., KUMAR, A., TANNER, E., ILLOUZ, S., RAHGOSHAY, D., BURNETT, J. *et al.* (2021). Identification of a therapeutic interfering particle— a single-dose SARS-CoV-2 antiviral intervention with a high barrier to resistance. *Cell*, **184**, 6022–6036. xxiii, xxx, 10, 152, 153, 156, 157, 166, 168, 179

CHEN, I.J., YUANN, J.M.P., CHANG, Y.M., LIN, S.Y., ZHAO, J., PERLMAN, S., SHEN, Y.Y., HUANG, T.H. & HOU, M.H. (2013). Crystal structure-based exploration of the important role of Arg106 in the RNA-binding domain of human coronavirus OC43 nucleocapsid protein. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, **1834**, 1054–1062. 211

CHIS, O.T., BANGA, J.R. & BALSA-CANTO, E. (2011). Structural identifiability of systems biology models: a critical comparison of methods. *PloS one*, **6**, e27755. 36

CHO, A., MUECKSCH, F., SCHAEFER-BABAJEW, D., WANG, Z., FINKIN, S., GAEBLER, C., RAMOS, V., CIPOLLA, M., MENDOZA, P., AGUDELO, M. *et al.* (2021). Anti-SARS-CoV-2 receptor-binding domain antibody evolution after mRNA vaccination. *Nature*, **600**, 517–522. 89, 90, 92

COBEY, S. & HENSLEY, S.E. (2017). Immune history and Influenza virus susceptibility. *Current opinion in virology*, **22**, 105–111. 90

COLAVITA, F., BIAVA, M., CASTILLETTI, C., LANINI, S., MICCIO, R., PORTELLA, G., VAIRO, F., IPPOLITO, G., CAPOBIANCHI, M.R., DI CARO, A. *et al.* (2019). Inflammatory and humoral immune response during Ebola virus infection in survivor and fatal cases occurred in Sierra Leone during the 2014–2016 outbreak in West Africa. *Viruses*, **11**, 373. 43, 68

COLSON, P., ROLAIN, J.M. & RAOULT, D. (2020). Chloroquine for the 2019 novel coronavirus SARS-CoV-2. 152

CREECH, C.B., WALKER, S.C. & SAMUELS, R.J. (2021). SARS-CoV-2 vaccines. *Jama*, **325**, 1318–1320. 4, 5, 10

CUBUK, J., ALSTON, J.J., INCICCO, J.J., SINGH, S., STUCHELL-BRERETON, M.D., WARD, M.D., ZIMMERMAN, M.I., VITHANI, N., GRIFFITH, D., WAGONER, J.A. *et al.* (2021). The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nature communications*, **12**, 1–17. 211

CUI, L., WANG, H., JI, Y., YANG, J., XU, S., HUANG, X., WANG, Z., QIN, L., TIEN, P., ZHOU, X. *et al.* (2015). The nucleocapsid protein of coronaviruses acts as a viral suppressor of RNA silencing in mammalian cells. *Journal of virology*, **89**, 9029–9043. 146

DAVIS, C.W., JACKSON, K.J., MCELROY, A.K., HALFMANN, P., HUANG, J., CHENNAREDDY, C., PIPER, A.E., LEUNG, Y., ALBARIÑO, C.G., CROZIER, I. *et al.* (2019). Longitudinal analysis of the human B cell response to Ebola virus infection. *Cell*, **177**, 1566–1582. 56, 57, 58, 67

DEMPSEY, P., VAIDYA, S. & CHENG, G. (2003). The art of war: Innate and adaptive immune responses. *Cellular and Molecular Life Sciences CMLS*, **60**, 2604–2621. 180

DIEKMANN, O., HEESTERBEEK, J. & ROBERTS, M.G. (2010). The construction of next-generation matrices for compartmental epidemic models. *Journal of the royal society interface*, **7**, 873–885. 195

DILLEY, K.A., VOORHIES, A.A., LUTHRA, P., PURI, V., STOCKWELL, T.B., LORENZI, H., BASLER, C.F. & SHABMAN, R.S. (2017). The Ebola virus VP35 protein binds viral immunostimulatory and host RNAs identified through deep sequencing. *PloS one*, **12**, e0178717. 125

DONELAN, N.R., BASLER, C.F. & GARCÍA-SASTRE, A. (2003). A recombinant Influenza A virus expressing an RNA-binding-defective NS1 protein induces high levels of beta-interferon and is attenuated in mice. *Journal of virology*, **77**, 13257–13266. 146, 147

DU, Y., CHEN, L. & SHI, Y. (2022). Booster COVID-19 vaccination against the SARS-CoV-2 omicron variant: A systematic review. *Human Vaccines & Immunotherapeutics*, **18**, 2062983. 152

REFERENCES

DUGAN, H.L., STAMPER, C.T., LI, L., CHANGROB, S., ASBY, N.W., HALF-MANN, P.J., ZHENG, N.Y., HUANG, M., SHAW, D.G., COBB, M.S. *et al.* (2021). Profiling B cell immunodominance after SARS-CoV-2 infection reveals antibody evolution to non-neutralizing viral targets. *Immunity*, **54**, 1290–1303. 91

DUGAS, M., GROTE-WESTRICK, T., VOLLENBERG, R., LORENTZEN, E., BRIX, T., SCHMIDT, H., TEPASSE, P.R. & KÜHN, J. (2021). Less severe course of COVID-19 is associated with elevated levels of antibodies against seasonal human coronaviruses OC43 and HKU1 (HCoV OC43, HCoV HKU1). *International Journal of Infectious Diseases*, **105**, 304–306. 70, 73, 91

EDRIDGE, A.W., KACZOROWSKA, J., HOSTE, A.C., BAKKER, M., KLEIN, M., LOENS, K., JEBBINK, M.F., MATSER, A., KINSELLA, C.M., RUEDA, P. *et al.* (2020). Seasonal coronavirus protective immunity is short-lasting. *Nature medicine*, **26**, 1691–1693. 91

EDWARDS, M.R., LIU, G., MIRE, C.E., SURESHCHANDRA, S., LUTHRA, P., YEN, B., SHABMAN, R.S., LEUNG, D.W., MESSAOUDI, I., GEISBERT, T.W. *et al.* (2016). Differential regulation of interferon responses by Ebola and Marburg virus VP35 proteins. *Cell reports*, **14**, 1632–1640. xxx, 130, 133, 146, 147

EFRON, B. (1982). *The jackknife, the bootstrap, and other resampling plans*, vol. 38. Siam. 22

ELLIOTT, R.M. & WEBER, F. (2009). Bunyaviruses and the type I interferon system. *Viruses*, **1**, 1003–1021. 99

ERBAN, R., CHAPMAN, S.J. & MAINI, P.K. (2007). A practical guide to stochastic simulations of reaction-diffusion process. 28

