Network models of T cell receptor repertoires, cross-reactivity, and viral infection

Daniel Felipe Luque Duque

Department of Applied Mathematics

University of Leeds

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

January, 2023

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Joint publications

The majority of the work in Chapter 4 has been submitted for publication, and refereed, as follows:

– **Daniel Luque Duque**, Martín López García, Jessica A. Gaevert, Grant Lythe, Paul G. Thomas, and Carmen Molina-París. Multivariate model of T cell clonotype competition and homeostasis. *Submitted to: Scientific Reports*, 2022.

The stochastic modelling and analysis presented in this Thesis and the manuscript are the work of Daniel Felipe Luque Duque.

Code availability: Codes used in Chapter 4 have been made publicly available (Luque, 2022a).

Part of the work in Chapter 5 is in preparation to be submitted, as follows:

– **Daniel Luque Duque**, Jessica Ann Gaevert, Grant Lythe, Carmen Molina-París, and Paul Thomas. Untitled manuscript on IAV heterologous infection on mice. *The Journal of Immunology, to be submitted.*

The statistical data analysis and modelling presented in this Thesis and the manuscript are the work of Daniel Felipe Luque Duque. The data used for analysis was obtained from experiments performed by Paul Thomas and Jessica Gaevert from St. Jude Children's Research Hospital.

Code availability: Codes used in Chapter 5 have been made publicly available (Luque, 2022b).

Additionally, part of the work in Chapters 5, and 6 has been refereed and published, as follows:

– Jessica Ann Gaevert‡, **Daniel Luque Duque**‡, Grant Lythe, Carmen Molina-París, and Paul Glyndwr Thomas. Quantifying T Cell Cross-Reactivity: Influenza and Coronaviruses. *Viruses*, volume 13, page 1786. Multidisciplinary Digital Publishing Institute, 2021. ‡: These authors contributed equally.

The analysis of the random network generation algorithms, the definition of the competition process, and the simulations presented in this Thesis and the publication are the work of Daniel Felipe Luque Duque. The biological implications of the results were guided by discussions with Paul Thomas and Jessica Gaevert from St. Jude Children's Research Hospital.

Code availability: Codes used in Chapter 6 have been made publicly available (Luque, 2021).

Acknowledgements

This research received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant agreement number 764698.

First, I would like to thank my supervisors Carmen Molina París, Grant Lythe, and Martín López García for their support and guidance throughout my research. Your advice and guidance has shaped me into the mathematician I am today.

From the Leeds Maths-Bio group, I would like to thank Macauley, Lea, Giulia, Bevelynn, Flavia, Polly, Van, Adam, and Marcus for all your help, support, and friendship throughout my PhD.

From St. Jude Children's Research Hospital, I would like to thank Jessica Gaevert, Paul Thomas, and all the members of the Thomas lab. Your help, support, and friendship has meant the world to me.

From the QuanTII network, I would like to thank All the ESRs and PIs for giving me countless hours of joy, both in and out of the office.

Finally, I would like to thank my family and friends, especially my parents, Maria and Horacio, my siblings, Diego and Aura, and my dear friends, Sisco, Brazz, Ito, Meeks, Che, and Pesca, I would not be where I am today if it wasn't for all of you.

Abstract

The mathematical models of this Thesis represent T cell population dynamics in homeostasis and during infection. In particular, T cell cross-reactivity is studied with a bipartite recognition network encoding the epitope recognition profiles of T cell receptors. The behaviour of extinction events is studied using stochastic models. Stochastic and deterministic techniques are used to study the late time behaviour of the system. Statistical methods are used to study immune responses in the context of influenza A virus infection in mice, providing insight into the effects of immunological history and cross-reactivity. Finally, network theoretical tools are used to study the dynamics of cross-reactive immune responses under different hypotheses for the structure of the bipartite recognition network.

Contents

List of Figures

Chapter 1

Introduction

Mathematical modelling in biology can be used as a quantitative tool to generate and test hypotheses of biological systems at all scales. That is, mathematical models are a versatile tool that can be used to study microscopic behaviours, such as cell signalling (Shinar *et al.*, 2007; Wilmes *et al.*, 2021), as well as macroscopic ones, *e.g.*, population dynamics, be it single-cell (Feliciangeli *et al.*, 2022; Zarnitsyna *et al.*, 2016) or multi-cellular organisms (Amador & Gómez-Corral, 2020; Baroyan *et al.*, 1971; Rvachev, 1968). Models can be broadly classified as being stochastic or deterministic. Deterministic models are those which do not consider random, or stochastic, effects in the system, while stochastic models, as their name suggests, are able to capture these random fluctuations.

Deterministic models are frequently described by systems of ordinary differential equations, and they have been used in the context of population growth since, at least, as far back as the 18$^{\text{th}}$ century, with Thomas Malthus' proposition that population growth is proportional to the size of the population (Malthus, 1986). While this initial proposition was not made in the context of mathematical modelling, it is an early description of what would later be known as an exponential growth model (Brauer *et al.*, 2012, Chapter 1). Not long after, however, more complex models for single populations with a carrying capacity were mathematically described and used to model population dynamics; namely the logistic growth model (Verhulst, 1838, 1845), and its multiple variations that where subsequently proposed (Buis, 1991; Kostitzin, 1940; Turner Jr *et al.*, 1969, 1976). Interest for mathematical modelling was not limited to single populations however, and

thus multi-dimensional models were developed to study the dynamics of multiple populations. Some examples include predator-prey population models (Lotka, 1925; Volterra, 1926), and epidemiological compartment models (Allen, 1994).

Stochastic models, on the other hand, are defined by random variables over a given state space, where each individual random variable describes the evolution of a population over time (Allen, 2010). This type of model can be used to describe direct competition between populations (Iglehart, 1964), as well as competition between populations for a shared survival resource (Stirk *et al.*, 2008, 2010). A common method to study stochastic models, is to simulate them using an exact stochastic simulation algorithm, such as the one proposed by Gillespie (1977). In general, biological systems exhibit some degree of stochasticity in their dynamics, which can be caused by intrinsic factors, randomness within the system, or extrinsic factors, randomness in other systems that interact with the system of interest (Tsimring, 2014). This stochastic dynamics has been observed at all scales, ranging from the microscopic scale, such as gene expression (Fraser & Kaern, 2009; Kaern *et al.*, 2005), to the macroscopic scale, such as cell-to-cell interactions (Simpson *et al.*, 2009) or cell population dynamics (Kussell & Leibler, 2005; Patra & Klumpp, 2013).

Altogether, this can be interpreted as stochastic modelling being a more appropriate approach (over deterministic modelling) in the context of biological systems. However, the analysis of stochastic models is limited by the scale and complexity of the system of interest; that is, as the complexity of the system or the number of possible states increase, the model becomes more computationally expensive to analyse (Fadai *et al.*, 2019; Fahse *et al.*, 1998; Simoni *et al.*, 2020). On the other hand, deterministic models are very versatile due to the extensive existing literature on ordinary differential equation analysis (Bock, 1983; Deuflhard, 1985; Gratie *et al.*, 2013), but they do not consider random fluctuations that occur in the system of interest, which can be of great importance (Hahl & Kremling, 2016). These limitations on both types of model can be considered, in a sense, complementary. Stochastic modelling allows for the study of noisy dynamics on a relatively small state space, while deterministic modelling allows for the study of larger populations by neglecting the random fluctuations that are part of the dynamics of the system (Hahl & Kremling, 2016).

1.1 Objectives of this Thesis

In Chapter 2 the immune cells that will be the focus of this Thesis, *T cells*, are introduced. Their population dynamics has been widely studied but is not yet fully understood (Doherty & Christensen, 2000; Henrickson & von Andrian, 2007). Different types, or *phenotypes*, of T cells are known to exist (Apetoh *et al.*, 2015; Farrant *et al.*, 1994; Newell *et al.*, 2012), and each of them plays a different role during and immune response (Murphy & Weaver, 2016, Chapter 11). In Section 2.2 it is explained how T cells of a given phenotype can *differentiate* to a different phenotype when certain conditions are met, and which of these phenotypes will be considered for the models presented here. In the context of T cell immunology, there is a phenomenon known as *cross-reactivity*, in which multiple different T cells can recognise several *epitopes* (small protein fragments derived from an invading pathogen or from cells belonging to the host organism), and there is an overlap of recognition between T cells (Elong Ngono & Shresta, 2019; Mateus *et al.*, 2021; Moris *et al.*, 2011; Webster & Askonas, 1980). The aim of this Thesis is to present a mathematical model, and methods to study the phenomenon of cross-reactivity in the context of multiple infections occurring over time. In Chapter 3 the mathematical and statistical methods used throughout this Thesis are introduced.

For *naive* phenotype T cells; that is, cells that have not taken part in an immune response before, stochastic models have been proposed to describe their population dynamics (Stirk *et al.*, 2008, 2010). However, these models only consider at most two populations of cells competing for survival stimulus. Thus, the effect of multiple populations of naive T cells competing for survival stimulus has not been studied in as much detail. In Chapter 4, a generalised version of the model proposed by Stirk *et al.* (2010) for $\eta \geq 3$ different T cell populations is presented, and stochastically analysed to study the effects of new populations of T cells being introduced to the system. Using this model, the time to extinction of the populations is calculated, together with the distribution of population sizes for surviving populations when an extinction event takes place. More than this, the distribution of the number of divisions is also calculated in order to gain better understanding of the proliferation of naive cells before they become extinct.

Using the method described in Section 4.2.3, a deterministic approximation of the stochastic model is found, and used to describe the long term behaviour of the populations before extinction. Then, the long term dynamics of the stochastic and deterministic models are compared.

During an infection, three phenotypes are usually considered to be of the greatest importance: *naive*, *effector*, and *memory* cells (Ahmed & Gray, 1996; Bevan & Fink, 2001; Doherty & Christensen, 2000). When an immune response is initiated, there is a perturbation of the populations of T cells involved in it, in the form of differentiation into the effector phenotype, followed by a rapid expansion of cells of this phenotype (Appay & Rowland-Jones, 2004; Wherry & Ahmed, 2004). This process has implication on future immunity to infections, since following the triggering of an immune response, immunological memory is generated (Gil *et al.*, 2015; Lanfermeijer *et al.*, 2020; Yang *et al.*, 2022). This is, therefore, important in the context of cross-reactivity, as it has been observed that infection with a given variant of influenza A virus can grant immunity to other variants (Duan *et al.*, 2015; Hillaire *et al.*, 2013; McMichael *et al.*, 1983; Sridhar *et al.*, 2013). In Chapter 5, an extension of the model defined in Chapter 4 is presented to include the three phenotypes previously mentioned, as well as initiation of the immune response due to the presence of foreign peptides. Given the size of the effector population during infection, which is orders of magnitude greater than naive cells (Lanzavecchia & Sallusto, 2001), a deterministic approximation of the model is found. Using the methods described in Hong *et al.* (2020), parameter identifiability analysis is performed on the deterministic model to determine whether it can be parametrised with novel experimental data provided by Jessica Gaevert, and Paul Thomas from the Paul Thomas laboratory in the Immunology department at St. Jude Children's Research Hospital. This data is then analysed using the statistical methods presented in Section 3.3, in order to elucidate the importance the order of infection has in the generation of memory to influenza A viruses.

The analysis performed in Chapter 5 is highly specific to the pathogen and variants being considered, requiring knowledge of the specific epitopes that elicit cross-reactive immune responses. However, this is information that is not generally known for all pathogens. For this reason, in Chapter 6 a general approach to

study mathematically generated hypotheses, in the form of randomly generated cross-reactivity strategies, is proposed. Three different random generation models are presented, and using the stochastic model defined in Chapter 5, stochastic realisations are used to study and compare the immune responses associated with the different hypotheses. The hypotheses considered are framed in the context of *focussed* and *unfocussed* cross-reactivity proposed by Mason (1998), and the results suggest that both types of cross-reactivity lead to T cell responses that are biologically plausible.

Chapter 2

Immunological background

In this Chapter a general background of the immunological concepts discussed throughout this Thesis is provided. The immune system is a complex network of different types of cells and molecules which work together to protect the body from infections caused by pathogens, such as viruses or bacteria (Murphy & Weaver, 2016, Chapter 1). Immune responses to infection can be split into two categories: innate responses, which provide non-specific protection by targeting invariant sections of commonly encountered pathogens (Murphy & Weaver, 2016, Chapter 2), and adaptive responses, which provide protection by the selection and expansion of specific immune cells that can recognise molecules associated with a given pathogen (Murphy & Weaver, 2016, Chapter 3). These molecules are commonly referred to as antigens. An important distinctive feature between these two types of immune responses is that adaptive immunity has the capability of generating long lasting immunity, therefore providing protection against future infections by the same pathogen (Lauvau & Soudja, 2015).

Within the adaptive immune system there are two cell populations of particular importance, they are called T and B lymphocytes (or T and B cells for short), and they are characterised by the presentation of the T or B cell receptor on the surface, respectively (Casola *et al.*, 2004; Varma, 2008). B cells generate immunity by producing antibodies (Cumbers *et al.*, 2002; Tonegawa, 1983). These antibodies can then be used to combat pathogens by binding to their antigen and preventing them from entering and infecting healthy cells (*neutralisation*) (Casadevall & Pirofski, 2004; Chan *et al.*, 2009), or by marking them to be cleared from the

body by macrophages (*opsonisation*) (Owens III & Peppas, 2006; Trotter *et al.*, 1986; Ziemssen & Ziemssen, 2005).

T cells can be categorised in two distinct types of cells: those which present the CD4 co-receptor on their surface, and those which present the CD8 co-receptor. These two types of T cells are respectively called $CD4^+$ and $CD8^+$ cells. $CD4^+$ are referred to as T helper cells, since they are involved in the activation of B cells, $CD8^+$ T cells, and other immune cells during infection (Qi *et al.*, 2014a; Weaver *et al.*, 2007; Zhu & Paul, 2008). On the other hand, $CD8^+$ cells are called cytotoxic T cells given their function of inducing the death of infected cells (Andersen *et al.*, 2006; Chisari *et al.*, 1997). In the context of infection, both $CD8^+$, and $CD4^+$ cells mount an immune response when they become activated by the recognition of an antigen they are able to recognise being presented in a major histocompatibility complex (MHC) of class I and II, respectively (Maryanski *et al.*, 1997; Rock *et al.*, 2016). The antigen being presented by MHC molecules is a short protein, called a peptide or epitope, produced by the infecting pathogen (Mantegazza *et al.*, 2013). The specific epitopes a T cell is able to recognise are called its *cognate epitopes*. Since a general aim of this Thesis is to model T cell immunity, the remainder of this chapter will focus on T cell immunology.

2.1 The T cell receptor

As alluded to previously, T cells are identified by the T cell receptor (TCR) presented in their surface, which allows them to recognise peptides presented in MHC molecules by antigen presenting cells (APCs). The collection of all the different TCRs presented by T cells in an individual is called its *T cell repertoire*. For humans, the size of the repertoire is estimated to be approximately 4×10^6–10^8 distinct TCRs (Arstila *et al.*, 1999; Nikolich-Žugich *et al.*, 2004; Qi *et al.*, 2014b; Robins *et al.*, 2009). Each T cell presents copies of only one specific TCR on its surface (Varma, 2008), which is able to recognise a variety of different epitopes (Lang *et al.*, 2002; Mason, 1998; Selin *et al.*, 1994; Sewell, 2012; van den Berg *et al.*, 2011; Yin & Mariuzza, 2009). All T cells that present the same TCR on their surface are said to constitute a *T cell clonotype*. Structurally, each TCR consists of paired α and β chains, which together with the CD4 and CD8

co-receptors are responsible for the recognition of presented epitopes (Reiser *et al.*, 2003). However, the TCR by itself is not able to signal the nucleus of the cell once it encounters a cognate epitope, and therefore it depends on CD3 molecules (composed of ε, γ, δ, and ζ chains), which can transfer the signals received by the TCR to the cell nucleus (Kuhns *et al.*, 2006). This interaction between a TCR and a peptide being presented in MHC is shown in Figure 2.1.