ESCUDERO-PÉREZ, B. & MUÑOZ-FONTELA, C. (2019). Role of type I interferons on filovirus pathogenesis. *Vaccines*, **7**, 22. 97, 98, 100, 103

FATEHI, F., BINGHAM, R.J., DECHANT, P.P., STOCKLEY, P.G. & TWAROCK, R. (2021). Therapeutic interfering particles exploiting viral replication and assembly mechanisms show promising performance: a modelling study. *Scientific reports*, **11**, 1–10. 152, 153

FELDMANN, H. & GEISBERT, T.W. (16th November 2010). Ebola haemorrhagic fever. *Lancet*, **377**, 849–862. 6, 7, 149

FOON, K.A., SHERWIN, S.A., ABRAMS, P.G., STEVENSON, H.C., HOLMES, P., MALUISH, A.E., OLDHAM, R.K. & HERBERMAN, R.B. (1985). A phase I trial of recombinant gamma interferon in patients with cancer. *Cancer Immunology, Immunotherapy*, **20**, 193–197. 40

FU, Y., CHENG, Y. & WU, Y. (2020). Understanding SARS-CoV-2-mediated inflammatory responses: from mechanisms to potential therapeutic tools. *Virologica Sinica*, **35**, 266–271. 152

GARCIA-BELTRAN, W.F., LAM, E.C., ASTUDILLO, M.G., YANG, D., MILLER, T.E., FELDMAN, J., HAUSER, B.M., CARADONNA, T.M., CLAYTON, K.L., NITIDO, A.D. *et al.* (2021). COVID-19-neutralizing antibodies predict disease severity and survival. *Cell*, **184**, 476–488. 81

GASTEIGER, E., GATTIKER, A., HOOGLAND, C., IVANYI, I., APPEL, R.D. & BAIROCH, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic acids research*, **31**, 3784–3788. 211

GAZZINELLI, R., HAKIM, F., HIENY, S., SHEARER, G. & SHER, A. (1991). Synergistic role of CD4+ and CD8+ t lymphocytes in IFN-gamma production and protective immunity induced by an attenuated toxoplasma gondii vaccine. *Journal of immunology (Baltimore, Md.: 1950)*, **146**, 286–292. 59

GEISBERT, T.W., HENSLEY, L.E., GIBB, T.R., STEELE, K.E., JAAX, N.K. & JAHRLING, P.B. (2000). Apoptosis induced in vitro and in vivo during infection by Ebola and Marburg viruses. *Laboratory investigation*, **80**, 171–186. 98

REFERENCES

GEISBERT, T.W., HENSLEY, L.E., LARSEN, T., YOUNG, H.A., REED, D.S., GEISBERT, J.B., SCOTT, D.P., KAGEN, E., JAHRLING, P.B. & DAVIS, K.J. (December 2003). Pathogenesis of Ebola hemorrhagic fever in cynomolgus macaques: Evidence that dendritic cells are early and sustained targets of infection. *The American Journal of Pathology*, **163**, 2346–2360. 6, 8, 149

GEORGE, C.X., LI, Z., OKONSKI, K.M., TOTH, A.M., WANG, Y. & SAMUEL, C.E. (2009). Tipping the balance: antagonism of PKR kinase and ADAR1 deaminase functions by virus gene products. *Journal of Interferon & Cytokine Research*, **29**, 477–487. 113

GIBBENS, J., WILESMITH, J., SHARPE, C., MANSLEY, L., MICHALOPOULOU, E., RYAN, J. & HUDSON, M. (2001). Descriptive epidemiology of the 2001 foot-and-mouth disease epidemic in Great Britain: the first five months. *Veterinary Record*, **149**, 729–743. 182

GILLESPIE, D.T. (2007). Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, **58**, 35–55. 28

GLYNN, J.R., BOWER, H., JOHNSON, S., HOULIHAN, C.F., MONTESANO, C., SCOTT, J.T., SEMPLE, M.G., BANGURA, M.S., KAMARA, A.J., KAMARA, O. *et al.* (2017). Asymptomatic infection and unrecognised Ebola virus disease in Ebola-affected households in Sierra Leone: a cross-sectional study using a new non-invasive assay for antibodies to Ebola virus. *The Lancet infectious diseases*, **17**, 645–653. 52, 54, 67

GOLDSTEIN, T., ANTHONY, S.J., GBAKIMA, A., BIRD, B.H., BANGURA, J., TREMEAU-BRAVARD, A., BELAGANAHALLI, M.N., WELLS, H.L., DHANOTA, J.K., LIANG, E. *et al.* (2018). The discovery of Bombali virus adds further support for bats as hosts of ebolaviruses. *Nature microbiology*, **3**, 1084. 6

GONZÀLEZ-PARRA, G., DE RIDDER, F., HUNTJENS, D., ROYMANS, D., ISPAS, G. & DOBROVOLNY, H.M. (2018). A comparison of RSV and Influenza in vitro kinetic parameters reveals differences in infecting time. *PloS one*, **13**, e0192645. 182, 183, 188, 195, 200, 202

GORDON, D.E., HIATT, J., BOUHADDOU, M., REZELJ, V.V., ULFERTS, S., BRABERG, H., JUREKA, A.S., OBERNIER, K., GUO, J.Z., BATRA, J. *et al.* (2020). Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science*, **370**. 211

GORSE, G.J., DONOVAN, M.M. & PATEL, G.B. (2020). Antibodies to coronaviruses are higher in older compared with younger adults and binding antibodies are more sensitive than neutralizing antibodies in identifying coronavirus-associated illnesses. *Journal of medical virology*, **92**, 512–517. 77

GREBENNIKOV, D., KHOLODAREVA, E., SAZONOV, I., KARSONOVA, A., MEYERHANS, A. & BOCHAROV, G. (2021). Intracellular life cycle kinetics of SARS-CoV-2 predicted using mathematical modelling. *Viruses*, **13**. xxiv, xxx, 10, 153, 161, 166, 168, 172, 179

GREBENNIKOV, D., KARSONOVA, A., LOGUINOVA, M., CASELLA, V., MEYERHANS, A. & BOCHAROV, G. (2022). Predicting the Kinetic Coordination of Immune Response Dynamics in SARS-CoV-2 Infection: Implications for Disease Pathogenesis. *Mathematics*, **10**, 3154. 153

GRUBMAN, M.J. & BAXT, B. (2004). Foot-and-mouth disease. *Clinical microbiology reviews*, **17**, 465–493. 182

GUTHMILLER, J.J. & WILSON, P.C. (2018). Harnessing immune history to combat Influenza viruses. *Current opinion in immunology*, **53**, 187–195. 90

HAYAMA, Y., YAMAMOTO, T., KOBAYASHI, S., MUROGA, N. & TSUTSUI, T. (2013). Mathematical model of the 2010 foot-and-mouth disease epidemic in Japan and evaluation of control measures. *Preventive veterinary medicine*, **112**, 183–193. 182

HELDT, F.S., KUPKE, S.Y., DORL, S., REICHL, U. & FRENSING, T. (2015). Single-cell analysis and stochastic modelling unveil large cell-to-cell variability in Influenza A virus infection. *Nature communications*, **6**, 1–12. 211, 212

REFERENCES

HICKS, J., KLUMPP-THOMAS, C., KALISH, H., SHUNMUGAVEL, A., MEHALKO, J., DENSON, J.P., SNEAD, K.R., DREW, M., CORBETT, K.S., GRAHAM, B.S. *et al.* (2021). Serologic cross-reactivity of SARS-CoV-2 with endemic and seasonal betacoronaviruses. *Journal of clinical immunology*, **41**, 906–913. 70

HO, T.H., KEW, C., LUI, P.Y., CHAN, C.P., SATOH, T., AKIRA, S., JIN, D.Y. & KOK, K.H. (2016). PACT-and RIG-I-dependent activation of type I interferon production by a defective interfering RNA derived from measles virus vaccine. *Journal of virology*, **90**, 1557–1568. xx, 110, 111, 112

HOEBE, K., JANSSEN, E. & BEUTLER, B. (2004). The interface between innate and adaptive immunity. *Nature immunology*, **5**, 971. 2

HOMMA, T. & SALTELLI, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, **52**, 1–17. 125

HONDA, K., YANAI, H., NEGISHI, H., ASAGIRI, M., SATO, M., MIZUTANI, T., SHIMADA, N., OHBA, Y., TAKAOKA, A., YOSHIDA, N. *et al.* (2005). IRF-7 is the master regulator of type-I interferon-dependent immune responses. *Nature*, **434**, 772–777. 148

HONG, H., OVCHINNIKOV, A., POGUDIN, G. & YAP, C. (2020). Global identifiability of differential models. *Communications on Pure and Applied Mathematics*, **73**, 1831–1879. 37

HOWEY, R., BANKOWSKI, B., JULEFF, N., SAVILL, N.J., GIBSON, D., FAZAKERLEY, J., CHARLESTON, B. & WOOLHOUSE, M.E. (2012). Modelling the within-host dynamics of the foot-and-mouth disease virus in cattle. *Epidemics*, **4**, 93–103. 182

ILMER, I., OVCHINNIKOV, A. & POGUDIN, G. (2021). Web-based structural identifiability analyzer. *arXiv preprint arXiv:2106.15066*. 129, 187

IRIGOYEN, N., FIRTH, A.E., JONES, J.D., CHUNG, B.Y.W., SIDDELL, S.G. & BRIERLEY, I. (2016). High-resolution analysis of coronavirus gene expression by RNA sequencing and ribosome profiling. *PLoS pathogens*, **12**, e1005473. 211

IVASHKIV, L.B. & DONLIN, L.T. (2014). Regulation of type I interferon responses. *Nature reviews Immunology*, **14**, 36–49. 97

JACK, A., FERRO, L.S., TRNKA, M.J., WEHRI, E., NADGIR, A., NGUYENLA, X., FOX, D., COSTA, K., STANLEY, S., SCHALETZKY, J. *et al.* (2021). SARS-CoV-2 nucleocapsid protein forms condensates with viral genomic RNA. *PLoS Biology*, **19**, e3001425. 211

JACKSON, B., BONI, M.F., BULL, M.J., COLLERAN, A., COLQUHOUN, R.M., DARBY, A.C., HALDENBY, S., HILL, V., LUCACI, A., MCCRONE, J.T. *et al.* (2021). Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell*, **184**, 5179–5188. 5

JAMESON, S.C. & MASOPUST, D. (2009). Diversity in T cell memory: an embarrassment of riches. *Immunity*, **31**, 859–871. 56

JANEWAY JR, C.A., TRAVERS, P., WALPORT, M. & SHLOMCHIK, M.J. (2001). The distribution and functions of immunoglobulin isotypes. In *Immunobiology: The Immune System in Health and Disease. 5th edition*, Garland Science. 76

JANGRA, S., YE, C., RATHNASINGHE, R., STADLBAUER, D., ALSHAMMARY, H., AMOAKO, A.A., AWAWDA, M.H., BEACH, K.F., BERMÚDEZ-GONZÁLEZ, M.C., CHERNET, R.L. *et al.* (2021). SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. *The Lancet Microbe*, **2**, e283–e284. 152

JOHN HOPKINS UNIVERSITY OF MEDICINE (March 2023). Mortality analyses. 4

KAK, G., RAZA, M. & TIWARI, B.K. (2018). Interferon-gamma (IFN-γ): Exploring its implications in infectious diseases. *Biomolecular concepts*, **9**, 64–79. 40

KATZE, M.G., HE, Y. & GALE, M. (2002). Viruses and interferon: a fight for supremacy. *Nature Reviews Immunology*, **2**, 675–687. 97

KAUFMAN, J.D. & DUNLAP, W.P. (2000). Determining the number of factors to retain: Q windows-based FORTRAN-IMSL program for parallel analysis. *Behavior Research Methods, Instruments, & Computers*, **32**, 389–395. 20

REFERENCES

KHURANA, S., RAVICHANDRAN, S., HAHN, M., COYLE, E.M., STONIER, S.W., ZAK, S.E., KINDRACHUK, J., DAVEY, R.T., DYE, J.M. & CHERTOW, D.S. (2020). Longitudinal human antibody repertoire against complete viral proteome from Ebola virus survivor reveals protective sites for vaccine design. *Cell host & microbe*, **27**, 262–276. 40

KIEPIELA, P., LESLIE, A.J., HONEYBORNE, I., RAMDUTH, D., THOBAK-GALE, C., CHETTY, S., RATHNAVALU, P., MOORE, C., PFAFFEROTT, K.J., HILTON, L. *et al.* (2004). Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature*, **432**, 769–775. 9, 59

KIJIMA, M. (1997). *Markov processes for stochastic modeling*, vol. 6. CRC Press. 14, 107

KIM, D., LEE, J.Y., YANG, J.S., KIM, J.W., KIM, V.N. & CHANG, H. (2020). The architecture of SARS-CoV-2 transcriptome. *Cell*, **181**, 914–921. 211

KIMBERLIN, C.R., BORNHOLDT, Z.A., LI, S., WOODS, V.L., MACRAE, I.J. & SAPHIRE, E.O. (2010). Ebolavirus VP35 uses a bimodal strategy to bind dsRNA for innate immune suppression. *Proceedings of the National Academy of Sciences*, **107**, 314–319. 125

KLEIN, S., CORTESE, M., WINTER, S.L., WACHSMUTH-MELM, M., NEUFELDT, C.J., CERIKAN, B., STANIFER, M.L., BOULANT, S., BARTENSCHLAGER, R. & CHLANDA, P. (2020). SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. *Nature communications*, **11**, 1–10. 211

KORKMAZ, S., GOKSULUK, D. & ZARARSIZ, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal*, **6**, 151–162. 20