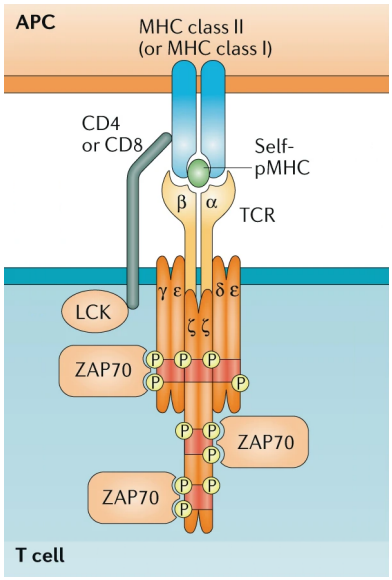


Figure 2.1: Interaction between a TCR, and a peptide being presented in an MHC molecule by an antigen presenting cell, adapted from Gaud *et al.* (2018). The α, and β chains, together with the CD4 or CD8 co-receptor are responsible for the recognition of the peptide being presented. While the ε, γ, δ, and ζ chains of the CD3 molecule are responsible for signalling the cell nucleus if a cognate epitope is encountered.

The α and β chains of the T cell receptor derive their great diversity from the V(D)J recombination process (Alt *et al.*, 1992; Schatz & Ji, 2011; Schatz & Swanson, 2011). During this process, through a complex set of genetic mechanisms (Lieber, 2010; Lieber & Wilson, 2010; Verkaik *et al.*, 2002), one of each V, D, and J genes is chosen and combined to generate the DNA sequence that produces a cell's TCR (Alt *et al.*, 1992; Thompson, 1995). For the β chain, specifically in humans, there are 48 possible choices of V genes, 2 possible D genes, and 13

possible J genes (Murugan *et al.*, 2012; Schatz & Swanson, 2011). It was shown by Murugan *et al.* (2012) that the functional β chains that can be generated from these genes have an entropy of about 47 bits; that is, about 2^{47} possible β chains can be generated from these genes. Without accounting for recombination events that result in the same β chain being generated, a phenomenon called convergent recombination, the entropy rises to 52 bits, which can be separated into contributions from gene choice, nucleotide insertion events, and nucleotide deletion events as follows: 9 bits come from the choice of gene, 30 bits from insertion events, and 13 bits from deletion events. This decomposition is summarised in Figure 2.2. This means that most of the diversity of the TCR repertoire comes from insertion and deletion events, and only a small portion of it is due to the variety of V, D, and J genes.

Figure 2.2: Decomposition of the entropy of human TCR β chains by contribution of gene choice, nucleotide insertion events, and nucleotide deletion events, taken from Murugan *et al.* (2012). The entropy of functional β chains (47 bits) is smaller than the entropy of possible DNA recombination events (52 bits) due to convergence of different events to the same chain. Gene choice represents 9 bits of the total entropy, and deletion events represent 13 bits. The majority of the TCR diversity comes from insertion events, with a contribution of 30 bits. The last row of the diagram further divides the entropy into gene choice, insertion events, and deletion events for V, D, and J genes separately.

T cells start their development in the bone marrow, similarly to B cells (Murphy & Weaver, 2016, Chapter 8). However, they migrate to the thymus where they begin the process of generating a TCR, and committing to a CD4 or a CD8 co-receptor (Murphy & Weaver, 2016, Chapter 8). Once a functional TCR β chain is generated, T cells undergo two important processes in the thymus to ensure that the α chain that is generated will result in a functional TCR that is also not

self-reactive; that is, it will not become activated when peptides derived from host cells are encountered. In other words, processes that ensure T cells will be able to initiate immune responses when they encounter their cognate epitope, but not when they recognise a peptide which has been derived from a healthy cell, called a *self-peptide*. Positive selection is the process by which cells with a non-functional TCR are removed from the pool of developing T cells (Anderson & Takahama, 2012; Starr *et al.*, 2003). At this stage of development T cells are programmed to die naturally by apoptosis if they do not receive the stimulus provided by TCR-epitope interactions. Thus, only T cells with a functional TCR will be able to receive survival stimulus from thymic epithelial cells, while cells without a functional TCR will be deprived of this stimulus and die by apoptosis (Alam *et al.*, 1996; Anderson & Takahama, 2012). Furthermore, after positive selection, T cells which present both the CD4 and CD8 co-receptors mature into single positive T cells, which present only CD4 or CD8 (Palmer, 2003; Starr *et al.*, 2003). This choice of co-receptor is caused by secondary signalling via cytokines (Luckey *et al.*, 2014; Park *et al.*, 2010; Singer *et al.*, 2008). In order to prevent T cells from being self-reactive, and cause autoimmune disease, the process of negative selection identifies cells that show high affinity to self-peptides presented in the context of MHC and causes them to die by apoptosis (Palmer, 2003).

2.2 T cell mediated immune responses

T cells that have undergone positive and negative selection exit the thymus into the periphery, and remain in a naive in-activated state until they come into contact with a cognate epitope that causes them to initiate an immune response (Goronzy *et al.*, 2015; Sprent & Surh, 2011; Weinreich & Hogquist, 2008). These naive cells circulate through the lymphatic system and can reside in secondary lymphoid organs (Mackay *et al.*, 1990; Surh & Sprent, 2008; Takada & Jameson, 2009). In homeostatic conditions, that is to say, when there is no infection, the repertoire of naive T cells is maintained by the homeostatic proliferation stimulus provided by self-peptides presented in the context of MHC (Boyman *et al.*, 2012; Rudd *et al.*, 2011; Sprent & Surh, 2011; Surh & Sprent, 2005).

During an infection, *professional* APCs present peptides derived from the infecting pathogen bound to MHC molecules, which together with other co-stimulatory molecules activate naive T cells specific to the given peptide, causing them to differentiate into the effector phenotype, and gain cytotoxic capabilities in the case of $CD8^+$ cells (Yewdell & Hill, 2002). Once activated and differentiated, effector phenotype cells begin dividing and migrating to the site of infection, where they are able to recognise infected cells by the epitopes they present in their MHCs (Mondino *et al.*, 1996). Once an effector T cell encounters an infected cell, it induces it to undergo apoptosis, therefore killing the infected cell, and preventing it from infecting other cells (Murphy & Weaver, 2016, Chapter 9).

After this differentiation from the naive to the effector phenotype, and the proliferation of effector cells that follows (called expansion of the clonotype), once the infection is cleared from the system (in the case of acute and not chronic infection) the population of effector cells begins to decline, since they are dependent on constant peptide stimulation to proliferate (Huppa *et al.*, 2003). However, not all effector cells die after the infection is cleared. A fraction of the effector cells present after the infection is cleared goes on to differentiate into the memory phenotype (Kaech & Cui, 2012). The process of memory generation is not yet fully understood, but it has been observed that between 5–10% of the effector population present when the infection is cleared differentiates into the memory phenotype (Ahmed & Gray, 1996). A general representation of this expansion into effector cells and contraction back to memory cells is shown in Figure 2.3.

T cells of the memory phenotype are long lived (Demkowicz Jr *et al.*, 1996; Okhrimenko *et al.*, 2014), and homeostatically maintained by a different type of signalling provided by cytokines (Harty & Badovinac, 2008). This allows the memory compartment to be maintained long after the infection has been cleared independently of TCR-peptide interactions, therefore providing immunity to future infections by the same pathogen. Furthermore, memory cells do not require co-stimulation in order to become activated and differentiate to the effector compartment (Rosenblum *et al.*, 2016; Yewdell & Hill, 2002), allowing for an immune response to be initiated more quickly during subsequent infections with the same pathogen. Figure 2.4 shows a visual representation of the differentiation pathway discussed so far, where naive cells differentiate into effector cells, which

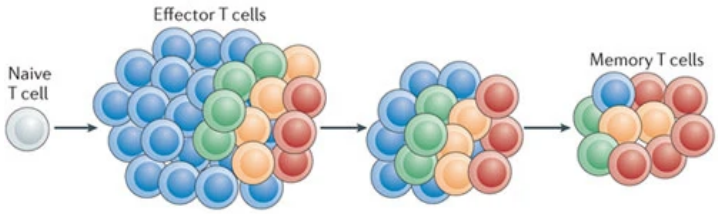


Figure 2.3: Example of the behaviour of different T cell phenotypes during the course of an infection, taken from Kaech & Cui (2012). When a naive cell encounters its cognate epitope it differentiates into the effector phenotype, and gains cytotoxic capabilities (if it is $CD8^+$) in order to fight the infection. After the infection is cleared from the body the population of effector cells declines, and a small portion of the remaining cells differentiates into the memory phenotype. Colours represent different T cell clonotypes, showing the diversity of T cells that become activated during infection.

in turn differentiate into memory cells. Memory cells can then differentiate back into effector cells if their cognate epitope is encountered again. Note that there exist more T cell phenotypes that are not considered in the models presented in this Thesis, for example, regulatory T cells (Vignali *et al.*, 2008), or central memory T cells (Pepper & Jenkins, 2011).

As previously mentioned, each TCR is able to recognise more than a single epitope in the context of MHC presentation. This can be evidenced by the fact that the total number of T cells in the human body (about 10^{12}) is outnumbered by the total number of possible epitopes of length 11 that can be presented by MHC (about 6×10^{12}), but the immune system is still able to provide protection against the majority of pathogens encountered throughout life (Mason, 1998). This phenomenon, where TCRs are able to recognise more than a single epitope, is called *cross-reactivity*. While it is not yet fully understood, T cell cross-reactivity plays an important role in pre-existing immunity and vaccination, as it has been observed, for example, that previous infections with influenza A virus can provide protection against other strains of the virus in subsequent infections (Duan *et al.*, 2015; Gras *et al.*, 2010; Sewell, 2012).

On a population level, it has been observed that in some cases immune responses from different individuals to the same pathogen are dominated by

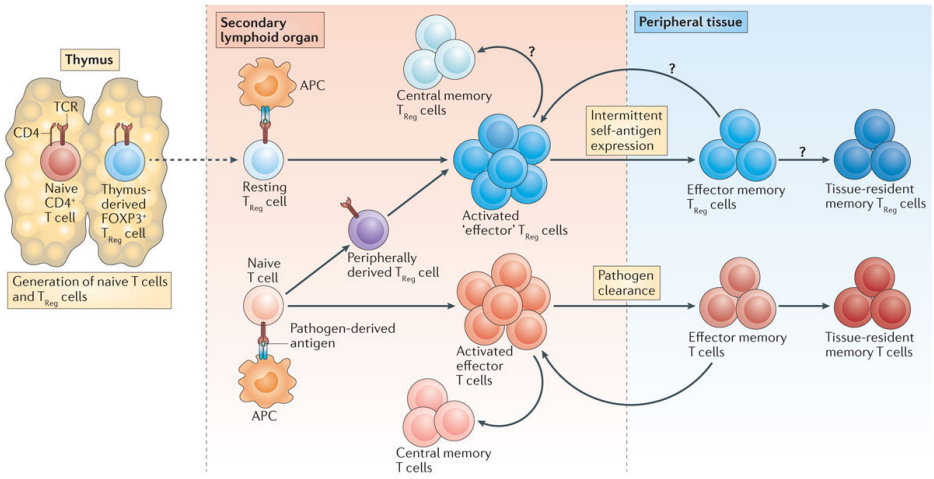


Figure 2.4: Differentiation pathway of T cells during infection, taken from Rosenblum *et al.* (2016). Naive cells can differentiate into effector cells once their cognate epitope is encountered. Then, once infection is cleared a fraction of effector cells differentiate into memory cells, which are homeostatically maintained in order to provide future immunity to re-infection with the same pathogen. Note that there are more T cell phenotypes which are not considered in the models proposed in this Thesis.

T cells presenting identical TCRs (Venturi *et al.*, 2008). This type of immune response, where multiple individuals have the same dominant TCR, is called a *public* T cell response (Li *et al.*, 2012; Venturi *et al.*, 2008). However, not all infections elicit this type of immune response. In many cases each individual immune response is dominated by a distinct TCR, which is not shared with other individuals (Cibotti *et al.*, 1994; Kim *et al.*, 2005). This type of response is called a *private* T cell response. The mechanisms that cause these distinct types of immune responses to arise are not yet fully understood, but several hypotheses exist on the biological reason why public T cell responses are observed. Some hypotheses hinge on the structure of the peptide (Miles *et al.*, 2005; Stewart-Jones *et al.*, 2003; Tynan *et al.*, 2005) or the TCR itself (Kjer-Nielsen *et al.*, 2003; Tynan *et al.*, 2007), suggesting that it is the 3D structure of the TCR-peptide interaction which causes this behaviour. Another hypothesis is that public TCRs exist because there are TCRs that are generated more frequently than others, causing them to

be more prevalent in the general population, and therefore causing public immune responses to be observed (Venturi *et al.*, 2006). This hypothesis is based on the previously mentioned phenomenon of convergent recombination, where different recombination events result in the same TCR being generated.

Chapter 3

Mathematical background

This Chapter introduces the mathematical and statistical methods to be used throughout this Thesis. In Section 3.1 the theory of stochastic processes is presented, followed in Section 3.2 by the theory of ordinary differential equations used to study deterministic models. Section 3.3 introduces the statistical techniques to be used on cell population data. Finally, in Section 3.4 the concepts from network theory used to define the models are presented.

3.1 Stochastic processes

In this Section the type of stochastic model used in Chapters 4, 5, and 6 is introduced, as well as some methods used to study stochastic models. The definitions provided are based on those found in Allen (2010), and Pinsky & Karlin (2010).

3.1.1 Continuous-time Markov chains

In order to define the specific type of stochastic process that will be considered in Chapters 4, 5, and 6, first a stochastic process must be formally defined.

Definition 1. A *stochastic process* is a collection

$$\mathcal{X} = \{X(t) : t \in T\}$$

of random variables, where T is an index set, and $X(t)$ is a random variable defined over a state space $\mathcal{S}$.

Remark. Note that a stochastic process can also be a collection of random vectors

$$\mathcal{X} = \{\mathbf{X}(t) \coloneqq (X_1(t), \ldots, X_\eta(t)) : t \in T\},$$

with $\eta \in \mathbb{N}$.

In the stochastic processes considered in this Thesis, the index set T will be used to represent time, thus the processes will track how a random variable or vector behaves over time. If the index set consists of discrete values, *e.g.*, $T = \{0, 1, 2, 3, \ldots\} = \mathbb{N}_0$, then the process is said to be a *discrete-time stochastic process*. On the other hand, if the index set takes continuous values, *e.g.*, $T = [0, +\infty)$, the process is said to be a *continuous-time stochastic process*.

Definition 2. A continuous-time stochastic process $\mathcal{X} = \{\mathbf{X}(t) : t \in [0, +\infty)\}$ is said to be a *continuous-time Markov chain (CTMC)* if it is defined over a discrete state space $\mathcal{S}$, and it satisfies the *Markov property*. That is, for any sequence of real numbers $0 \leq t_0 < t_1 < \cdots < t_n < t_{n+1}$, the following is true

$$\begin{aligned}\mathbb{P}\{\mathbf{X}(t_{n+1}) &= \mathbf{n}_{n+1} | \mathbf{X}(t_0) = \mathbf{n}_0, \mathbf{X}(t_1) = \mathbf{n}_1, \ldots, \mathbf{X}(t_n) = \mathbf{n}_n\} \\ &= \mathbb{P}\{\mathbf{X}(t_{n+1}) = \mathbf{n}_{n+1} | \mathbf{X}(t_n) = \mathbf{n}_n\},\end{aligned}$$

for any $\mathbf{n}_i \in \mathcal{S}$.

3.1.2 Transition probabilities

For a given CTMC $\mathcal{X} = \{\mathbf{X}(t) : t \in [0, +\infty)\}$, every random variable $\mathbf{X}(t)$ has an associated probability distribution $\{p_{\mathbf{n}}(t) : \mathbf{n} \in \mathcal{S}\}$, where these probabilities are given by

$$p_{\mathbf{n}}(t) = \mathbb{P}\{\mathbf{X}(t) = \mathbf{n}\}.$$

Transition probabilities provide a relation between the state of the stochastic process at different timepoints, and they are defined as follows:

Definition 3. The transition probability from state $\mathbf{n}$ at time s, to state $\mathbf{m}$ at time t is defined as

$$p_{\mathbf{nm}}(s, t) = \mathbb{P}\{\mathbf{X}(t) = \mathbf{m} | \mathbf{X}(s) = \mathbf{n}\},$$

with $s < t$. If the transition probabilities depend only on the length of the interval, $t - s$; that is

$$\begin{aligned} p_{\mathbf{nm}}(s,t) &= \mathbb{P}\left\{\mathbf{X}(t) = \mathbf{m} | \mathbf{X}(s) = \mathbf{n}\right\} \\ &= \mathbb{P}\left\{\mathbf{X}(t-s) = \mathbf{m} | \mathbf{X}(0) = \mathbf{n}\right\} \\ &= p_{\mathbf{nm}}(0, t-s) \\ &= p_{\mathbf{nm}}(t-s), \end{aligned}$$

the CTMC is said to be *stationary* or *time-homogenous*.