KOTLIAR, D., LIN, A.E., LOGUE, J., HUGHES, T.K., KHOURY, N.M., RAJU, S.S., WADSWORTH II, M.H., CHEN, H., KURTZ, J.R., DIGHERO-KEMP, B. *et al.* (2020). Single-cell profiling of Ebola virus disease in vivo reveals viral and host dynamics. *Cell*, **183**, 1383–1401. xii, xxix, 98, 99, 128, 129, 149

LADNER, J.T., HENSON, S.N., BOYLE, A.S., ENGELBREKTSON, A.L., FINK, Z.W., RAHEE, F., D'AMBROZIO, J., SCHAECHER, K.E., STONE, M., DONG, W. *et al.* (2021). Epitope-resolved profiling of the SARS-CoV-2 antibody response identifies cross-reactivity with endemic human coronaviruses. *Cell Reports Medicine*, **2**, 100189. 77

LAMERS, M.M. & HAAGMANS, B.L. (2022). SARS-CoV-2 pathogenesis. *Nature reviews microbiology*, **20**, 270–284. xv, 4, 5, 9

LEROY, E.M., BAIZE, S., VOLCHKOV, V., FISHER-HOCH, S., GEORGES-COURBOT, M., LANSOUD-SOUKATE, J., CAPRON, M., DEBRE, P., GEORGES, A. & MCCORMICK, J. (2000). Human asymptomatic Ebola infection and strong inflammatory response. *The Lancet*, **355**, 2210–2215. 51

LEUNG, L.W., MARTINEZ, O., REYNARD, O., VOLCHKOV, V.E. & BASLER, C.F. (2011). Ebola virus failure to stimulate plasmacytoid dendritic cell interferon responses correlates with impaired cellular entry. *The Journal of infectious diseases*, **204**, S973–S977. 148

LI, D., LIN, M.H., RAWLE, D.J., JIN, H., WU, Z., WANG, L., LOR, M., HUSSAIN, M., AASKOV, J. & HARRICH, D. (2021). Dengue virus-free defective interfering particles have potent and broad anti-Dengue virus activity. *Communications biology*, **4**, 1–11. 179

LI, X., WANG, W., ZHAO, X., ZAI, J., ZHAO, Q., LI, Y. & CHAILLON, A. (2020a). Transmission dynamics and evolutionary history of 2019-nCoV. *Journal of medical virology*, **92**, 501–511. 3, 70, 152

LI, Y.D., CHI, W.Y., SU, J.H., FERRALL, L., HUNG, C.F. & WU, T.C. (2020b). Coronavirus vaccine development: from SARS and MERS to COVID-19. *Journal of biomedical science*, **27**, 1–23. 4

LIAO, L.E., CARRUTHERS, J., SMITHER, S.J., TEAM, C.V., WELLER, S.A., WILLIAMSON, D., LAWS, T.R., GARCÍA-DORIVAL, I., HISCOX, J., HOLDER, B.P. *et al.* (2020). Quantification of Ebola virus replication kinetics in vitro. *PLoS Computational Biology*, **16**, e1008375. 151, 182, 183, 185, 188, 200, 202

REFERENCES

LIN, C.Y., WOLF, J., BRICE, D.C., SUN, Y., LOCKE, M., CHERRY, S., CASTELLAW, A.H., WEHENKEL, M., CRAWFORD, J.C., ZARNITSYNA, V.I. *et al.* (2022). Pre-existing humoral immunity to human common cold coronaviruses negatively impacts the protective SARS-CoV-2 antibody response. *Cell host & microbe*, **30**, 83–96. 70, 72

LIU, D.X., LIANG, J.Q. & FUNG, T.S. (2021). Human coronavirus-229E,-OC43,-NL63, and-HKU1 (Coronaviridae). *Encyclopedia of virology*, 428. 4

LOCKE, M., LYTHE, G., LÓPEZ-GARCÍA, M., MUÑOZ-FONTELA, C., CARROLL, M. & MOLINA-PARÍS, C. (2021). Quantification of type I Interferon Inhibition by Viral Proteins: Ebola Virus as a case study. *Viruses*, **13**, 2441. 97

LOPEZ-GARCIA, M., NOWICKA, M., BENDTSEN, C., LYTHE, G., PONNAMBALAM, S. & MOLINA-PARIS, C. (2016). Stochastic models of the binding kinetics of VEGF-A to VEGFR1 and VEGFR2 in endothelial cells. 106

LOU, Y., LIU, L. & GAO, D. (2017). Modeling co-infection of Ixodes tick-borne pathogens. *Mathematical Biosciences & Engineering*, **14**, 1301. 129, 130, 133

LOUTEN, J. (2016a). Virus replication. *Essential human virology*, 49. 2

LOUTEN, J. (2016b). Virus transmission and epidemiology. *Essential human virology*, 71. 1

LU, C., XU, H., RANJITH-KUMAR, C., BROOKS, M.T., HOU, T.Y., HU, F., HERR, A.B., STRONG, R.K., KAO, C.C. & LI, P. (2010). The structural basis of 5 triphosphate double-stranded RNA recognition by RIG-I C-terminal domain. *Structure*, **18**, 1032–1043. xxx, 130, 133

LUTHRA, P., RAMANAN, P., MIRE, C.E., WEISEND, C., TSUDA, Y., YEN, B., LIU, G., LEUNG, D.W., GEISBERT, T.W., EBIHARA, H. *et al.* (2013). Mutual antagonism between the Ebola virus VP35 protein and the RIG-I activator PACT determines infection outcome. *Cell host & microbe*, **14**, 74–84. xx, 110, 111, 112

MacNeil, A., Reed, Z. & Rollin, P.E. (2011). Serologic cross-reactivity of human IgM and IgG antibodies to five species of Ebola virus. *PLoS neglected tropical diseases*, **5**, e1175. 40

Manly, B.F. & Alberto, J.A.N. (2016). *Multivariate statistical methods: a primer*. Chapman and Hall/CRC. 17, 20, 21

Mardia, K.V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics, Series B*, 115–128. 20

Martellucci, C.A., Flacco, M.E., Cappadona, R., Bravi, F., Mantovani, L. & Manzoli, L. (2020). SARS-CoV-2 pandemic: An overview. *Advances in biological regulation*, **77**, 100736. 152

Martinez-Flores, D., Zepeda-Cervantes, J., Cruz-Resendiz, A., Aguirre-Sampieri, S., Sampieri, A. & Vaca, L. (2021). SARS-CoV-2 vaccines based on the spike glycoprotein and implications of new viral variants. *Frontiers in immunology*, **12**, 701501. 85

Mate, S.E., Kugelman, J.R., Nyenswah, T.G., Ladner, J.T., Wiley, M.R., Cordier-Lassalle, T., Christie, A., Schroth, G.P., Gross, S.M., Davies-Wayne, G.J. *et al.* (2015). Molecular evidence of sexual transmission of Ebola virus. *New England Journal of Medicine*, **373**, 2448–2454. 149

McAllister, C.S., Toth, A.M., Zhang, P., Devaux, P., Cattaneo, R. & Samuel, C.E. (2010). Mechanisms of protein kinase PKR-mediated amplification of beta interferon induction by C protein-deficient measles virus. *Journal of virology*, **84**, 380–386. 113

McAllister, C.S., Taghavi, N. & Samuel, C.E. (2012). Protein kinase PKR amplification of interferon β induction occurs through initiation factor eIF-2α-mediated translational control. *Journal of Biological Chemistry*, **287**, 36384–36392. 113

REFERENCES

McFadden, G., Mohamed, M.R., Rahman, M.M. & Bartee, E. (2009). Cytokine determinants of viral tropism. *Nature Reviews Immunology*, **9**, 645–655. 2

McLean, G., Kamil, J., Lee, B., Moore, P., Schulz, T.F., Muik, A., Sahin, U., Türeci, Ö. & Pather, S. (2022). The impact of evolving SARS-CoV-2 mutations and variants on COVID-19 vaccines. *Mbio*, **13**, e02979–21. 5

Medzhitov, R. & Janeway Jr, C. (2000). Innate immunity. *New England Journal of Medicine*, **343**, 338–344. 2

Mendonça, L., Howe, A., Gilchrist, J.B., Sheng, Y., Sun, D., Knight, M.L., Zanetti-Domingues, L.C., Bateman, B., Krebs, A.S., Chen, L. *et al.* (2021). Correlative multi-scale cryo-imaging unveils SARS-CoV-2 assembly and egress. *Nature Communications*, **12**, 1–10. 159

Messaoudi, I., Amarasinghe, G.K. & Basler, C.F. (2015). Filovirus pathogenesis and immune evasion: insights from Ebola virus and Marburg virus. *Nature Reviews Microbiology*, **13**, 663–676. 98

Messaoudi, I., Amarasinghe, G.K. & Basler, C.F. (6th October 2015). Filovirus pathogenesis and immune evasion: insights from Ebola virus and Marburg virus. *Nature Reviews Microbiology*, **13**, 663–676. xv, 7, 8

Metcalf, C.J.E., Morris, D.H. & Park, S.W. (2020). Mathematical models to guide pandemic response. *Science*, **369**, 368–369. 151

Mohapatra, S., Priyanka, V., Mohapatra, S., Kohli, I. & Mishra, R.K. (2020). Impact of corona virus COVID-19 on the global economy. *Int. J. Agric. Stat. Sci*, **16**, 771–778. 152

Mooney, J., Thakur, S., Kahng, P., Trapani, J.G. & Poccia, D. (2014). Quantification of exocytosis kinetics by DIC image analysis of cortical lawns. *Journal of chemical biology*, **7**, 43–55. 211, 212

MORSE, J.S., LALONDE, T., XU, S. & LIU, W.R. (2020). Learning from the past: possible urgent prevention and treatment options for severe acute respiratory infections caused by 2019-nCoV. *Chembiochem*, **21**, 730–738. 152

MSEMBURI, W., KARLINSKY, A., KNUTSON, V., ALESHIN-GUENDEL, S., CHATTERJI, S. & WAKEFIELD, J. (2023). The WHO estimates of excess mortality associated with the COVID-19 pandemic. *Nature*, **613**, 130–137. 4

MUSHAYABASA, S., BHUNU, C.P. & DHLAMINI, M. (2011). Impact of vaccination and culling on controlling foot and mouth disease: a mathematical modelling approach. 182

NAQVI, A.A.T., FATIMA, K., MOHAMMAD, T., FATIMA, U., SINGH, I.K., SINGH, A., ATIF, S.M., HARIPRASAD, G., HASAN, G.M. & HASSAN, M.I. (2020). Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, **1866**, 165878. 4, 152

NELSON, D.L., LEHNINGER, A.L. & COX, M.M. (2008). *Lehninger principles of biochemistry*. Macmillan. 211, 212

NEUMAN, B.W., KISS, G., KUNDING, A.H., BHELLA, D., BAKSH, M.F., CONNELLY, S., DROESE, B., KLAUS, J.P., MAKINO, S., SAWICKI, S.G. *et al.* (2011). A structural analysis of M protein in coronavirus assembly and morphology. *Journal of structural biology*, **174**, 11–22. 211

NG, K.W., FAULKNER, N., CORNISH, G.H., ROSA, A., HARVEY, R., HUSSAIN, S., ULFERTS, R., EARL, C., WROBEL, A.G., BENTON, D.J. *et al.* (2020). Preexisting and de novo humoral immunity to SARS-CoV-2 in humans. *Science*, **370**, 1339–1343. 70, 77, 81

OLIVAL, K.J., CRYAN, P.M., AMMAN, B.R., BARIC, R.S., BLEHERT, D.S., BROOK, C.E., CALISHER, C.H., CASTLE, K.T., COLEMAN, J.T., DASZAK, P. *et al.* (2020). Possibility for reverse zoonotic transmission of SARS-CoV-2 to free-ranging wildlife: A case study of bats. *PLoS pathogens*, **16**, e1008758. 149

REFERENCES

Ozono, S., Zhang, Y., Ode, H., Sano, K., Tan, T.S., Imai, K., Miyoshi, K., Kishigami, S., Ueno, T., Iwatani, Y. *et al.* (2021). SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. *Nature communications*, **12**, 1–9. 210, 211

Padma, G. & Vijayalakshmi, C. (2011). An analysis of continuous time Markov chains using generator matrices. *International Journal of Computer Applications*, **35**, 20–24. 22, 24

Paweska, J.T., Storm, N., Grobbelaar, A.A., Markotter, W., Kemp, A. & Jansen van Vuren, P. (2016). Experimental inoculation of Egyptian fruit bats (Rousettus aegyptiacus) with Ebola virus. *Viruses*, **8**, 29. 149

Peralta, A.F. & Toral, R. (2018). System-size expansion of the moments of a master equation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **28**, 106303. 117

Perelson, A.S. (1999). Viral kinetics and mathematical models. *The American journal of medicine*, **107**, 49–52. 151

Perelson, A.S. (2002). Modelling viral and immune system dynamics. *Nature reviews immunology*, **2**, 28–36. 148

Perelson, A.S. & Ke, R. (2021). Mechanistic modeling of SARS-CoV-2 and other infectious diseases and the effects of therapeutics. *Clinical Pharmacology & Therapeutics*, **109**, 829–840. 153

Perez-Martin, E., Beechler, B., Zhang, F., Scott, K., de Klerk-Lorist, L.M., Limon, G., Dugovich, B., Gubbins, S., Botha, A., Hetem, R. *et al.* (2022). Viral dynamics and immune responses to foot-and-mouth disease virus in African buffalo (syncerus caffer). *Veterinary Research*, **53**, 1–14. 191, 196, 202

Perry, A.K., Chen, G., Zheng, D., Tang, H. & Cheng, G. (June 2005). *The Host Type I Interferon Response to Viral and Bacterial Infections*, vol. 15. Cell Research. 97