The transition probabilities can be represented naturally as a matrix, $\mathbf{P}(t)$, called the *transition probability matrix*. This is a square matrix whose order is given by the cardinality of $\mathcal{S}$, denoted by $|\mathcal{S}|$, where the entry in the $\mathbf{n}$-th row and $\mathbf{m}$-th column is the transition probability $p_{\mathbf{nm}}(t)$. It is easy to see that each row of $\mathbf{P}(t)$ adds up to one, since for any state $\mathbf{n}$ the process must either travel to another state $\mathbf{m} \in \mathcal{S}$, or remain in $\mathbf{n}$.

3.1.3 Infinitesimal generator matrix

The transition probabilities, $p_{\mathbf{nm}}(t)$, can be used to derive transition rates $q_{\mathbf{nm}}$ between states. Assuming that the transition probabilities $p_{\mathbf{nm}}(t)$ are continuous and differentiable for $t \geq 0$, and satisfy

$$p_{\mathbf{nm}}(0) = 0, \mathbf{n} \neq \mathbf{m}, \qquad p_{\mathbf{nn}}(0) = 1,$$

then the transition rates are defined as

$$q_{\mathbf{nm}} = \begin{cases} \lim\limits_{\Delta t \to 0^+} \dfrac{p_{\mathbf{nm}}(\Delta t) - p_{\mathbf{nm}}(0)}{\Delta t} = \lim\limits_{\Delta t \to 0^+} \dfrac{p_{\mathbf{nm}}(\Delta t)}{\Delta t} & \text{for } \mathbf{n} \neq \mathbf{m}, \\ \lim\limits_{\Delta t \to 0^+} \dfrac{p_{\mathbf{nn}}(\Delta t) - p_{\mathbf{nn}}(0)}{\Delta t} = \lim\limits_{\Delta t \to 0^+} \dfrac{p_{\mathbf{nn}}(\Delta t) - 1}{\Delta t} & \text{for } \mathbf{n} = \mathbf{m}. \end{cases}$$

Since every row of the transition probability matrix adds up to one, it can be shown that

$$q_{\mathbf{nn}} = -\sum_{\mathbf{m} \in \mathcal{S} \setminus \{\mathbf{n}\}} q_{\mathbf{nm}}.$$

Similarly to the transition probabilities, the transition rates can be represented as a matrix $\mathbf{Q}$, called the *infinitesimal generator matrix*. The entry in the $\mathbf{n}$-th row and $\mathbf{m}$-th column is the transition rate $q_{\mathbf{nm}}$. From the previous equation it is easy to see that each row of $\mathbf{Q}$ adds up to zero.

3.1.4 Kolmogorov differential equations

The forward and backward Kolmogorov differential equations describe the rate of change of the transition probabilities $p_{\mathbf{nm}}(t)$. In matrix form, the forward Kolmogorov equation is given by

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

where $\mathbf{I}$ is the identity matrix. The backward Kolmogorov differential equation can be expressed in matrix form as

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

3.1.5 Linear noise approximation

The solution to the forward Kolmogorov equation is given by

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

with the matrix exponential function defined as (Moler & Van Loan, 1978, 2003)

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

This solution describes the probability of being in each state of the state space $\mathcal{S}$ at time t. However, this solution is usually computationally intractable to calculate due to the size of the state space, since a large state space would require the computation of large (or even infinite in the case on an infinite state space) matrix exponentials. Therefore, approximations to the solution of the master equation have been developed. The *linear noise approximation*, developed by van Kampen (2007), is one such method to approximate the solution of the forward Kolmogorov equation chiefly used when considering chemical reactions (Bortolussi *et al.*, 2016; Cardelli *et al.*, 2016).

This approximation assumes that the reactions are occurring within a fixed volume Ω (Elf & Ehrenberg, 2003). It provides a second order approximation of the forward Kolmogorov equation, by considering a large volume expansion around the steady state of its solution, and finding differential equations that describe

the first and second order moments of the random noise around the solution (van Kampen, 2007). The method consists of calculating the Taylor expansion of the forward Kolmogorov equation in terms of $\Omega^{-1/2}$. The terms of first-order in this expansion result in the deterministic equations describing the concentrations of the populations in the system. While the terms of second order result in a linear Fokker-Planck equation describing the fluctuations around the steady state of the deterministic approximation (van Kampen, 2007, Chapter 8).

3.1.6 Stochastic simulation algorithms

As previously mentioned, finding the solution to the forward Kolmogorov equation is usually not computationally tractable. Therefore, numerical realisations of the stochastic model are usually employed in order to verify other analytical results for the stochastic process in question. An exact algorithm to simulate a Markov process was developed by Gillespie (1977).

For a given CTMC, $\mathcal{X}$, described by a series of possible reactions which depend on the current state x, $R_i(x)$, with $1 \leq i \leq M$, where $r_i(x)\Delta t$ is the probability that reaction $R_i(x)$ will occur within the time interval $(t, \Delta t)$, taking the CTMC from state x_0 at time t to state x_i at time $t + \Delta t$, the exact stochastic simulation algorithm is executed as follows: First the total rate at which any reaction occurs, $r(x) = \sum_{i=1}^{M} r_i(x)$, is calculated. Then, a random number sampled uniformly from $[0, 1]$ is used to determine which of the reactions will occur, and a second random number sampled uniformly from $[0, 1]$ is used to determine the time elapsed until this reaction takes place through inverse transform sampling. Finally, the state and time of the system get updated, and the algorithm is continued until the desired maximum time is reached. An outline of this algorithm is shown in Algorithm 3.1.

3.2 Ordinary differential equations

In this section the basic notions of deterministic modelling with ordinary differential equations (ODEs) are introduced for the models presented in Chapters 4, 5, and 6. As discussed in Chapter 1, deterministic models can be interpreted as

Algorithm 3.1: Outline of the exact stochastic simulation algorithm developed by Gillespie (1977).

Input: $t_{max} \leftarrow$ maximum time of the simulation,
$x_0 \leftarrow$ initial state of the simulation.

$t \leftarrow 0$ – Current time is set to 0;

$x \leftarrow x_0$ – Current state is set to x_0;

while $t \leq t_{max}$ **do**

$r(x) = \sum_{i=1}^{M} r_i(x)$;

Sample $u_1, u_2 \sim U(0,1)$;

$R_k \leftarrow$ reaction k such that $\sum_{i=1}^{k-1} \frac{r_i(x)}{r(x)} < u_1 \leq \sum_{i=1}^{k} \frac{r_i(x)}{r(x)}$;

$x \leftarrow x_k$ – Current state is updated to reflect reaction R_k taking place;

$t = t - \frac{\log u_2}{r(x)}$;

end

approximations of stochastic models where random fluctuations are ignored (Hahl & Kremling, 2016), and thus they will be used when the stochastic processes under consideration become computationally intractable to study. The definitions provided here are based on those found in Allen (2007).

Differential equations are classified by their order, where a differential equation of order n is of the form

$$f\left(x, \frac{dx}{dt}, \frac{d^2x}{dt^2}, \ldots, \frac{d^nx}{dt^n}, t\right) = 0.$$

In the models presented in this Thesis, the functions $x(t)$ will be used to represent the expected number of cells of a given T cell population at time t, and first order ODEs will be considered to describe the rate of change of those populations of cells over time. In other words, a first order ODE will describe the dynamics of the T cell population over time. In a first order equation, such as

$$c_1(t)\frac{dx}{dt} + c_0(t)x = g(t),$$

if the coefficients $c_0(t)$ and $c_1(t)$ are constant, or a function of only t and not x or $\frac{dx}{dt}$, then the equation is said to be *linear*. If the function $g(t)$ is identically zero, then the equation is said to be *homogeneus*.

In most use cases, instead of a single ODE, a system of ODEs is considered, in this way (for the purposes of this Thesis) the population of more that a single T cell clonotype can be tracked simultaneously over time. This approach with a system of ODEs also allows for the modelling of interaction between populations, such as competition for a shared resource. For a given set of initial conditions, it is possible in some cases to obtain an analytic solution of the system of ODEs (Murphy, 2011). However, in the case of more complex systems of equations, an analytical solution is not always possible. In those cases numerical methods can be used to evolve the system over time from an initial condition by using numerical integration (Griffiths & Higham, 2010; Kang & Cheek, 1972).

3.2.1 Steady states

An important solution for a system of ODEs is the constant solution, $\mathbf{x}^*$, which satisfies

$$\mathbf{f}(\mathbf{x}^*) = 0,$$

and is called a *steady state* of the system. This type of solution is of interest in the context of modelling since it represents the state in which the rate of change of all populations being modelled is exactly 0, meaning that the populations are no longer changing. In general, a system of ODEs can have multiple steady states. However, when considering biological quantities, such as a cellular population, steady states that are not biologically relevant, such as those with negative values, are not taken into account.

3.2.2 Stability analysis

The steady states of an ODE system can be categorised according to their stability as *locally stable*, *asymptotically stable*, or *unstable*, depending on the behaviour of the solutions around them. In layman's terms, a steady state $\mathbf{x}^*$ is locally stable if the solution to the system with initial conditions close to the steady state remains

close the steady state as $t \to +\infty$, or it is asymptotically stable if it is both locally stable, and solutions with initial conditions close to the steady state approach it as $t \to +\infty$. Formally stability is defined as follows:

Definition 4. A steady state $\mathbf{x}^*$ of a system of ODEs, $\mathbf{f}$, is *locally stable* if for every $\varepsilon > 0$, there exists $\delta > 0$ such that for every solution $\mathbf{x}(t)$ with initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$, the following is true

$$\|\mathbf{x}_0 - \mathbf{x}^*\| < \delta \Rightarrow \|\mathbf{x}(t) - \mathbf{x}^*\| < \varepsilon,$$

for all $t \geq t_0$. Furthermore, if the following condition is satisfied

$$\|\mathbf{x}_0 - \mathbf{x}^*\| < \delta \Rightarrow \lim_{t \to +\infty} \|\mathbf{x}(t) - \mathbf{x}^*\| = 0,$$

then the steady state is said to be *asymptotically stable*. If neither of these conditions is met the steady state is said to be *unstable*.

3.2.3 Identifiability analysis

Structural identifiability is a method by which it can be verified whether, given a set of data, unique parameter values can be found for a deterministic model that represents the dynamics of the data. The formal definition of structural identifiability presented here is based on that found in Chis *et al.* (2011). In this definition a system of ODEs, $\dot{\mathbf{x}}$, together with a vector of experimentally observed quantities, $\mathbf{y}$, are considered to depend on the state of the system $\mathbf{x}$, and an unknown-parameter vector $\mathbf{p}$. This system is denoted by $\sum(\mathbf{p})$ in order to make explicit its dependance on the parameter vector.

Definition 5. Given a biological system described by

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

where $\mathbf{x} = (x_1, \ldots, x_{n_x}) \in \mathbf{M} \subset \mathbb{R}^{n_x}$ is the state of the system, with $\mathbf{M}$ a subset of $\mathbb{R}^{n_x}$ which contains the initial state, $\mathbf{u} = (u_1, \ldots, u_{n_u}) \in \mathbb{R}^{n_u}$ an input control vector, and $\mathbf{y} = (y_1, \ldots, y_{n_y})$ is the vector of experimentally observed quantities.

The vector of unknown parameters $\mathbf{p} = (p_1, \ldots, p_{n_p}) \in \mathbf{P}$ is assumed to belong to an open connected subset of $\mathbb{R}^{n_p}$. The entries $\mathbf{f}, \mathbf{h}, \mathbf{g}_1, \ldots, \mathbf{g}_{n_u}$ are analytic functions of their arguments.

Then, a parameter p_i, $i = 1, \ldots, n_p$ is:

- *structurally globally (uniquely) identifiable* if for almost any $\mathbf{p}^* \in \mathbf{P}$,

$$\sum(\mathbf{p}) = \sum(\mathbf{p}^*) \Rightarrow p_i = p_i^*,$$

- *structurally locally identifiable* if for almost any $\mathbf{p}^* \in \mathbf{P}$, there exists a neighbourhood $\mathbf{V}(\mathbf{p}^*)$ such that

$$\mathbf{p} \in \mathbf{V}(\mathbf{p}^*) \text{ and } \sum(\mathbf{p}) = \sum(\mathbf{p}^*) \Rightarrow p_i = p_i^*,$$

- *structurally non-identifiable* if for almost any $\mathbf{p}^* \in \mathbf{P}$, there exists no neighbourhood $\mathbf{V}(\mathbf{p}^*)$ such that

$$\mathbf{p} \in \mathbf{V}(\mathbf{p}^*) \text{ and } \sum(\mathbf{p}) = \sum(\mathbf{p}^*) \Rightarrow p_i = p_i^*.$$

In order to computationally assess the structural identifiability of a model, the SIAN structural identifiability toolbox (Hong *et al.*, 2019), which is based on the methods described in Hong *et al.* (2020), can be used to test for identifiability of parameters given a data set.

3.3 Statistical analysis

In this section, the statistical tests used in Chapter 5 to analise the T cell population data are introduced. The definitions presented here are based on those found in De Sá (2007), Montgomery (2017), and Abdi & Williams (2010).

3.3.1 Analysis of variance

Analysis of variance (ANOVA) is a statistical method, which tests, given k independent samples, whether the null hypothesis that the means of all groups are equal, *i.e.*, $\mu_1 = \mu_2 = \cdots = \mu_k$, against the alternative hypothesis that the means of at least one pair of samples are not equal, *i.e.*, $\mu_i \neq \mu_j$, for some pair

i, j. When there is only one categorical grouping level the test is called one-way ANOVA. In this case the variable being tested is called the *dependent variable*, while the variable with the groupings of sample is called the *independent variable*. The general idea of the method is to decompose the total variance of the data into variance within, and between groups.

Consider a sample of size n, split into k groups of sizes $n_1, n_2, \ldots, n_k$, and with sample means $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k$. Any value in the total sample can be denoted by x_{ij}, where $i = 1, \ldots, k$ indexes the groups within the sample, and $j = 1, \ldots, n_i$ indexes the value within the group it belongs to. Then, the total variance is related to the total sum of squares of deviations, SS, from the total sample mean, μ, by

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \mu)^2.$$

Now, adding and subtracting $\overline{x}_i$ to the deviations $x_{ij} - \mu$, the following expression can be found

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{x}_i - \mu)^2 + 2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)(\overline{x}_i - \mu),$$

where the last term is equal to zero, and thus

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{x}_i - \mu)^2,$$

where the first term of the equation is the *within-group sum of squares* (SS_W), and the second term is the *between-group sum of squares* (SS_B). Then, the total sum of squares can be written as

$$SS = SS_W + SS_B.$$

Now, each sum of squares can be written in terms of variances as

$$(n-1)\sigma^2 = (n-k)\sigma_W^2 + (k-1)\sigma_B^2,$$

where σ^2 is the total variance, σ_W^2 is the within-group variance, and σ_B^2 is the between-group variance. The total variance has $n-1$ degrees of freedom, while the

within-group and between-group variances have $n-k$, and $k-1$ degrees of freedom respectively. If the null hypothesis is false, that is at least one pair of means is not equal, then the variation between means would be large in comparison to the variation within means. The within-group variance can be written as

$$\sigma_W^2 = \frac{SS_W}{n-k},$$

and the between-group variance is given by

$$\sigma_B^2 = \frac{SS_B}{k-1}.$$

Then, if the null hypothesis is true, the ratio σ_B^2/σ_W^2 is expected to be close to one, while if the null hypothesis is false this ratio would be expected to be larger than one. This ratio is what is used to define the test statistic, by comparing it with the F distribution to test the validity of the null hypothesis, H_0. First, note that the sum of squares of k independent random variables, with standard normal distribution, follows the chi-squared distribution, $\chi^2(k)$. Then, the F distribution can be defined as the ratio of two independent χ^2 random variables as follows: if two independent random variables X_1, and X_2 are χ^2 distributed with d_1, and d_2 degrees of freedom, respectively, then the random variable

$$\frac{X_1/d_1}{X_2/d_2},$$

follows and $F(d_1, d_2)$ distribution. Then the test statistic for ANOVA is

$$F^* = \frac{\sigma_B^2}{\sigma_W^2} \sim F(k-1, n-k),$$

this means that if the test statistic F^* takes a value greater than one the null hypothesis is rejected, since this implies that the between group variance is greater than the within group variance, and thus there is a significant difference, and at least one pair of means is not equal.

3.3.2 Tukey's honest significant difference test

Once the ANOVA test is performed, if the null hypothesis is rejected it is not known exactly which of the groups have means significantly different to each other.

Tukey's honest significant difference (HSD) is the smallest amount by which the means must differ between two groups for them to be significantly different. This test uses a *studentised range statistic*

$$q = \frac{\overline{x}_{max} - \overline{x}_{min}}{\sqrt{\sigma_W^2/n}},$$

where $\overline{x}_{min}$ and $\overline{x}_{max}$ are the smallest and largest, respectively, of the sample means of k samples of size n from the same distribution. Then, Tukey's test statistic is defined as

$$T_\alpha = q_\alpha(a, f)\sqrt{\frac{\sigma_W^2}{n}},$$

for a significance level α, where a is the number of observations in each group, n is the total number of observations, and f are the degrees of freedom associated with σ_W^2. The value of $q_\alpha(a, f)$ can be obtained from a studentised range distribution table. Then, taking two means $\overline{x}_i$, and $\overline{x}_j$ of groups i, and j, respectively, Tukey's HSD test states that if the following condition is satisfied

$$|\overline{x}_i - \overline{x}_j| \geq T_\alpha,$$

then there is a significant difference between the two means at the level α. This test is used on each pair of groups when the ANOVA results in a rejection of the null hypothesis.

3.4 Network theory

In Chapters 4, 5, and 6 a network is used to define a recognition profiles of peptides and TCRs. In this section, the basic concepts of network theory used throughout this Thesis are introduced. The definitions presented here are based on those found in Bollobás (2002), Diestel (2006), and Newman (2018).

Definition 6. A *graph* or *network* is a pair $G = (E, V)$ of sets, such that $E \subseteq V \times V$. That is, the elements of E are pairs of elements from V.

The set V is called the *vertex set*, and its elements are called the *vertices*, *nodes*, or *points* of the graph, while the set E is called the *edge set*, and its elements are called the *edges*, or *lines* of the graph. The usual way to picture a graph is

to draw a dot for each element in the vertex set, and join two dots with a line if the corresponding edge is in the edge set. For example, the graph defined by $V = \{1, 2, 3, 4, 5, 6\}$, and $E = \{\{1, 2\}, \{1, 5\}, \{2, 5\}, \{3, 4\}, \{5, 6\}\}$ is shown in Figure 3.1. In order to make the notation less cumbersome, an edge $\{x, y\}$ is usually denoted by xy.

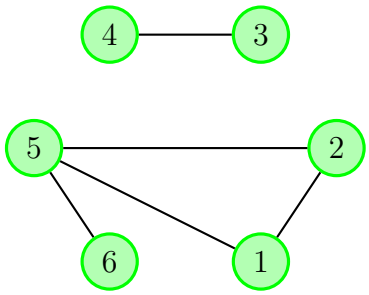


Figure 3.1: Graph defined by the vertex set $V = \{1, 2, 3, 4, 5, 6\}$, and the edge set $E = \{\{1, 2\}, \{1, 5\}, \{2, 5\}, \{3, 4\}, \{5, 6\}\}$.

A vertex $v \in V$ is said to be *incident* with an edge $e \in E$, if $v \in e$. The two vertices in an edge are called its *ends*, and the edge is said to *join* its ends. Two vertices are *adjacent*, or *neighbours*, if they are ends of the same edge; that is, there is an edge that connects them. Two edges, on the other hand, are *adjacent* if the have an end in common. A graph in which all vertices are pairwise adjacent is called a *complete graph*, and it is denoted by K^n, where n is the number of vertices in the graph, $n = |V|$.

Definition 7. For a given vertex v, the set of all edges in E for which v is an end, is denoted by $E(v)$. The *degree* of a vertex v is defined as $d(v) = |E(v)|$. If the degree of a vertex is equal to zero, the vertex is said to be *isolated*.

Definition 8. A *path* is a non-empty graph $P = (V, E)$ of the form

$$V = \{x_0, x_1, \ldots, x_k\} \qquad E = \{x_0x_1, x_1x_2, \ldots, x_{k-1}x_k\},$$

where the x_i are all distinct nodes. The vertices x_0 and x_k are *linked* by P, and are called its *ends*, while the vertices $x_1, x_2, \ldots, x_{k-1}$ are called the *inner vertices* of P. The *length* of a path is the number of edges that comprise it, and a path of length k is denoted by P^k.

Since a path P can be defined by the ordered sequence of its vertices, it is usually referred to as $P = x_0x_1 \dots x_k$. Furthermore, given two paths $P = x_0x_1 \dots x_k$, and $P' = x'_0x'_1 \dots x'_\ell$, such that $x_k = x'_0$, then the union of the paths, $P \cup P'$, is denoted by $P + P' = x_0x_1 \dots x_kx'_1 \dots x'_\ell$.

Definition 9. If $P = x_0 \dots x_{k-1}$ is a path, and $k \geq 3$, then the graph $C = P + x_{k-1}x_0$ is called a *cycle* or a *closed path*.

In the case of a cycle, the number of vertices is equal to the number of edges, thus the *length*, or *size*, of a cycle is the number of edges, or vertices, that comprise it. An example of a cycle of length 3, C^3, which is equivalent to a closed path of length 2, P^2, is shown in Figure 3.2.

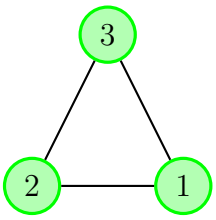


Figure 3.2: A closed path of length 2, P^2, or equivalently a cycle of size 3, C^3.

Definition 10. For a graph $G = (E, V)$, the *distance* between two nodes $x, y \in V$, $d_G(x, y)$, is defined as the length of the shortest path between x and y. If there exist no paths between x and y, the distance is defined to be $d_G(x, y) := +\infty$. The greatest distance between two vertices in G is called the *diameter* of G.

Definition 11. Given a graph G with n vertices of degrees $d_1 \leq \dots \leq d_n$, then the n-tuple $(d_1, \dots, d_n)$ is called the *degree sequence* of G.

Definition 12. Let $k \geq 2$ be an integer. A graph $G = (E, V)$ is called a *k-partite* graph if V admits a partition into k disjoint sets, such that every edge in E has its ends in different sets, or equivalently, no edge in E has both edges in the same subset.

It can be shown that an equivalent condition for a graph to be *bipartite*; that is, 2-partite, is that it contains no cycles of odd length (Diestel, 2006, Chapter 1).

Chapter 4

Multi-variate model of T cell clonotype competition and homeostasis

An adult human has approximately 4×10^{11} T cells (Alberts *et al.*, 2002; Jenkins *et al.*, 2009), each of them expressing about 3×10^4 identical T cell receptors (TCRs) on its surface (Varma, 2008). These receptors recognise self-peptides bound to major histocompatibility complexes (MHCs), which as bound complexes are called self-pMHCs. The interaction between TCRs and self-pMHCs signals a T cell to synthesise proteins important for survival and proliferation (Ferreira *et al.*, 2000; Kawabe *et al.*, 2021; Kieper & Jameson, 1999; Seddon & Zamoyska, 2002). Human naive T cell repertoires are estimated to consist of approximately 4×10^6–10^8 different TCR families (Arstila *et al.*, 1999; Nikolich-Žugich *et al.*, 2004; Qi *et al.*, 2014b; Robins *et al.*, 2009). This diversity implies that each TCR must be present on more than a single T cell. These sub-populations of T cells sharing the same TCR molecular structure are called T cell clonotypes. Previous deterministic models of T cell populations dynamics during infection, such as that proposed by De Boer & Perelson (1994), have shown that the principle of competitive exclusion applies to T cell populations; that is, two different clonotypes cannot inhabit the same stimulus niche, as one will always out-compete the other and cause it to become extinct.

4. MULTI-VARIATE MODEL OF T CELL CLONOTYPE COMPETITION AND HOMEOSTASIS

If a single naive T cell clonotype is considered, such that it experiences relatively little competition for self-pMHCs with other clonotypes, then its population dynamics can be modelled as a uni-variate birth and death process as was done by Stirk *et al.* (2008). Thus, for these TCR clonotypes self-pMHC stimulation promotes their homeostatic establishment in the periphery; that is, once they exit the thymus (Boyman *et al.*, 2012; Rudd *et al.*, 2011; Sprent & Surh, 2011; Surh & Sprent, 2005). The mathematical model of Stirk *et al.* also shows that clonotypes are susceptible to extinction events. This suggests that there must be an overlap in the sets of self-pMHCs that stimulate different T cell clonotypes. Otherwise the natural extinction of a clonotype would decrease the coverage of the T cell repertoire over the space of foreign peptides, which is known to be maintained even in the presence of such extinction events (Correia-Neves *et al.*, 2001; Naylor *et al.*, 2005). A similar mathematical model can be used for two clonotypes which compete for self-pMHC survival stimuli. In this case, a bi-variate Markov competition process can be defined as in (Stirk *et al.*, 2010). This bi-variate model can be used to show that extinction is certain for both clones for sufficiently late times, *i.e.*, after a transient time one clonotype will become extinct and the remaining one can be described by the uni-variate model (Stirk *et al.*, 2008). This is a closer representation of the competition for survival stimuli experienced by the naive T cell repertoire. However, the highly oligoclonal nature of immune responses (MacDonald *et al.*, 1993) and the occurrence of similar TCRs (Wynn *et al.*, 2008) serve as evidence that the self-pMHC recognition profile overlap will typically extend to more than two clonotypes.

In this chapter, a generalisation of the model presented in Stirk *et al.* (2010) is proposed in order to characterise the competition of η different T cell clonotypes ($\eta \geq 3$) with non-negligible self-pMHC recognition profile overlap. This means that the number of peptides shared by the clonotypes under consideration is large enough that their competition cannot be modelled as a single clonotype (Stirk *et al.*, 2008), nor as multiple competitions between two different clonotypes (Stirk *et al.*, 2010). It is assumed that naive T cells of a given clonotype exit the thymus at roughly the same time (Lythe *et al.*, 2016). After this point they are not generated again by the thymus, given the potential diversity of recombination (Murugan *et al.*, 2012; Zarnitsyna *et al.*, 2013). Thus, the population dynamics of a given

naive clonotype in the periphery depends on its homeostatic birth and death rates, and its extinction is possible. In addition, in this model extinction of any clonotype is certain for sufficiently late times, and thus, some time after its thymic output into the periphery, there will be one fewer clonotype competing for homeostatic proliferation stimuli. Mathematically, this decrease in the number of competing clonotypes will continue until two remain, and finally until only one remains, taking us back to the models described in Stirk *et al.* (2008, 2010).

The main interest in this chapter is the study of perturbations of homeostatically established clonotypes in the periphery, specifically by the introduction of a new clonotype that competes with them. The dynamics of competition before the extinction of the first clonotype is studied, as well as the population distribution after the first extinction event. This is of special importance if the first extinction event corresponds to the clonotype that most recently arrived in the periphery, as it informs us on how its introduction perturbs already homeostatically established clonotypes at both short (before extinction), and long (after extinction) timescales.

In Section 4.1, the competition model which describes the population of η different naive T cell clonotypes is introduced, as well as the recognition network of self-pMHCs used to calculate their homeostatic proliferation (birth) rates. Two special cases of competition with clonotypes in the periphery are also described in this section. Section 4.2 focuses on the quasi-stationary probability distribution (QSD), which is approximated stochastically with two different processes, one where extinction is not possible, and another where each clonotype has one immortal cell (Nåsell, 1991, 2001), and deterministically using the linear noise approximation (Elf & Ehrenberg, 2003; van Kampen, 2007). In Section 4.3 it is proven that for sufficiently late (but finite) times, all clonotypes will become extinct. Also, the stochastic descriptors used to study the behaviour of the competition around these extinction events are defined. Finally, in Section 4.4 the approximations of the QSD together with the stochastic descriptors defined in Section 4.3 are used to study the perturbation exerted on two established clonotypes by a new clonotype entering the periphery. For this, four different competition scenarios for the three clonotypes, and three different values for the homeostatic proliferation stimuli available are considered. Furthermore, the effects

of the new clonotype being in both special cases discussed in Section 4.1 are also studied.

4.1 Stochastic model of multiple naive T cell clonotype competition for self-pMHC stimuli

Consider the following two sets: the set $\mathscr{C}$, of η different clonotypes with a significant overlap in the self-pMHCs they recognise, and $\mathscr{Q}$, the set of all self-pMHCs which can stimulate clonotypes in $\mathscr{C}$. The number of T cells belonging to each of the η clonotypes at time t is described by a continuous-time multivariate Markov process, $\mathcal{X} = \{(X_1(t), \ldots, X_\eta(t)) : t \geq 0\}$, over the state space $\mathcal{S} = \{(n_1, \ldots, n_\eta) : n_i \geq 0, \forall i\} = \mathbb{N}_0^\eta$, where $X_i(t)$ represents the number of cells of clonotype i at time t (for $1 \leq i \leq \eta$), and $\mathbf{X}(t) = (X_1(t), \ldots, X_\eta(t))$ is the random vector describing the population of all clonotypes being modelled at time t.

Now, consider the following assumption of the model: all cells of a particular clonotype exit the thymus at roughly the same time. However, different clonotypes can exit the thymus at different times. Since the main interest of this chapter is to model the competition dynamics of all clonotypes in $\mathscr{C}$, the initial time $t = 0$ in the process $\mathcal{X}$ is such that all clonotypes in $\mathscr{C}$ are already present in the periphery.

The birth rate of a clonotype is defined so that it accounts for the competition between clonotypes for shared self-pMHC stimuli. To this end a bipartite recognition network is considered (see Figure 4.1). In a bipartite recognition network each clonotype (green circle) is able to receive stimuli from a set of self-pMHCs (blue circles), and this ability is represented by an edge between the clonotype and the self-pMHC. The set of all peripheral naive T cell clonotypes is partitioned as follows: clonotypes in the periphery are in $\mathscr{C}$, if they are explicitly modelled, or in $\mathcal{M}$, if they are not explicitly modelled. Note that this definition implies $\mathscr{C} \cap \mathcal{M} = \emptyset$. Each clonotype $i \in \mathscr{C}$ has an associated set of self-pMHCs that stimulate it, denoted by $\mathscr{Q}_i$ (see Figure 4.1).

All self-pMHCs are assumed to provide the same rate of homeostatic proliferation stimulus, γ (Stirk *et al.*, 2008, 2010), which is considered to be constant in time (van den Berg *et al.*, 2001). Then, the total homeostatic proliferation

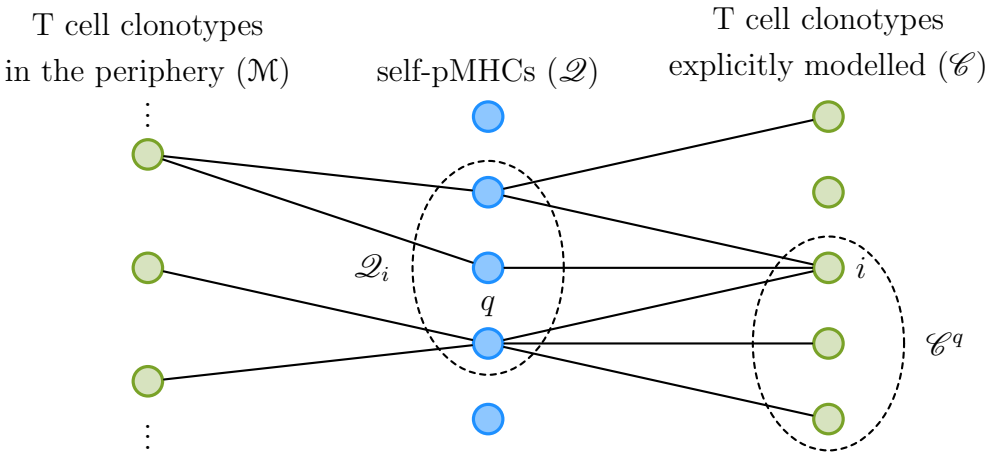


Figure 4.1: Bipartite recognition network of TCR-self-pMHC recognition. Each blue circle represents a self-pMHC and each green circle a T cell clonotype. A clonotype is explicitly modelled if it is in $\mathscr{C}$, or part of the periphery (and not explicitly modelled) if it is in $\mathcal{M}$. This implies $\mathscr{C} \cap \mathcal{M} = \emptyset$. An edge between a blue and a green circle represents the ability of that T cell clonotype to receive stimulus from the self-pMHC. For a given self-pMHC, $q \in \mathcal{Q}$, the set of clonotypes it can stimulate in $\mathscr{C}$ is $\mathscr{C}^q$, and for a given clonotype, $i \in \mathscr{C}$, the set of self-pMHCs that can stimulate it is $\mathcal{Q}_i$.

stimulus each naive T cell of clonotype i receives, if the system of η clonotypes is in state $\mathbf{n} = (n_1, \ldots, n_\eta) \in \mathcal{S}$, can be written as follows

$$\Lambda^{(i)}(\mathbf{n}) = \sum_{q \in \mathcal{Q}_i} \frac{\gamma}{h_q}, \tag{4.1}$$

where h_q is the total number of naive T cells in the periphery ($\mathscr{C} \cup \mathcal{M}$) that are stimulated by self-pMHC q (see Figure 4.1). Now, using this stimulus rate, the birth rate of clonotype i in state $\mathbf{n} \in \mathcal{S}$ is defined as the transition rate from state $\mathbf{n}$ to state $\mathbf{n}^{(+i)} := (n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_\eta)$, and it is given by

$$\lambda_{\mathbf{n}}^{(i)} = n_i \Lambda^{(i)}(\mathbf{n}). \tag{4.2}$$

Similarly, the transition rate from state $\mathbf{n} \in \mathcal{S}$ to state $\mathbf{n}^{(-i)} := (n_i, \ldots, n_{i-1}, n_i - 1, n_{i+1}, \ldots, n_\eta)$ is the death rate of clonotype i, and it is given by

$$\mu_{\mathbf{n}}^{(i)} = \mu_i \, n_i. \tag{4.3}$$

If $n_i = 0$ for any clonotype i, with $1 \leq i \leq \eta$, its birth and death rates will both be zero, in agreement with the assumption that thymic production cannot reproduce already extinct clonotypes. Thus, the set of states with at least one entry equal to zero, $\mathcal{A} = \{(n_1, \ldots, n_\eta) \in \mathcal{S} : n_i = 0 \text{ for any } i\}$, with $1 \leq i \leq \eta$, is an absorbing set, and the state $(0, \ldots, 0)$ is an absorbing state representing the extinction of all η clonotypes (Allen, 2010; Howard M & Karlin, 1998).