PIESMAN, J. & EISEN, L. (2008). Prevention of tick-borne diseases. *Annu. Rev. Entomol.*, **53**, 323–343. 2

PINILLA, J., BARBER, P., VALLEJO-TORRES, L., RODRÍGUEZ-MIRELES, S., LÓPEZ-VALCÁRCEL, B.G. & SERRA-MAJEM, L. (2021). The economic impact of the SARS-CoV-2 (COVID-19) pandemic in Spain. *International journal of environmental research and public health*, **18**, 4708. 4

POSTON, D., WEISBLUM, Y., WISE, H., TEMPLETON, K., JENKS, S., HATZI-IOANNOU, T. & BIENIASZ, P. (2021). Absence of severe acute respiratory syndrome coronavirus 2 neutralizing activity in prepandemic sera from individuals with recent seasonal coronavirus infection. *Clinical Infectious Diseases*, **73**, e1208–e1211. 73

RAGONNET-CRONIN, M., NUTALAI, R., HUO, J., DIJOKAITE-GURALIUC, A., DAS, R., TUEKPRAKHON, A., SUPASA, P., LIU, C., SELVARAJ, M., GROVES, N. *et al.* (2023). Generation of SARS-CoV-2 escape mutations by monoclonal antibody therapy. *Nature Communications*, **14**, 3334. 5

RAND, U., RINAS, M., SCHWERK, J., NÖHREN, G., LINNES, M., KRÖGER, A., FLOSSDORF, M., KÁLY-KULLAI, K., HAUSER, H., HÖFER, T. *et al.* (2012). Multi-layered stochasticity and paracrine signal propagation shape the type-I interferon response. *Molecular systems biology*, **8**, 584. 99, 146

RAND, U., KUPKE, S.Y., SHKARLET, H., HEIN, M.D., HIRSCH, T., MARICHAL-GALLARDO, P., CICIN-SAIN, L., REICHL, U. & BRUDER, D. (2021). Antiviral activity of Influenza A virus defective interfering particles against SARS-CoV-2 replication in vitro through stimulation of innate immunity. *bioRxiv*. 179, 181

RANDALL, R.E. & GOODBOURN, S. (2008). Interferons and viruses: an interplay between induction, signalling, antiviral responses and virus countermeasures. *Journal of general virology*, **89**, 1–47. 97, 147

REUTER, G. (1961). Competition processes. In *Proc. 4th Berkeley Symp. Math. Statist. Prob*, vol. 2, 421–430. 99

REFERENCES

Rhein, B.A., Powers, L.S., Rogers, K., Anantpadma, M., Singh, B.K., Sakurai, Y., Bair, T., Miller-Hunt, C., Sinn, P., Davey, R.A. *et al.* (2015). Interferon-γ inhibits Ebola virus infection. *PLoS pathogens*, **11**, e1005263. 40

Rimoin, A.W., Lu, K., Bramble, M.S., Steffen, I., Doshi, R.H., Hoff, N.A., Mukadi, P., Nicholson, B.P., Alfonso, V.H., Olinger, G. *et al.* (2018). Ebola virus neutralizing antibodies detectable in survivors of Theyambuku, Zaire outbreak 40 years after infection. *The Journal of infectious diseases*, **217**, 223–231. 46, 58, 67

Rinas, M. (2016). *Data-driven modeling of the dynamic competition between virus infection and the antiviral interferon response*. Ph.D. thesis. 99, 146

Ross, S. (2014). *A first course in probability*. Pearson. 14, 29

Roux, L., Simon, A.E. & Holland, J.J. (1991). Effects of defective interfering viruses on virus replication and pathogenesis in vitro and in vivo. *Advances in virus research*, **40**, 181–211. 179

Rouzine, I.M. & Weinberger, L.S. (2013). Design requirements for interfering particles to maintain coadaptive stability with HIV-1. *Journal of virology*, **87**, 2081–2093. 153

Sanghera, J., Pattani, N., Hashmi, Y., Varley, K.F., Cheruvu, M.S., Bradley, A. & Burke, J.R. (2020). The impact of SARS-CoV-2 on the mental health of healthcare workers in a hospital setting- A systematic review. *Journal of occupational health*, **62**, e12175. 3

Sarry, M., Vitour, D., Zientara, S., Bakkali Kassimi, L. & Blaise-Boisseau, S. (2022). Foot-and-mouth disease virus: molecular interplays with IFN response and the importance of the model. *Viruses*, **14**, 2129. 203

Sazonov, I., Grebennikov, D., Meyerhans, A. & Bocharov, G. (2022). Sensitivity of SARS-CoV-2 Life Cycle to IFN Effects and ACE2 Binding Unveiled with a Stochastic Model. *Viruses*, **14**, 403. 171

SCHMID, B., RINAS, M., RUGGIERI, A., ACOSTA, E.G., BARTENSCHLAGER, M., REUTER, A., FISCHL, W., HARDER, N., BERGEEST, J.P., FLOSSDORF, M. *et al.* (2015). Live cell analysis and mathematical modeling identify determinants of attenuation of Dengue virus 2'-O-methylation mutant. *PLoS pathogens*, **11**, e1005345. 99, 146

SCHULTZE, J.L. & ASCHENBRENNER, A.C. (2021). COVID-19 and the human innate immune system. *Cell*. 99

SCHÜMANN, M., GANTKE, T. & MÜHLBERGER, E. (2009). Ebola virus VP35 antagonizes PKR activity through its C-terminal interferon inhibitory domain. *Journal of virology*, **83**, 8993–8997. 113

SEALY, R.E. & HURWITZ, J.L. (2021). Cross-reactive immune responses toward the common cold human coronaviruses and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): Mini-review and a murine study. *Microorganisms*, **9**, 1643. 70, 73

SHCHERBATOVA, O., GREBENNIKOV, D., SAZONOV, I., MEYERHANS, A. & BOCHAROV, G. (2020). Modeling of the HIV-1 life cycle in productively infected cells to predict novel therapeutic targets. *Pathogens*, **9**, 255. 211, 212

SHROCK, E., FUJIMURA, E., KULA, T., TIMMS, R.T., LEE, I.H., LENG, Y., ROBINSON, M.L., SIE, B.M., LI, M.Z., CHEN, Y. *et al.* (2020). Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. *Science*, **370**, eabd4250. 70, 73, 91

SIMON-LORIERE, E., FAYE, O., FAYE, O., KOIVOGUI, L., MAGASSOUBA, N., KEITA, S., THIBERGE, J.M., DIANCOURT, L., BOUCHIER, C., VANDENBOGAERT, M. *et al.* (2015). Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature*, **524**, 102–104. 43, 48, 51, 66

SINGH, J., PANDIT, P., MCARTHUR, A.G., BANERJEE, A. & MOSSMAN, K. (2021). Evolutionary trajectory of SARS-CoV-2 and emerging variants. *Virology journal*, **18**, 1–21. 5