Consider two states, $\mathbf{n} = (n_1, \ldots, n_\eta)$ and $\mathbf{m} = (m_1, \ldots, m_\eta)$ in $\mathcal{S}$. The transition probability from $\mathbf{n}$ to $\mathbf{m}$ in a small time interval, Δt, is defined as

$$p_{\mathbf{nm}}(\Delta t) = \mathbb{P}\left(\mathbf{X}(t + \Delta t) = \mathbf{m} \mid \mathbf{X}(t) = \mathbf{n}\right),$$

and in the limit $\Delta t \to 0^+$, this transition probability satisfies

$$p_{\mathbf{nm}}(\Delta t) = \begin{cases} \lambda_{\mathbf{n}}^{(i)} \Delta t + o(\Delta t), & \text{if } \mathbf{m} = \mathbf{n}^{(+i)}, \\ \mu_{\mathbf{n}}^{(i)} \Delta t + o(\Delta t), & \text{if } \mathbf{m} = \mathbf{n}^{(-i)}, \\ 1 - \sum_{i=1}^{\eta} (\lambda_{\mathbf{n}}^{(i)} + \mu_{\mathbf{n}}^{(i)}) \Delta t + o(\Delta t), & \text{if } \mathbf{m} = \mathbf{n}, \\ o(\Delta t), & \text{otherwise.} \end{cases}$$

The clonotypes in $\mathscr{C}$ are explicitly modelled by the process $\mathcal{X}$, yet as mentioned before, there are other clonotypes in the naive T cell repertoire which can also receive stimuli from self-pMHCs in $\mathscr{Q}$, but which do not overlap significantly with the clonotypes in $\mathscr{C}$, namely those in $\mathcal{M}$. These clonotypes are contributing to the competition for stimuli as a "sink", in the sense that they are taking a portion of the stimuli, but their population dynamics is not explicitly modelled. The cardinality of $\mathcal{M}$ will be denoted by M; that is, $|\mathcal{M}| \coloneqq M$. Now, define $\mathscr{C}^q$ as the set of clonotypes in $\mathscr{C}$ which can receive stimuli from self-pMHC $q \in \mathscr{Q}$, this allows for the separation of h_q into the number of cells in $\mathscr{C}^q$ which receive stimuli from self-pMHC q, and the number of cells in $\mathcal{M}$ which receive stimuli from self-pMHC q. Since cells in $\mathcal{M}$ are not explicitly modelled, their populations are assumed to be in homeostatic steady state (Stirk *et al.*, 2008), and thus, have a constant size (see Eq. (4.12)).

4.1.1 Approximation of the transition rates

First, Eq. (4.1) will be re-written making use of the following definition: h_{iq} is the number of cells that are not of clonotype i and receive stimuli from self-pMHC q;

that is, $h_{iq} \coloneqq h_q - n_i$. Then, Eq. (4.1) becomes

$$\Lambda^{(i)}(\mathbf{n}) = \sum_{q \in \mathscr{Q}_i} \frac{\gamma}{n_i + h_{iq}}. \tag{4.4}$$

By writing the birth rate in this manner it is easy to see that it depends not only on clonotype i, but on all other clonotypes which compete for stimuli from self-pMHCs in $\mathscr{Q}_i$.

For a fixed clonotype i, define $\mathscr{C}_i$ to be the set of all clonotypes in $\mathscr{C}$ except for i, $\mathscr{C}_i \coloneqq \mathscr{C} \setminus \{i\}$. Now, consider the subsets I_{ij} of the power set, $\mathcal{P}(\mathscr{C}_i)$, which consist of all elements with cardinality j; that is, I_{ij} contains all the possible subsets of $\mathscr{C}$ with j elements which do not contain clonotype i. Define I_{ij}^k as the k-th element of I_{ij} under the lexicographical order. Note that $|I_{ij}| = \binom{\eta-1}{j}$, therefore k can only take the values $k = 1, 2, \ldots, \binom{\eta-1}{j}$. As an illustrative example consider the set $\mathscr{C}$ of clonotypes presented in Figure 4.1, with labels $i-2$, $i-1$, i, $i+1$, $i+2$ from top to bottom. Then, for $j = 3$, the sets in $I_{i,3}$ are

$$\begin{aligned} I_{i,3} &= \left\{ I_{i,3}^1, I_{i,3}^2, I_{i,3}^3, I_{i,3}^4 \right\} \\ &= \Big\{ \{i-2, i-1, i+1\}, \{i-2, i-1, i+2\}, \\ &\qquad \{i-2, i+1, i+2\}, \{i-1, i+1, i+2\} \Big\}. \end{aligned} \tag{4.5}$$

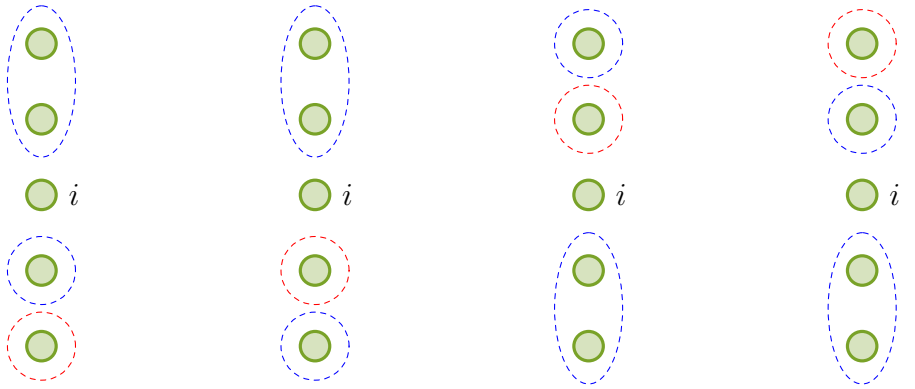
The sets described in this example are shown in Figure 4.2.

These sets allow for the partition of $\mathscr{Q}_i$, the set of self-pMHCs that stimulate clonotype i, into sets, $\mathscr{Q}_{ij}^k$, of self-pMHCs that stimulate precisely clonotype i and those in I_{ij}^k. These $\mathscr{Q}_{ij}^k$ sets are defined as follows

$$\mathscr{Q}_{ij}^k = \underbrace{\mathscr{Q}_i}_{\substack{\text{stimulates} \\ \text{clonotype } i}} \cap \underbrace{\left(\bigcap_{l \in I_{ij}^k} \mathscr{Q}_l \right)}_{\substack{\text{stimulates all} \\ \text{clonotypes in } I_{ij}^k}} \cap \underbrace{\left(\bigcap_{l \in \overline{I_{ij}^k}} \overline{\mathscr{Q}_l} \right)}_{\substack{\text{does not stimulate} \\ \text{any other clonotypes}}}, \tag{4.6}$$

where $\overline{I_{ij}^k}$ is the complement of I_{ij}^k in $\mathscr{C}_i$ and $\overline{\mathscr{Q}_i}$ is the complement of $\mathscr{Q}_i$ in $\mathscr{Q}$. By construction the sets $\mathscr{Q}_{ij}^k$ are disjoint and their union is $\mathscr{Q}_i$; that is

$$\mathscr{Q}_i = \bigsqcup_{j=0}^{\eta-1} \bigsqcup_{k=1}^{\binom{\eta-1}{j}} \mathscr{Q}_{ij}^k, \tag{4.7}$$



(a) $I^1_{i,3}$ **(b)** $I^2_{i,3}$ **(c)** $I^3_{i,3}$ **(d)** $I^4_{i,3}$

Figure 4.2: Example of the I^k_{ij} sets for $\eta = 5$ and $i = 3$. For a given clonotype i, the elements of $I_{i,3}$ are ordered sets of three clonotypes in $\mathscr{C}$ different to i. These elements are named I^k_{ij}, where k denotes their position in $I_{i,3}$ under the lexicographical order of their elements. The complement of these sets in $\mathscr{C}_i = \mathscr{C} \setminus \{i\}$ is denoted by $\overline{I^k_{i,3}}$. In this illustrative example the set $\mathscr{C}$ consists of the clonotypes $i-2$, $i-1$, i, $i+1$, and $i+2$, wich are shown in increasing order from top to bottom. Figures (a), (b), (c), and (d) show to the sets defined in Eq. (4.5) in blue, and their complement $\mathscr{C}_i$ in red.

where $\bigsqcup$ denotes the disjoint union of sets. With this partition, Eq. (4.4) can be re-written as the sum of stimuli from self-pMHCs that are shared with $j = 0, 1, \ldots, \eta - 1$ other clonotypes. This allows for a better understanding of the competition for self-pMHCs between the η different clonotypes from Eq. (4.4), by separating the competition into several competitions with different subsets of clonotypes. The resulting equation for the stimulus received by a cell of clonotype i is

$$\Lambda^{(i)}(\mathbf{n}) = \gamma \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \sum_{q \in \mathscr{Q}^k_{ij}} \Lambda^k_{ijq}(\mathbf{n}), \tag{4.8}$$

where $\Lambda^k_{ijq}(\mathbf{n})$ is the fraction of stimulus provided to clonotype i by a self-pMHC q that it shares only with clonotypes in I^k_{ij}, which is given by

$$\Lambda^k_{ijq}(\mathbf{n}) = \frac{1}{n_i + \sum\limits_{l \in I^k_{ij}} n_l + h^k_{ijq}},$$

with h^k_{ijq} defined as the number of cells not in $\mathscr{C}$ that receive stimulus from $q \in \mathscr{Q}^k_{ij}$; that is, $h^k_{ijq} \coloneqq h_q - n_i - \sum_{l\in I^k_{ij}} n_l$. The product of $\Lambda^k_{ijq}(\mathbf{n})$ and γ is the homeostatic proliferation stimulus provided to clonotype i by self-pMHC q, which also stimulates all clonotypes in I^k_{ij}. Now, the $\mathscr{Q}^k_{ij}$ sets can be subdivided further by considering the number of clonotypes not in $\mathscr{C}$ that can receive stimulus from $q \in \mathscr{Q}^k_{ij}$. Let $\mathscr{Q}^k_{ijr}$ with $r = 0, 1, \ldots, M$ denote the set of self-pMHCs that stimulate clonotype i, the clonotypes in I^k_{ij}, and r other clonotypes in $\mathcal{M}$. With these sets Eq. (4.8) can be written as

$$\Lambda^{(i)}(\mathbf{n}) = \gamma \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \sum_{r=0}^{M} \sum_{q\in\mathscr{Q}^k_{ijr}} \Lambda^k_{ijq}(\mathbf{n}). \tag{4.9}$$

To simplify Eq. (4.9), and for practical purposes, a mean field approximation will be used, as discussed in Stirk *et al.* (2008). First, define the following quantities, mean and variance of h^k_{ijq} over the set of $q \in \mathscr{Q}^k_{ijr}$:

$$\mathbb{E}^k_{ijr}\left[h^k_{ijq}\right] \coloneqq \frac{1}{|\mathscr{Q}^k_{ijr}|} \sum_{\ell\in\mathscr{Q}^k_{ijr}} n^k_{ij\ell}, \tag{4.10a}$$

$$\mathbb{V}^k_{ijr}\left[h^k_{ijq}\right] \coloneqq \frac{1}{|\mathscr{Q}^k_{ijr}|} \sum_{\ell\in\mathscr{Q}^k_{ijr}} \left(n^k_{ij\ell} - \mathbb{E}^k_{ijr}\left[h^k_{ijq}\right]\right)^2, \tag{4.10b}$$

respectively. Then, making use of Eq. (4.10a), Eq. (4.10b), the Taylor series of $\Lambda^k_{ijq}(\mathbf{n})$ as a function of h^k_{ijq}, and the properties of the expected value (Bain & Engelhardt, 2000, Chapter 2), the sum of $\Lambda^k_{ijq}(\mathbf{n})$ can be re-written as follows

$$\begin{aligned}\sum_{q\in\mathscr{Q}^k_{ijr}} \Lambda^k_{ijq}(\mathbf{n}) = |\mathscr{Q}^k_{ijr}| \Bigg(& \frac{1}{n_i + \sum_{l\in I^k_{ij}} n_l + \mathbb{E}^k_{ijr}\left[h^k_{ijq}\right]} \\ & + \frac{\mathbb{V}^k_{ijr}\left[h^k_{ijq}\right]}{\left(n_i + \sum_{l\in I^k_{ij}} n_l + \mathbb{E}^k_{ijr}\left[h^k_{ijq}\right]\right)^3} + \cdots \Bigg).\end{aligned} \tag{4.11}$$

The first assumption of the mean field approximation is that the first term in Eq. (4.11) dominates the sum. This assumption stems from the fact that in

carrying out this approximation, $\mathbb{V}^k_{ijr}\left[h^k_{ijq}\right]$ is considered to be small. Since it appears as a factor of the second and subsequent terms, they are considered to be small as well. The second assumption is that given r, the mean number of cells per clonotype in $\mathcal{M}$ competing with i and all clonotypes in I^k_{ij} for stimuli from self-pMHC $q \in \mathscr{Q}^k_{ijr}$ is the same as the average clonotype size; that is

$$\mathbb{E}^k_{ijr}\left[h^k_{ijq}\right] = r\langle n\rangle, \tag{4.12}$$

where $\langle n\rangle$ is the average clonotype size in $\mathcal{M}$ (Stirk *et al.*, 2008). From Eq. (4.12) and the assumptions from the mean field approximation, Eq. (4.11) is approximated as

$$\sum_{q\in\mathscr{Q}^k_{ijr}} \Lambda^k_{ijq}(\mathbf{n}) \approx \frac{|\mathscr{Q}^k_{ijr}|}{n_i + \sum\limits_{l\in I^k_{ij}} n_l + r\langle n\rangle}. \tag{4.13}$$

The next step is to find an expression for $|\mathscr{Q}^k_{ijr}|$, but in order to do so, an expression for $|\mathscr{Q}^k_{ij}|$ needs to be found first. Let p^k_{ij} denote the probability that a randomly chosen self-pMHC in $\mathscr{Q}_i$ will provide stimuli to all clonotypes in I^k_{ij}. This probability can be defined in terms of the cardinality of the $\mathscr{Q}^k_{ij}$ sets as follows

$$p^k_{ij} = \frac{|\mathscr{Q}^k_{ij}|}{|\mathscr{Q}_i|},$$

and from Eq. (4.7) it follows that

$$\sum_{j=0}^{\eta-1}\sum_{k=1}^{\binom{\eta-1}{j}} p^k_{ij} = 1.$$

Now, consider the probability $p_{\cdot|ijk}$ that a clonotype in $\mathcal{M}$ is stimulated by a self-pMHC in $\mathscr{Q}^k_{ij}$. That is, any clonotype chosen at random from $\mathcal{M}$ is stimulated by a self-pMHC in $\mathscr{Q}^k_{ij}$, with probability $p_{\cdot|ijk}$. It is clear that the number of self-pMHCs in $\mathscr{Q}^k_{ijr}$ follows a binomial distribution, Binomial$(M, p_{\cdot|ijk})$, since it is the sum of M independent Bernoulli experiments, each with probability $p_{\cdot|ijk}$. Thus, the cardinality of $\mathscr{Q}^k_{ijr}$ is given by

$$|\mathscr{Q}^k_{ijr}| = |\mathscr{Q}^k_{ij}|\binom{M}{r}\left(p_{\cdot|ijk}\right)^r\left(1-p_{\cdot|ijk}\right)^{M-r}.$$

Under the assumption that the number of modelled clonotypes is small compared to the total number of clonotypes, *i.e.*, $\eta \ll M$, the Poisson approximation can be used by introducing the parameter $\nu_{ij}^k = Mp_{\cdot|ijk}$, so that

$$|\mathscr{Q}_{ijr}^k| \approx p_{ij}^k |\mathscr{Q}_i| \frac{\left(\nu_{ij}^k\right)^r e^{-\nu_{ij}^k}}{r!}. \tag{4.14}$$

The parameter ν_{ij}^k is called the mean niche overlap for self-pMHCs that stimulate clonotype i and all clonotypes in I_{ij}^k (Stirk *et al.*, 2008). Let $\varphi_i = \gamma|\mathscr{Q}_i|$, then, using Eq. (4.13), and Eq. (4.14) the following approximation of Eq. (4.9) is obtained

$$\begin{aligned}\Lambda^{(i)}(\mathbf{n}) &\approx \gamma \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \sum_{r=0}^{M} p_{ij}^k |\mathscr{Q}_i| \frac{\left(\nu_{ij}^k\right)^r e^{-\nu_{ij}^k}}{r!} \frac{1}{n_i + \sum\limits_{l \in I_{ij}^k} n_l + r\langle n \rangle} \\ &= \varphi_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} p_{ij}^k e^{-\nu_{ij}^k} \sum_{r=0}^{M} \frac{\left(\nu_{ij}^k\right)^r}{r!} \frac{1}{n_i + \sum\limits_{l \in I_{ij}^k} n_l + r\langle n \rangle}.\end{aligned} \tag{4.15}$$

Using this result, the birth rate of clonotype i, defined in Eq. (4.2), can be approximated as follows

$$\lambda_{\mathbf{n}}^{(i)} = n_i \Lambda^{(i)}(\mathbf{n}) \approx \varphi_i n_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} p_{ij}^k e^{-\nu_{ij}^k} \sum_{r=0}^{M} \frac{\left(\nu_{ij}^k\right)^r}{r!} \frac{1}{n_i + \sum\limits_{l \in I_{ij}^k} n_l + r\langle n \rangle}. \tag{4.16}$$

Also, note that from the definition of φ_i it follows that

$$\gamma|\mathscr{Q}_{ij}^k| = \gamma p_{ij}^k |\mathscr{Q}_i| = p_{ij}^k \varphi_i.$$

Then, for any pair of clonotypes i, i' in $\mathscr{C}$ and a pair of sets I_{ij}^k, $I_{i'j}^{k'}$, such that $I_{ij}^k \cup \{i\} = I_{i'j}^{k'} \cup \{i'\}$, the following relation holds

$$\varphi_i p_{ij}^k = \varphi_{i'} p_{i'j}^{k'}. \tag{4.17}$$

This constraint comes naturally from the fact that, while the sizes of $\mathscr{Q}_i$ and $\mathscr{Q}_{i'}$ can be different, the stimuli provided by $\mathscr{Q}_{ij}^k$ and $\mathscr{Q}_{i'j}^{k'}$ is the same if $I_{ij}^k \cup \{i\} = I_{i'j}^{k'} \cup \{i'\}$. Furthermore, for any pair of clonotypes i, i' in $\mathscr{C}$, and a pair of sets I_{ij}^k, $I_{i'j}^{k'}$, such

that $I_{ij}^k \cup \{i\} = I_{i'j}^{k'} \cup \{i'\}$, the following constraint for their mean niche overlap parameters holds

$$\nu_{ij}^k = \nu_{i'j}^{k'}. \tag{4.18}$$

This is because the mean niche overlap is a characteristic of self-pMHCs and not of clonotypes, and therefore its value depends only on the sets $\mathscr{Q}_{ij}^k$ and $\mathscr{Q}_{i'j}^{k'}$, and not on the specific clonotypes i and i'.

The birth and death rates of the process $\mathcal{X}$, which have been defined in Eq. (4.3) and Eq. (4.16), respectively, can be simplified for two limiting cases. The first one is that in which $\nu_{ij}^k \ll 1$ for all ν_{ij}^k. This is called the "hard niche" case, which is characterised by little competition with clonotypes in $\mathcal{M}$. In this case Eq. (4.16) simplifies to

$$\lambda_{\mathbf{n}}^{(i)} \approx \varphi_i n_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{ij}^k}{n_i + \sum\limits_{l \in I_{ij}^k} n_l}. \tag{4.19}$$

The second case, when $\nu_{ij}^k \gg 1$ for all ν_{ij}^k, is called the "soft niche" case, where there is greater competition with clonotypes in $\mathcal{M}$. In this case, Eq. (4.16) is approximated by

$$\lambda_{\mathbf{n}}^{(i)} \approx \varphi_i n_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{ij}^k}{n_i + \sum\limits_{l \in I_{ij}^k} n_l + \nu_{ij}^k \langle n \rangle}. \tag{4.20}$$

4.2 Quasi-stationary probability distribution

This section focuses on the study of the behaviour of η competing clonotypes before the first extinction event occurs. In order to do so, the quasi-stationary probability distribution (QSD) is used, which describes the late time behaviour of the process conditioned on non-extinction (Guillemin & Sericola, 2007; Nåsell, 1991; van Doorn & Scheinhardt, 1997).

4.2.1 Direct calculation of the quasi-stationary probability distribution

To derive a closed-form expression for the QSD, first consider the probability $p_{\mathbf{n}}(t)$ that at time t the competition process $\mathcal{X}$ is in state $\mathbf{n}$, given that it started in state $\mathbf{n}_0$, *i.e.*,

$$p_{\mathbf{n}}(t) = \mathbb{P}\left(\mathbf{X}(t) = \mathbf{n} \mid \mathbf{X}(0) = \mathbf{n}_0\right),$$

and note that these probabilities satisfy the Kolmogorov differential equations (Kolmogoroff, 1931)

$$\frac{\mathrm{d}p_{\mathbf{n}}(t)}{\mathrm{d}t} = \sum_{i=1}^{\eta} \lambda^{(i)}_{\mathbf{n}^{(-i)}} p_{\mathbf{n}^{(-i)}}(t) + \sum_{i=1}^{\eta} \mu^{(i)}_{\mathbf{n}^{(+i)}} p_{\mathbf{n}^{(+i)}}(t) - \sum_{i=1}^{\eta} \left(\lambda^{(i)}_{\mathbf{n}} + \mu^{(i)}_{\mathbf{n}}\right) p_{\mathbf{n}}(t). \tag{4.21}$$

Define $\mathcal{A}$ to be the absorbing set $\{(n_1, \ldots, n_\eta) : n_i = 0 \text{ for any } i\}$, and denote by $p_{\overline{\mathcal{A}}}(t)$ the probability that at time t the process is not in $\mathcal{A}$. Now, define the probability that the process is in state $\mathbf{n} \in \mathcal{S} \setminus \mathcal{A}$ at time t given that absorption into $\mathcal{A}$ has not occurred yet, as follows

$$g_{\mathbf{n}}(t) \coloneqq \mathbb{P}\left(\mathbf{X}(t) = \mathbf{n} \mid \mathbf{X}(t) \notin \mathcal{A}\right) = \frac{p_{\mathbf{n}}(t)}{p_{\overline{\mathcal{A}}}(t)}. \tag{4.22}$$

From the definition of $g_{\mathbf{n}}(t)$ in Eq. (4.22) the following system of differential equations can be derived

$$\begin{aligned} \frac{\mathrm{d}g_{\mathbf{n}}(t)}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\frac{p_{\mathbf{n}}(t)}{p_{\overline{\mathcal{A}}}(t)} \\ &= \frac{1}{p_{\overline{\mathcal{A}}}(t)}\frac{\mathrm{d}p_{\mathbf{n}}(t)}{\mathrm{d}t} - \frac{g_{\mathbf{n}}(t)}{p_{\overline{\mathcal{A}}}(t)}\frac{\mathrm{d}p_{\overline{\mathcal{A}}}(t)}{\mathrm{d}t}. \end{aligned} \tag{4.23}$$

Then, given that the $p_{\mathbf{n}}(t)$ probabilities follow the Kolmogorov equations, see Eq. (4.21), the first term of Eq. (4.23) is re-written as follows

$$\begin{aligned} \frac{1}{p_{\overline{\mathcal{A}}}(t)}\frac{\mathrm{d}p_{\mathbf{n}}(t)}{\mathrm{d}t} = &\sum_{i=1}^{\eta} \lambda^{(i)}_{\mathbf{n}^{(-i)}} g_{\mathbf{n}^{(-i)}}(t) + \sum_{i=1}^{\eta} \mu^{(i)}_{\mathbf{n}^{(+i)}} g_{\mathbf{n}^{(+i)}}(t) \\ &- \sum_{i=1}^{\eta} \left(\lambda^{(i)}_{\mathbf{n}} + \mu^{(i)}_{\mathbf{n}}\right) g_{\mathbf{n}}(t). \end{aligned} \tag{4.24}$$

In order to solve the second term of Eq. (4.23), the law of total probability (Howard M & Karlin, 1998, Chapter 1) is used first to rewrite $p_{\overline{\mathcal{A}}}(t)$ as

$$p_{\overline{\mathcal{A}}}(t) = 1 - \sum_{\mathbf{n}\in\mathcal{A}} p_{\mathbf{n}}(t). \tag{4.25}$$

However, since the competition process under consideration is conditioned on non-extinction, the boundary is the set of states where only one clonotype has become extinct. Thus, when calculating $\frac{\mathrm{d}p_{\overline{\mathcal{A}}}(t)}{\mathrm{d}t}$ the only relevant transitions are those to states in which only one clonotype has become extinct. Then, $\mathcal{A}_i$ is defined as the set of states where only clonotype i has gone extinct; that is, for $1 \leq i \leq \eta$

$$\mathcal{A}_i \coloneqq \{\mathbf{n} \in \mathcal{S} : n_i = 0, n_k > 0 \quad \forall k \neq i \text{ with } 1 \leq k \leq \eta\}. \tag{4.26}$$

For the sake of computing $p_{\overline{\mathcal{A}}}(t)$, one can consider that states in $\mathcal{A}_i$ are absorbing for all $1 \leq i \leq \eta$, and then rewrite Eq. (4.25) as

$$p_{\overline{\mathcal{A}}}(t) = 1 - \sum_{i=1}^{\eta} \sum_{\mathbf{n}\in\mathcal{A}_i} p_{\mathbf{n}}(t). \tag{4.27}$$

Taking the derivative of Eq. (4.27) with respect to t results in

$$\frac{\mathrm{d}p_{\overline{\mathcal{A}}}(t)}{\mathrm{d}t} = -\sum_{i=1}^{\eta} \sum_{\mathbf{n}\in\mathcal{A}_i} \mu^{(i)}_{\mathbf{n}^{(+i)}} p_{\mathbf{n}^{(+i)}}(t). \tag{4.28}$$

Now, making use of Eq. (4.22) and Eq. (4.28), the second term of Eq. (4.23) can be written as

$$\frac{g_{\mathbf{n}}(t)}{p_{\overline{\mathcal{A}}}(t)} \frac{\mathrm{d}p_{\overline{\mathcal{A}}}(t)}{\mathrm{d}t} = -g_{\mathbf{n}}(t) \sum_{i=1}^{\eta} \sum_{\mathbf{m}\in\mathcal{A}_i} \mu^{(i)}_{\mathbf{m}^{(+i)}} g_{\mathbf{m}^{(+i)}}(t). \tag{4.29}$$

Note that the only death events considered are those of clonotypes of size one. Then, by Eq. (4.3), Eq. (4.29) can be further simplified to obtain

$$\frac{g_{\mathbf{n}}(t)}{p_{\overline{\mathcal{A}}}(t)} \frac{\mathrm{d}p_{\overline{\mathcal{A}}}(t)}{\mathrm{d}t} = -g_{\mathbf{n}}(t) \sum_{i=1}^{\eta} \sum_{\mathbf{m}\in\mathcal{A}_i} \mu_i g_{\mathbf{m}^{(+i)}}(t). \tag{4.30}$$

Finally, substituting Eq. (4.24) and Eq. (4.30) in Eq. (4.23) the following differential equation is obtained

$$\begin{aligned}\frac{\mathrm{d}g_{\mathbf{n}}(t)}{\mathrm{d}t} =& \sum_{i=1}^{\eta} \lambda^{(i)}_{\mathbf{n}^{(-i)}} g_{\mathbf{n}^{(-i)}}(t) + \sum_{i=1}^{\eta} \mu^{(i)}_{\mathbf{n}^{(+i)}} g_{\mathbf{n}^{(+i)}}(t) \\ &- \sum_{i=1}^{\eta} \left(\lambda^{(i)}_{\mathbf{n}} + \mu^{(i)}_{\mathbf{n}}\right) g_{\mathbf{n}}(t) + g_{\mathbf{n}}(t) \sum_{i=1}^{\eta} \sum_{\mathbf{m}\in\mathcal{A}_i} \mu_i g_{\mathbf{m}^{(+i)}}(t).\end{aligned}$$

The limiting conditional distribution (LCD) of the process is the limit as $t \to +\infty$ of $g_{\mathbf{n}}(t)$ (Darroch & Seneta, 1967). If this limit exists, the resulting probability distribution is a QSD of the process, which in the case of a process with finite state space is unique and therefore equal to the LCD (Darroch & Seneta, 1967). Now, if there exists a probability distribution $\tilde{g}$ such that $\sum_{\mathbf{n}\in\mathcal{S}\setminus\mathcal{A}} \tilde{g}_{\mathbf{n}} = 1$, that also satisfies

$$\begin{aligned}0 =& \sum_{i=1}^{\eta} \lambda^{(i)}_{\mathbf{n}^{(-i)}} \tilde{g}_{\mathbf{n}^{(-i)}} + \sum_{i=1}^{\eta} \mu^{(i)}_{\mathbf{n}^{(+i)}} \tilde{g}_{\mathbf{n}^{(+i)}} \\ &- \sum_{i=1}^{\eta} \left(\lambda^{(i)}_{\mathbf{n}} + \mu^{(i)}_{\mathbf{n}}\right) \tilde{g}_{\mathbf{n}} + \tilde{g}_{\mathbf{n}} \sum_{i=1}^{\eta} \sum_{\mathbf{m}\in\mathcal{A}_i} \mu_i \tilde{g}_{\mathbf{m}^{(+i)}},\end{aligned} \tag{4.31}$$

then, it is a QSD of the process. Finding an analytical solution of Eq. (4.31) is in general not possible, and thus, the QSD will be numerically approximated. Two useful approximations are discussed in Section 4.2.2 (Nåsell, 1991, 2001).

4.2.2 Approximation of the quasi-stationary distribution: two auxiliary processes

The two auxiliary competition processes presented in Nåsell (2001) will be used to approximate the QSD in this section. In the first approximation the multivariate Markov process $\mathcal{X}^{(1)} = \left\{\left(X_1^{(1)}(t), \ldots, X_\eta^{(1)}(t)\right) : t \geq 0\right\}$ is considered, where $X_i^{(1)}(t)$ is the number of cells of clonotype i at time t. The birth rate of clonotype i in state $\mathbf{n} = (n_1, n_2, \ldots, n_\eta)$, $\lambda^{1,(i)}_{\mathbf{n}}$, is given by Eq. (4.16), and its death rate by

$$\mu^{1,(i)}_{\mathbf{n}} = \begin{cases} \mu_i n_i & \text{if } n_i > 1, \\ 0 & \text{if } n_i = 1. \end{cases}$$

Thus, the state space of $\mathcal{X}^{(1)}$ is the set of states where no extinction has occurred

$$\mathcal{A}^0 := \mathcal{S} \setminus \mathcal{A} = \{\mathbf{n} \in \mathcal{S} : n_i > 0 \text{ for all } 1 \leq i \leq \eta\}. \tag{4.32}$$

The second auxiliary process considered is $\mathcal{X}^{(2)} = \left\{\left(X_1^{(2)}(t), \ldots, X_\eta^{(2)}(t)\right) : t \geq 0\right\}$, where the birth rates, $\lambda_{\mathbf{n}}^{2,(i)}$, are the same as those for $\mathcal{X}$ and $\mathcal{X}^{(1)}$, and each clonotype is considered to have an immortal cell; that is, the death rates are given by

$$\mu_{\mathbf{n}}^{2,(i)} = \mu_i(n_i - 1),$$

and the state space of $\mathcal{X}^{(2)}$ is also $\mathcal{A}^0$, as defined in Eq. (4.32).