REFERENCES

SIU, K.L., YEUNG, M.L., KOK, K.H., YUEN, K.S., KEW, C., LUI, P.Y., CHAN, C.P., TSE, H., WOO, P.C., YUEN, K.Y. *et al.* (2014). Middle east respiratory syndrome coronavirus 4a protein is a double-stranded RNA-binding protein that suppresses PACT-induced activation of RIG-I and MDA5 in the innate antiviral response. *Journal of virology*, **88**, 4866–4876. 112

SMITHER, S.J., GARCIA-DORIVAL, I., EASTAUGH, L., FINDLAY, J.S., O'BRIEN, L.M., CARRUTHERS, J., WILLIAMSON, E.D., MOLINA-PARÍS, C., HISCOX, J.A. & LAWS, T.R. (2020). An investigation of the effect of transfected defective, Ebola virus genomes on Ebola replication. *Frontiers in cellular and infection microbiology*, **10**, 159. 153

SOBOL, I.M. (1993). Sensitivity analysis for non-linear mathematical models. *Mathematical modelling and computational experiment*, **1**, 407–414. 34, 35, 36

SPENCER, K.A. & HISCOX, J.A. (2006). Characterisation of the RNA binding properties of the coronavirus infectious bronchitis virus nucleocapsid protein amino-terminal region. *FEBS letters*, **580**, 5993–5998. 211

SPENCER, K.A., DEE, M., BRITTON, P. & HISCOX, J.A. (2008). Role of phosphorylation clusters in the biology of the coronavirus infectious bronchitis virus nucleocapsid protein. *Virology*, **370**, 373–381. 211

SPERANZA, E., RUIBAL, P., PORT, J.R., FENG, F., BURKHARDT, L., GRUND-HOFF, A., GÜNTHER, S., OESTEREICH, L., HISCOX, J.A., CONNOR, J.H. *et al.* (2018). T-cell receptor diversity and the control of T-cell homeostasis mark Ebola virus disease survival in humans. *The Journal of infectious diseases*, **218**, S508–S518. 64, 68

STICH, M., BENNING, L., SPEER, C., GARBADE, S.F., BARTENSCHLAGER, M., KIM, H., GLEICH, F., JELTSCH, K., HAASE, B., JANDA, A. *et al.* (2022). Waning immunity 14 months after SARS-CoV-2 infection. *Pediatrics*, **150**. 40

SWAIN, S.L. (1983). T cell subsets and the recognition of MHC class. *Immunological reviews*, **74**, 129–142. 3

Székely, G.J., Rizzo, M.L. *et al.* (2004). Testing for equal distributions in high dimension. *InterStat*, **5**, 1249–1272. 22

Tawaratsumida, K., Phan, V., Hrincius, E.R., High, A.A., Webby, R., Redecke, V. & Häcker, H. (2014). Quantitative proteomic analysis of the Influenza A virus nonstructural proteins NS1 and NS2 during natural cell infection identifies PACT as an NS1 target protein and antiviral host factor. *Journal of virology*, **88**, 9038–9048. 112

tenOever, B. (2016). The Evolution of Antiviral Defense Systems. *Cell Host & Microbe*, **19**, 142–149. 180

Thom, R., Tipton, T., Strecker, T., Hall, Y., Bore, J.A., Maes, P., Koundouno, F.R., Fehling, S.K., Krähling, V., Steeds, K. *et al.* (2021). Longitudinal antibody and T cell responses in Ebola virus disease survivors and contacts: an observational cohort study. *The Lancet Infectious Diseases*, **21**, 507–516. 40, 41, 43, 59, 60, 67

Thompson, D., Muriel, P., Russell, D., Osborne, P., Bromley, A., Rowland, M., Creigh-Tyte, S. & Brown, C. (2002). Economic costs of the foot and mouth disease outbreak in the United Kingdom in 2001. *Revue scientifique et technique (International Office of Epizootics)*, **21**, 675–687. 182

Timothy, J., Hall, Y., Akoi-Bore, J., Diallo, B., Tipton, T., Bower, H., Glynn, J. & Carroll, M. (2019). Seroprevalence study of anti-ebolavirus (ebov) immunoglobulin g (IgG) at the index site of 2013-2016 West African EBOV outbreak: Insight into early transmission and case fatality rate. In *TRANSACTIONS OF THE ROYAL SOCIETY OF TROPICAL MEDICINE AND HYGIENE*, vol. 113, S48–S48, OXFORD UNIV PRESS GREAT CLARENDON ST, OXFORD OX2 6DP, ENGLAND. 40, 41

Tipton, T., Hall, Y., Bore, J., White, A., Sibley, L., Sarfas, C., Yuki, Y., Martin, M., Longet, S., Mellors, J. *et al.* (2021). Characterisation of the T-cell response to Ebola virus glycoprotein amongst survivors of the 2013–16 West Africa epidemic. *Nature Communications*, **12**, 1153. 42

REFERENCES

TOBÍAS, A. (2020). Evaluation of the lockdowns for the SARS-CoV-2 epidemic in Italy and Spain after one month follow up. *Science of the Total Environment*, **725**, 138539. 3, 70

TOLLE, M.A. (2009). Mosquito-borne diseases. *Current problems in pediatric and adolescent health care*, **39**, 97–140. 2

TONI, T., WELCH, D., STRELKOWA, N., IPSEN, A. & STUMPF, M.P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6**, 187–202. 31, 32, 33, 34, 130

TRINCHIERI, G. (2010). Type I interferon: friend of foe? **207**, 2053–2063. 97

VAHERI, A., STRANDIN, T., HEPOJOKI, J., SIRONEN, T., HENTTONEN, H., MÄKELÄ, S. & MUSTONEN, J. (2013). Uncovering the mysteries of hantavirus infections. *Nature Reviews Microbiology*, **11**, 539–550. 99

VAN DEN DRIESSCHE, P. (2017). Reproduction numbers of infectious disease models. *Infectious Disease Modelling*, **2**, 288–303. 195

VAN KAMPEN, N.G. (1992). *Stochastic processes in physics and chemistry*, vol. 1. Elsevier. 117

VIEHWEGER, A., KRAUTWURST, S., LAMKIEWICZ, K., MADHUGIRI, R., ZIEBUHR, J., HÖLZER, M. & MARZ, M. (2019). Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome research*, **29**, 1545–1554. 211

VILLINGER, F., ROLLIN, P.E., BRAR, S.S., CHIKKALA, N.F., WINTER, J., SUNDSTROM, J.B., ZAKI, S.R., SWANEPOEL, R., ANSARI, A.A. & PETERS, C.J. (1999). Markedly elevated levels of interferon (IFN)-γ, IFN-α, interleukin (IL)-2, IL -10, and tumor necrosis factor-α associated with fatal Ebola virus infection. *The Journal of infectious diseases*, **179**, S188–S191. 40