In order to approximate the QSD for the system of interest, the stationary probability distribution of the two auxiliary processes will be calculated following (Nåsell, 1991, 2001). To this end, an order on the set of non-absorbing states, $\mathcal{A}^0$, must be defined. First, separate $\mathcal{A}^0$ into levels of constant total cells

$$L^0(k) = \left\{(n_1, \ldots, n_\eta) : \sum_{i=1}^{\eta} n_i = k \text{ and } n_i > 0 \text{ for all } i\right\},$$

for $k = \eta, \eta+1, \eta+2, \ldots$. These levels can be ordered as follows

$$L^0(\eta) \prec L^0(\eta+1) \prec L^0(\eta+2) \prec L^0(\eta+3) \prec \cdots,$$

and the states in each level can be ordered using the colexicographical order, sometimes called reverse lexicographical order (Cameron, 1994). Note that the levels start at $L^0(\eta)$, since any state with fewer than η cells in total does not belong to $\mathcal{A}^0$ by definition. Using a combinatorial argument (Feller, 1957), it can be seen that the cardinality of $L^0(k)$ is given by

$$L_k^0 := |L^0(k)| = \binom{k-1}{\eta-1}. \tag{4.33}$$

Now, in order to truncate the state space, the plane $\sum_{i=1}^{\eta} n_i = N$ is introduced as a reflecting boundary on the complete state space $\mathcal{S}$. This means that only states which have at most N cells in total are considered. In practice, this truncation value N can be chosen so that the probability of exceeding a total number N of

cells in the population is negligible (see Section 4.3). Then, the number of states in $\mathcal{A}^0$ is

$$|\mathcal{A}^0| = \sum_{i=\eta}^{N} L_i^0 = \sum_{i=\eta}^{N} \binom{i-1}{\eta-1} = \binom{N}{\eta},$$

and the position of state $\mathbf{n} = (n_1, n_2, \ldots, n_\eta)$ in level $L^0(k)$ is given by

$$\begin{aligned} \mathrm{pos}_k(\mathbf{n}, \eta) =& \frac{1}{k-\eta} \left[\sum_{i=1}^{\eta} (n_i - 1)(k - \eta + 1)^{i-1} \right. \\ &\left. + \sum_{i=3}^{\eta} (n_i - 1)(1 - (k - \eta + 1)^{i-1}) \right] \\ &+ \sum_{i=3}^{\eta} \left[\binom{k - 1 - \sum_{j=i+1}^{\eta} n_j}{i-1} - \binom{k - \sum_{j=i}^{\eta} n_j}{i-1} \right], \end{aligned} \tag{4.34}$$

for $k > \eta \geq 3$. Since there is only one state in $L^0(\eta)$, $\mathbf{n} = (1, \ldots, 1)$, $\mathrm{pos}_\eta(\mathbf{n}, \eta)$ is defined to be 1 for $\mathbf{n} \in L^0(\eta)$. Then, from Eq. (4.33) and Eq. (4.34), the position in $\mathcal{A}^0$ of a state $\mathbf{n} \in L^0(k)$ is given by

$$\mathrm{pos}_{\mathcal{A}^0}(\mathbf{n}, \eta) = \mathrm{pos}_k(\mathbf{n}, \eta) + \sum_{\ell=\eta}^{k-1} L_\ell^0.$$

From any state $\mathbf{n} \in L^0(k)$ the process can only move to states in the adjacent levels $L^0(k-1)$ and $L^0(k+1)$. This means that the infinitesimal generator matrix of process $\mathcal{X}^{(j)}$, $\mathbf{Q}^{(j)}$ for $j = 1, 2$, is of quasi-birth-and-death type (Gómez-Corral & López-García, 2018; Kulkarni, 2017), and can be written as

$$\mathbf{Q}^{(j)} = \begin{bmatrix} \mathbf{A}^{(j)}_{\eta,\eta} & \mathbf{A}^{(j)}_{\eta,\eta+1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{A}^{(j)}_{\eta+1,\eta} & \mathbf{A}^{(j)}_{\eta+1,\eta+1} & \mathbf{A}^{(j)}_{\eta+1,\eta+2} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{(j)}_{\eta+2,\eta+1} & \mathbf{A}^{(j)}_{\eta+2,\eta+2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}^{(j)}_{N-1,N} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}^{(j)}_{N,N} \end{bmatrix}, \tag{4.35}$$

where $\mathbf{0}$ are zero matrices of the appropriate sizes, and the $\mathbf{A}^{(j)}_{k,k'}$ sub-matrices are defined as follows: First, let $\mathbf{n}_k^i$ denote the state in $L^0(k)$ such that $\mathrm{pos}_k(\mathbf{n}_k^i, \eta) = i$. Then, define the sub-matrices as

- for $\eta + 1 \leq k \leq N$

$$\left(\mathbf{A}^{(j)}_{k,k-1}\right)_{ip} = \begin{cases} \mu^{j,(\ell)}_{\mathbf{n}^i_k} & \text{if } \mathbf{n}^p_{k-1} = \mathbf{n}^{i(-\ell)}_k, \\ 0 & \text{otherwise.} \end{cases}$$

- for $\eta \leq k \leq N - 1$

$$\left(\mathbf{A}^{(j)}_{k,k+1}\right)_{ip} = \begin{cases} \lambda^{j,(\ell)}_{\mathbf{n}^i_k} & \text{if } \mathbf{n}^p_{k+1} = \mathbf{n}^{i(+\ell)}_k, \\ 0 & \text{otherwise.} \end{cases}$$

- for $\eta \leq k \leq N$

$$\left(\mathbf{A}^{(j)}_{k,k}\right)_{ip} = \begin{cases} -\left(\sum_{\ell=1}^{\eta} \lambda^{j,(\ell)}_{\mathbf{n}^i_k} + \mu^{j,(\ell)}_{\mathbf{n}^i_k}\right) & \text{if } i = p, \\ 0 & \text{otherwise.} \end{cases}$$

As previously mentioned, to approximate the QSD of the competition process the stationary distribution of the auxiliary processes will be calculated. The stationary distribution of a process with infinitesimal generator matrix $\mathbf{Q}^{(j)}$ is defined as a vector $\boldsymbol{\pi}^{(j)} = (\boldsymbol{\pi}^{(j)}_\eta, \ldots, \boldsymbol{\pi}^{(j)}_N)$ such that

$$\boldsymbol{\pi}^{(j)}\mathbf{Q}^{(j)} = \mathbf{0}, \tag{4.36a}$$

$$\boldsymbol{\pi}^{(j)}\mathbf{e} = 1, \tag{4.36b}$$

where each $\boldsymbol{\pi}^{(j)}_i$ is the vector of stationary probabilities for each state in $L(i)$. From Eq. (4.36a) and Eq. (4.36b) the following relations for the $\boldsymbol{\pi}^{(j)}_i$ are obtained

$$\begin{aligned}
\boldsymbol{\pi}^{(j)}_\eta \mathbf{A}^{(j)}_{\eta,\eta} + \boldsymbol{\pi}^{(j)}_{\eta+1}\mathbf{A}^{(j)}_{\eta+1,\eta} &= 0 \\
\boldsymbol{\pi}^{(j)}_\eta \mathbf{A}^{(j)}_{\eta,\eta+1} + \boldsymbol{\pi}^{(j)}_{\eta+1}\mathbf{A}^{(j)}_{\eta+1,\eta+1} + \boldsymbol{\pi}^{(j)}_{\eta+2}\mathbf{A}^{(j)}_{\eta+2,\eta+1} &= 0 \\
\boldsymbol{\pi}^{(j)}_{\eta+1}\mathbf{A}^{(j)}_{\eta+1,\eta+2} + \boldsymbol{\pi}^{(j)}_{\eta+2}\mathbf{A}^{(j)}_{\eta+2,\eta+2} + \boldsymbol{\pi}^{(j)}_{\eta+3}\mathbf{A}^{(j)}_{\eta+3,\eta+2} &= 0 \\
&\vdots \\
\boldsymbol{\pi}^{(j)}_{N-2}\mathbf{A}^{(j)}_{N-2,N-1} + \boldsymbol{\pi}^{(j)}_{N-1}\mathbf{A}^{(j)}_{N-1,N-1} + \boldsymbol{\pi}^{(j)}_N \mathbf{A}^{(j)}_{N,N-1} &= 0 \\
\boldsymbol{\pi}^{(j)}_{N-1}\mathbf{A}^{(j)}_{N-1,N} + \boldsymbol{\pi}^{(j)}_N \mathbf{A}^{(j)}_{N,N} &= 0 \\
\sum_{i=\eta}^{N}\sum_{j=1}^{T_{i-2}} \boldsymbol{\pi}^{(j)}_{i,j} &= 1.
\end{aligned} \tag{4.37}$$

By re-writing the penultimate equation in Eq. (4.37) one obtains

$$\boldsymbol{\pi}_N^{(j)} = -\boldsymbol{\pi}_{N-1}^{(j)} \mathbf{A}_{N-1,N}^{(j)} \left(\mathbf{A}_{N,N}^{(j)}\right)^{-1}. \tag{4.38}$$

Now, let $\mathbf{A}_{N,N}^{(j)} = \mathbf{H}_N^{(j)}$, then Eq. (4.38) becomes

$$\boldsymbol{\pi}_N^{(j)} = -\boldsymbol{\pi}_{N-1}^{(j)} \mathbf{A}_{N-1,N}^{(j)} \left(\mathbf{H}_N^{(j)}\right)^{-1}. \tag{4.39}$$

Replacing Eq. (4.39) in the preceding relation results in

$$\begin{aligned} 0 =& \boldsymbol{\pi}_{N-2}^{(j)} \mathbf{A}_{N-2,N-1}^{(j)} + \boldsymbol{\pi}_{N-1}^{(j)} \mathbf{A}_{N-1,N-1}^{(j)} - \boldsymbol{\pi}_{N-1}^{(j)} \mathbf{A}_{N-1,N}^{(j)} \left(\mathbf{H}_N^{(j)}\right)^{-1} \mathbf{A}_{N,N-1}^{(j)} \\ 0 =& \boldsymbol{\pi}_{N-2}^{(j)} \mathbf{A}_{N-2,N-1}^{(j)} + \boldsymbol{\pi}_{N-1}^{(j)} \underbrace{\left(\mathbf{A}_{N-1,N-1}^{(j)} - \mathbf{A}_{N-1,N}^{(j)} \left(\mathbf{H}_N^{(j)}\right)^{-1} \mathbf{A}_{N,N-1}^{(j)}\right)}_{\mathbf{H}_{N-1}^{(j)}}, \end{aligned}$$

then

$$\boldsymbol{\pi}_{N-1}^{(j)} = -\boldsymbol{\pi}_{N-2}^{(j)} \mathbf{A}_{N-2,N-1}^{(j)} \left(\mathbf{H}_{N-1}^{(j)}\right)^{-1}. \tag{4.40}$$

Continuing this recursive calculation results in the following expression for the $\mathbf{H}_k^{(j)}$ matrices

$$\mathbf{H}_k^{(j)} = \mathbf{A}_{k,k}^{(j)} - \mathbf{A}_{k,k+1}^{(j)} \left(\mathbf{H}_{k+1}^{(j)}\right)^{-1} \mathbf{A}_{k+1,k}^{(j)}, \tag{4.41}$$

and for the stationary probability vectors the following recursive expression is obtained

$$\boldsymbol{\pi}_k^{(j)} = -\boldsymbol{\pi}_{k-1}^{(j)} \mathbf{A}_{k-1,k}^{(j)} \left(\mathbf{H}_k^{(j)}\right)^{-1}. \tag{4.42}$$

Since the system of equations in Eq. (4.37) has $N - \eta + 2$ equations and $N - \eta + 1$ unknowns, it is over-determined when they are all considered, thus the first equation of the system will be ignored. This means that the only expression where $\boldsymbol{\pi}_\eta^{(j)}$ appears in is

$$\boldsymbol{\pi}_{\eta+1}^{(j)} = -\boldsymbol{\pi}_\eta^{(j)} \mathbf{A}_{\eta,\eta+1}^{(j)} \left(\mathbf{H}_{\eta+1}^{(j)}\right)^{-1}, \tag{4.43}$$

and it can be set to an arbitrary value. Since $L_\eta^0 = 1$ this value is a scalar, which is chosen to be $\boldsymbol{\pi}_\eta^{(j)*} = 1$. Then, Eq. (4.41) and Eq. (4.42) can be used to find

$\boldsymbol{\pi}_i^{(j)*}$ values, which are correct relative to each other. Finally, Eq. (4.36b) is used to normalise the $\boldsymbol{\pi}_i^{(j)*}$, to find the values of $\boldsymbol{\pi}_i^{(j)}$ as follows

$$\boldsymbol{\pi}_i^{(j)} = \frac{\boldsymbol{\pi}_i^{(j)*}}{\sum_{i=\eta}^{N}\sum_{j=1}^{L_i}\boldsymbol{\pi}_{i,j}^{(j)*}}. \tag{4.44}$$

This method for calculating the stationary distribution of a Markov process is called the linear level-reduction algorithm (Gaver *et al.*, 1984), an outline of which is given in Algorithm 4.1. This algorithm was implemented in code presented in Appendix A.1.1, which will be used in Section 4.5.

Algorithm 4.1: Linear level-reduction algorithm to calculate the stationary probability distribution of the approximating process $\mathcal{X}^{(j)}$.

$\mathbf{H}_N^{(j)} = \mathbf{A}_{N,N}^{(j)}$;
for $k = N-1, N-2, \ldots, \eta$ **do**
  $\mathbf{H}_k^{(j)} = \mathbf{A}_{k,k}^{(j)} - \mathbf{A}_{k,k+1}^{(j)}\left(\mathbf{H}_{k+1}^{(j)}\right)^{-1}\mathbf{A}_{k+1,k}^{(j)}$
end
$\boldsymbol{\pi}_\eta^{(j)*} = 1$;
for $k = \eta+1, \ldots, N$ **do**
  $\boldsymbol{\pi}_k^{(j)*} = -\boldsymbol{\pi}_{k-1}^{(j)*}\mathbf{A}_{k-1,k}^{(j)}\left(\mathbf{H}_k^{(j)}\right)^{-1}$
end
for $k = \eta, \ldots, N$ **do**
  $\boldsymbol{\pi}_k^{(j)} = \frac{\boldsymbol{\pi}_k^{(j)*}}{\sum_{i=\eta}^{N}\sum_{j=1}^{L_i}\boldsymbol{\pi}_{i,j}^{(j)*}}$
end

4.2.3 Linear noise approximation

Van Kampen's linear noise approximation (van Kampen, 2007) allows for the approximation of a discrete stochastic model with a continuous deterministic one in the form of a mean plus fluctuations. This, in turn, allows for the approximation of the QSD of the stochastic model by studying the steady state behaviour of the deterministic approximation (Elf & Ehrenberg, 2003). To begin the calculation of the approximation, the step operators $\mathbb{S}^i$ and $\mathbb{S}^{-i}$ are defined, such that on a

generic function f they act as follows

$$\mathbb{S}^i f(\mathbf{n}) = f(\mathbf{n}^{(+i)}), \tag{4.45a}$$

$$\mathbb{S}^{-i} f(\mathbf{n}) = f(\mathbf{n}^{(-i)}). \tag{4.45b}$$

Then, using these operators the Kolmogorov equation of the competition process, Eq. (4.21), can be re-written as

$$\frac{\mathrm{d}p_{\mathbf{n}}(t)}{\mathrm{d}t} = \sum_{i=1}^{\eta} (\mathbb{S}^{-i} - 1)\left[\lambda_{\mathbf{n}}^{(i)} p_{\mathbf{n}}(t)\right] + (\mathbb{S}^{i} - 1)\left[\mu_{\mathbf{n}}^{(i)} p_{\mathbf{n}}(t)\right]. \tag{4.46}$$