VORA, S.M., LIEBERMAN, J. & WU, H. (2021). Inflammasome activation at the crux of severe COVID-19. *Nature Reviews Immunology*, **21**, 694–703. 5, 6

V'KOVSKI, P., KRATZEL, A., STEINER, S., STALDER, H. & THIEL, V. (2021). Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology*, **19**, 155–170. 159

WAINWRIGHT, H.M., FINSTERLE, S., JUNG, Y., ZHOU, Q. & BIRKHOLZER, J.T. (2014). Making sense of global sensitivity analyses. *Computers & Geosciences*, **65**, 84–94. 125, 182

WALLS, A.C., PARK, Y.J., TORTORICI, M.A., WALL, A., McGUIRE, A.T. & VEESLER, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, **181**, 281–292. 210, 211

WANG, M.Y., ZHAO, R., GAO, L.J., GAO, X.F., WANG, D.P. & CAO, J.M. (2020). SARS-CoV-2: structure, biology, and structure-based therapeutics development. *Frontiers in cellular and infection microbiology*, **10**, 587269. 5

WANG, R., CHEN, J. & WEI, G.W. (2021). Mechanisms of SARS-CoV-2 evolution revealing vaccine-resistant mutations in Europe and America. *The journal of physical chemistry letters*, **12**, 11850–11857. 5

WEBER, F., BRIDGEN, A., FAZAKERLEY, J.K., STREITENFELD, H., KESSLER, N., RANDALL, R.E. & ELLIOTT, R.M. (2002). Bunyamwera bunyavirus nonstructural protein NSs counteracts the induction of alpha/beta interferon. *Journal of virology*, **76**, 7949–7955. 98, 100

WEINBERG, P.D., HOUNSHELL, J., SHERMAN, L.A., GODWIN, J., ALI, S., TOMORI, C. & BENNETT, C.L. (2002). Legal, financial, and public health consequences of HIV contamination of blood and blood products in the 1980s and 1990s. *Annals of internal medicine*, **136**, 312–319. 2

WHO (2015). Blueprint for R&D preparedness and response to public health emergencies due to highly infectious pathogens. In *Workshop on prioritization of pathogens*. 1

REFERENCES

WILLIAMS, B., LÓPEZ-GARCÍA, M., GILLARD, J.J., LAWS, T.R., LYTHE, G., CARRUTHERS, J., FINNIE, T. & MOLINA-PARÍS, C. (2021). A stochastic intracellular model of anthrax infection with spore germination heterogeneity. *Frontiers in Immunology*, **12**, 688257. 151, 203

WORLD HEALTH ORGANIZATION (May 2019). Ebola virus disease EVD. https://www.who.int/news-room/fact-sheets/detail/Ebola-virus-disease. 6

WORLD HEALTH ORGANIZATION (November 2023). SARS-CoV-2. https://covid19.who.int/. 3

WORLD HEALTH ORGANIZATION *et al.* (2020). Coronavirus disease 2019 (COVID-19): situation report, 73. 3, 152

WU, D., WU, T., LIU, Q. & YANG, Z. (2020). The SARS-CoV-2 outbreak: what we know. *International Journal of Infectious Diseases*, **94**, 44–48. 1, 152

WU, M., ZHOU, E., SHENG, R., FU, X., LI, J., JIANG, C. & SU, W. (2022). Defective interfering particles of influenza virus and their characteristics, impacts, and use in vaccines and antiviral strategies: A systematic review. *Viruses*, **14**, 2773. 10

XIA, W., LI, M., WANG, Y., KAZIS, L.E., BERLO, K., MELIKECHI, N. & CHIKLIS, G.R. (2021). Longitudinal analysis of antibody decay in convalescent COVID-19 patients. *Scientific reports*, **11**, 16796. 56

YANG, H.M. (2014). The basic reproduction number obtained from Jacobian and next generation matrices–a case study of Dengue transmission modelling. *Biosystems*, **126**, 52–75. 182

YANG, P.C., CHU, R.M., CHUNG, W.B. & SUNG, H. (1999). Epidemiological characteristics and financial costs of the 1997 foot-and-mouth disease epidemic in Taiwan. *Veterinary Record*, **145**, 731–734. 182

YAO, H., SONG, Y., CHEN, Y., WU, N., XU, J., SUN, C., ZHANG, J., WENG, T., ZHANG, Z., WU, Z. *et al.* (2020). Molecular architecture of the SARS-CoV-2 virus. *Cell*, **183**, 730–738. 211

Yuen, C.K., Lam, J.Y., Wong, W.M., Mak, L.F., Wang, X., Chu, H., Cai, J.P., Jin, D.Y., To, K.K.W., Chan, J.F.W. *et al.* (2020). SARS-CoV-2 NSP13, NSP14, NSP15 and ORF6 function as potent interferon antagonists. *Emerging microbes & infections*, **9**, 1418–1428. 5

Zhang, W. & Zou, X. (2013). Systematic analysis of the mechanisms of virus-triggered type I IFN signaling pathways through mathematical modeling. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **10**, 771–779. 133

Zhang, X.Y., Trame, M.N., Lesko L, J. & Schmidt, S. (25th February 2015). Sobol sensitivity analysis: A tool to guide the development and evaluation of systems pharmacology models. *CPT: Pharmacometrics and Systems Pharmacology*, 69–79. 35, 36

Zhang, Y.Z. & Holmes, E.C. (2020). A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell*, **181**, 223–227. 4

Zhao, Y. & Xing, Y. (2021). A delayed dynamical model for COVID-19 therapy with defective interfering particles and artificial antibodies. *Discrete & Continuous Dynamical Systems-B*. 153

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R. *et al.* (2020a). A novel coronavirus from patients with pneumonia in China, 2019. *New England journal of medicine*. 3

Zhu, Y., Yu, D., Yan, H., Chong, H. & He, Y. (2020b). Design of potent membrane fusion inhibitors against SARS-CoV-2, an emerging coronavirus with high fusogenic activity. *Journal of virology*, **94**, e00635–20. 211

Zinzula, L. & Tramontano, E. (2013). Strategies of highly pathogenic RNA viruses to block dsRNA detection by RIG-I-like receptors: hide, mask, hit. *Antiviral research*, **100**, 615–635. 3, 146

Zinzula, L., Basquin, J., Bohn, S., Beck, F., Klumpe, S., Pfeifer, G., Nagy, I., Bracher, A., Hartl, F.U. & Baumeister, W. (2021). High-resolution structure and biophysical characterization of the nucleocapsid phosphoprotein dimerization domain from the COVID-19 severe acute respiratory

syndrome coronavirus 2. *Biochemical and biophysical research communications*, **538**, 54–62. 211

Zou, X., Xiang, X., Chen, Y., Peng, T., Luo, X. & Pan, Z. (2010). Understanding inhibition of viral proteins on type I IFN signaling pathways with modeling and optimization. *Journal of theoretical biology*, **265**, 691–703. xxx, 99, 133, 146, 147