Now, the variables $x_i(t)$ and $\xi_i(t)$ are defined as follows

$$X_i(t) = \Omega x_i(t) + \Omega^{1/2}\xi_i(t), \tag{4.47}$$

where the parameter Ω represents the size of the system, the fluctuations $\xi_i(t)$ are of order $\Omega^{1/2}$, and $\Omega x_i(t) = \mathbb{E}\left[X_i(t)\right]$. From Eq. (4.47) it is easy to see that the probability density $p_{\mathbf{n}}(t)$ needs to be re-written to study the system in terms of the variables $x_i(t)$ and $\xi_i(t)$. Thus, the density $\Pi(\xi;t)$, with $\xi = (\xi_1, \ldots, \xi_\eta)$, is considered. This density satisfies (van Kampen, 2007)

$$\begin{aligned}\frac{\partial \Pi}{\partial t} - \Omega^{1/2}\sum_{i=1}^{\eta}\frac{\mathrm{d}x_i}{\mathrm{d}t}\frac{\partial \Pi}{\partial \xi_i} = &\sum_{i=1}^{\eta}\left(-\Omega^{-1/2}\frac{\partial}{\partial \xi_i} + \frac{1}{2}\Omega^{-1}\frac{\partial^2}{\partial \xi_i^2}\right)\lambda_{\mathbf{n}}^{(i)}\Pi \\ &+ \left(\Omega^{-1/2}\frac{\partial}{\partial \xi_i} + \frac{1}{2}\Omega^{-1}\frac{\partial^2}{\partial \xi_i^2}\right)\mu_{\mathbf{n}}^{(i)}\Pi.\end{aligned} \tag{4.48}$$

Define $x_{ij}^k = x_i + \sum_{\ell \in I_j^k} x_\ell$ and $\xi_{ij}^k = \xi_i + \sum_{\ell \in I_j^k} \xi_\ell$, and substitute Eq. (4.47) in Eq. (4.19) and Eq. (4.3) to obtain the following expressions for the birth and death rates of the approximating process

$$\begin{aligned}\lambda_{\mathbf{n}}^{(i)} &= \Omega\tilde{\varphi}_i\left(x_i + \Omega^{-1/2}\xi_i\right)\sum_{j=0}^{\eta-1}\sum_{k=1}^{\binom{\eta-1}{j}}\frac{p_{ij}^k}{x_{ij}^k + \Omega^{-1/2}\xi_{ij}^k} \\ &= \Omega\tilde{\varphi}_i\left(x_i + \Omega^{-1/2}\xi_i\right)\sum_{j=0}^{\eta-1}\sum_{k=1}^{\binom{\eta-1}{j}}\frac{p_{ij}^k}{x_{ij}^k}\frac{1}{1 + \Omega^{-1/2}\frac{\xi_{ij}^k}{x_{ij}^k}},\end{aligned} \tag{4.49a}$$

$$\mu_{\mathbf{n}}^{(i)} = \mu_i\left(\Omega x_i + \Omega^{1/2}\xi_i\right), \tag{4.49b}$$

where $\varphi_i = \Omega\tilde{\varphi}_i$. Then, the Kolmogorov equation Eq. (4.48) becomes

$$\begin{aligned}\frac{\partial \Pi}{\partial t} - \Omega^{1/2}\sum_{i=1}^{\eta}\frac{\mathrm{d}x_i}{\mathrm{d}t}\frac{\partial \Pi}{\partial \xi_i} = &\sum_{i=1}^{\eta}\left(-\Omega^{1/2}\frac{\partial}{\partial \xi_i} + \frac{1}{2}\Omega^0\frac{\partial^2}{\partial \xi_i^2}\right)\tilde{\varphi}_i\left(x_i + \Omega^{-1/2}\xi_i\right) \\ &\times\left(\sum_{j=0}^{\eta-1}\sum_{k=1}^{\binom{\eta-1}{j}}\frac{p_{ij}^k}{x_{ij}^k} - \Omega^{-1/2}\sum_{j=0}^{\eta-1}\sum_{k=1}^{\binom{\eta-1}{j}}\frac{p_{ij}^k\xi_{ij}^k}{\left(x_{ij}^k\right)^2}\right)\Pi \\ &+\left(\Omega^{1/2}\frac{\partial}{\partial \xi_i} + \frac{1}{2}\Omega^0\frac{\partial^2}{\partial \xi_i^2}\right)\mu_i\left(x_i + \Omega^{-1/2}\xi_i\right)\Pi.\end{aligned} \tag{4.50}$$

Now, collecting the terms of order $\Omega^{1/2}$ from Eq. (4.50) results in the following deterministic approximation of the competition process (Elf & Ehrenberg, 2003)

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \tilde{\varphi}_i x_i\sum_{j=0}^{\eta-1}\sum_{k=1}^{\binom{\eta-1}{j}}\frac{p_{ij}^k}{x_{ij}^k} - \mu_i x_i, \tag{4.51}$$

and collecting the terms of order Ω^0 yields a Fokker-Planck equation that describes the fluctuations around the steady state of Eq. (4.51)

$$\frac{\partial \Pi}{\partial t} = -\sum_{i,j}A_{ij}\frac{\partial \xi_j \Pi}{\partial \xi_i} + \frac{1}{2}\sum_{i,j}B_{ij}\frac{\partial^2 \Pi}{\partial \xi_i \partial \xi_j}, \tag{4.52}$$

where

$$\begin{aligned}-\sum_{i,j}A_{ij}\frac{\partial \xi_j \Pi}{\partial \xi_i} = &\sum_{i=1}^{\eta}\left(\mu_i - \tilde{\varphi}_i\sum_{j=0}^{\eta-1}\sum_{k=1}^{\binom{\eta-1}{j}}\frac{p_{ij}^k}{x_{ij}^k}\right)\frac{\partial \xi_i \Pi}{\partial \xi_i} \\ &+\tilde{\varphi}_i x_i\sum_{j=0}^{\eta-1}\sum_{k=1}^{\binom{\eta-1}{j}}\frac{p_{ij}^k}{\left(x_{ij}^k\right)^2}\frac{\partial \xi_{ij}^k \Pi}{\partial \xi_i},\end{aligned} \tag{4.53a}$$

$$\frac{1}{2}\sum_{i,j}B_{ij}\frac{\partial^2 \Pi}{\partial \xi_i \partial \xi_j} = \frac{1}{2}\sum_{i=1}^{\eta}\left[\mu_i x_i + \tilde{\varphi}_i x_i\sum_{j=0}^{\eta-1}\sum_{k=1}^{\binom{\eta-1}{j}}\frac{p_{ij}^k}{x_{ij}^k}\right]\frac{\partial^2 \Pi}{\partial \xi_i^2}. \tag{4.53b}$$

It is easy to see from Eq. (4.53a) that

$$\begin{aligned} A_{ii} &= \tilde{\varphi}_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{ij}^k}{x_{ij}^k} - \tilde{\varphi}_i x_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{ij}^k}{\left(x_{ij}^k\right)^2} - \mu_i \\ &= \tilde{\varphi}_i \left(\sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{ij}^k}{x_{ij}^k} - x_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{ij}^k}{\left(x_{ij}^k\right)^2} \right) - \mu_i \\ &= \tilde{\varphi}_i \left(\sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{ij}^k (x_{ij}^k - x_i)}{(x_{ij}^k)^2} \right) - \mu_i, \end{aligned} \tag{4.54a}$$

$$A_{ij} = -\tilde{\varphi}_i x_i \sum_{\ell=0}^{\eta-2} \sum_{k=1}^{\binom{\eta-2}{\ell}} \frac{p_{\{i,j\}\ell}^k}{\left(x_{\{i,j\}\ell}^k\right)^2}. \tag{4.54b}$$

Similarly, from Eq. (4.53b) it is clear that

$$B_{ii} = \mu_i x_i + \tilde{\varphi}_i x_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{ij}^k}{x_{ij}^k}, \tag{4.55a}$$

$$B_{ij} = 0. \tag{4.55b}$$

When evaluated at the steady state of Eq. (4.51), Eq. (4.52) becomes a linear multi-variate Fokker-Planck equation whose solution is a multi-variate Gaussian distribution, and is therefore fully determined by its first and second moments. Multiplying Eq. (4.52) by ξ_i and integrating results in the following equations describing the first moments of the distribution (van Kampen, 2007)

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\langle \xi_i \rangle &= \sum_{m=1}^{\eta} A_{im} \langle \xi_m \rangle \\ &= \left(\tilde{\varphi}_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{ij}^k (x_{ij}^k - x_i)}{(x_{ij}^k)^2} - \mu_i \right) \langle \xi_i \rangle \\ &\quad - \tilde{\varphi}_i x_i \sum_{\ell \neq i} \sum_{j=0}^{\eta-2} \sum_{k=1}^{\binom{\eta-2}{j}} \frac{p_{\{i,\ell\}j}^k}{\left(x_{\{i,\ell\}j}^k\right)^2} \langle \xi_\ell \rangle, \end{aligned} \tag{4.56}$$

where $\langle \cdot \rangle$ denotes $\mathbb{E}[\cdot]$. Similarly, multiplying by $\xi_i \xi_\ell$ one obtains the following

differential equations for the second moments of the distribution (van Kampen, 2007), if $i = \ell$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\langle \xi_i^2 \rangle &= 2\sum_{m=1}^{\eta} A_{im}\langle \xi_i \xi_m \rangle + B_{ii} \\
&= 2\left(\tilde{\varphi}_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{ij}^k (x_{ij}^k - x_i)}{(x_{ij}^k)^2} - \mu_i \right) \langle \xi_i^2 \rangle \\
&\quad - 2\tilde{\varphi}_i x_i \sum_{\ell \neq i} \sum_{j=0}^{\eta-2} \sum_{k=1}^{\binom{\eta-2}{j}} \frac{p_{\{i,\ell\}j}^k}{\left(x_{\{i,\ell\}j}^k\right)^2} \langle \xi_i \xi_\ell \rangle \\
&\quad + \mu_i x_i + \tilde{\varphi}_i x_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{ij}^k}{x_{ij}^k},
\end{aligned} \tag{4.57}$$

and if $i \neq \ell$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\langle \xi_i \xi_\ell \rangle &= \sum_{m=1}^{\eta} A_{im}\langle \xi_m \xi_\ell \rangle + \sum_{m=1}^{\eta} A_{\ell m}\langle \xi_i \xi_m \rangle + B_{i\ell} \\
&= \left(\tilde{\varphi}_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{ij}^k (x_{ij}^k - x_i)}{(x_{ij}^k)^2} - \mu_i \right) \langle \xi_i^2 \rangle \\
&\quad + \left(\tilde{\varphi}_\ell \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{\ell j}^k (x_{\ell j}^k - x_\ell)}{(x_{\ell j}^k)^2} - \mu_\ell \right) \langle \xi_\ell^2 \rangle \\
&\quad - \tilde{\varphi}_i x_i \sum_{m \neq i} \sum_{j=0}^{\eta-2} \sum_{k=1}^{\binom{\eta-2}{j}} \frac{p_{\{i,m\}j}^k}{\left(x_{\{i,m\}j}^k\right)^2} \langle \xi_m \xi_\ell \rangle \\
&\quad - \tilde{\varphi}_\ell x_\ell \sum_{m \neq i} \sum_{j=0}^{\eta-2} \sum_{k=1}^{\binom{\eta-2}{j}} \frac{p_{\{\ell,m\}j}^k}{\left(x_{\{\ell,m\}j}^k\right)^2} \langle \xi_i \xi_m \rangle.
\end{aligned} \tag{4.58}$$

In order to study the steady states of this approximation, and their stability, the code presented in Appendix A.1.2 will be used in Section 4.5. Other features of the linear noise approximation described by Equations (4.56), (4.57), and (4.58) are not considered in this Thesis, however, they are presented here in order to facilitate future research of this model.

4.3 Study of clonal extinction

In this section, the behaviour of the competition process $\mathcal{X}$ and its extinction events is studied with the use of stochastic descriptors and first step arguments. Some of these descriptors are calculated for the competition process with an infinite state space (Sections 4.3.1 and 4.3.2). In Sections 4.3.3 and 4.3.4 a finite (*i.e.*, truncated) state space is considered to make numerical calculations possible.

4.3.1 Total extinction in finite time

Using the method described in Iglehart (1964), it can be shown that the competition process reaches the absorbing state, $(0,\ldots,0)$, with certainty. To prove this, first the complete state space $\mathcal{S}$ is partitioned into levels defined by

$$L(k) = \left\{ (n_1,\ldots,n_\eta) \in \mathcal{S} : \sum_{i=1}^{\eta} n_i = k \right\},$$

for $k = 0,1,2,\ldots$, so that $\mathcal{S} = \bigcup_{k=0}^{+\infty} L(k)$. Now, define

$$\lambda'_k = \max_{\mathbf{n}\in L(k)} \left\{ \sum_{i=1}^{\eta} \lambda_{\mathbf{n}}^{(i)} \right\}, \quad \mu'_k = \min_{\mathbf{n}\in L(k)} \left\{ \sum_{i=1}^{\eta} \mu_{\mathbf{n}}^{(i)} \right\}, \tag{4.59}$$

for $k \geq 1$, and since $L(0) = \{(0,\ldots,0)\}$, then $\lambda'_0 = \mu'_0 = 0$. Equipped with these rates, a uni-variate birth and death process can be defined on the state space $\mathcal{S}' = \{L(k) : k = 0,1,2,\ldots\}$, which considers each level $L(k)$ as a single state, with birth rates λ'_k and death rates μ'_k (see Figure 4.3). From the definition of these rates, this uni-variate birth and death process moves towards $L(0)$ at a slower rate than the original process (Iglehart, 1964).

Reference (Iglehart, 1964, Theorem 3) states that a sufficient condition for absorption at $L(0)$ to be certain is that the sum

$$\sum_{k=1}^{+\infty} \frac{1}{\lambda'_k \sigma_k}, \tag{4.60}$$

where $\sigma_k = \frac{\lambda'_1\lambda'_2\ldots\lambda'_{k-1}}{\mu'_2\mu'_3\ldots\mu'_k}$ for $k \geq 2$ and $\sigma_1 = 1$, is divergent. First, note that $\lambda_{\mathbf{n}}^{(i)}$ is

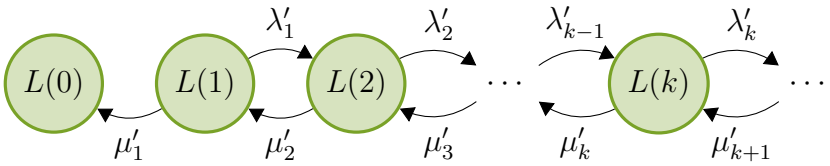


Figure 4.3: Uni-variate bounding competition process of the multi-variate competition process defined on the state space $\mathcal{S}' = \{L(k) : k = 0, 1, 2, \ldots\}$ with the rates defined in Eq. (4.59).

bounded by

$$\begin{aligned}
\lambda_{\mathbf{n}}^{(i)} &= \varphi_i n_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} p_{ij}^k e^{-\nu_{ij}^k} \sum_{r=0}^{M} \frac{\left(\nu_{ij}^k\right)^r}{r!} \frac{1}{n_i + \sum\limits_{l \in I_{ij}^k} n_l + r\langle n \rangle} \\
&\leq \varphi_i n_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} p_{ij}^k e^{-\nu_{ij}^k} \sum_{r=0}^{+\infty} \frac{\left(\nu_{ij}^k\right)^r}{r!} \frac{1}{n_i} \\
&= \varphi_i n_i \sum_{j=0}^{\eta-1} \sum_{k=1}^{\binom{\eta-1}{j}} p_{ij}^k \frac{1}{n_i} = \varphi_i,
\end{aligned} \tag{4.61}$$

so that

$$\lambda'_k = \max_{\mathbf{n} \in L(k)} \left\{ \sum_{i=1}^{\eta} \lambda_{\mathbf{n}}^{(i)} \right\} \leq \sum_{i=1}^{\eta} \varphi_i, \tag{4.62}$$

and

$$\mu'_k = \min_{\mathbf{n} \in L(k)} \left\{ \sum_{i=1}^{\eta} \mu_{\mathbf{n}}^{(i)} \right\} = \min_{\mathbf{n} \in L(k)} \left\{ \sum_{i=1}^{\eta} \mu_i n_i \right\} = k\mu^*, \tag{4.63}$$

where $\mu^* = \min_{i=1,2,\ldots,\eta} \{\mu_i\}$. Then, from Eq. (4.62) and Eq. (4.63) the following inequality is obtained

$$\sum_{k=1}^{+\infty} \frac{1}{\lambda'_k \sigma_k} \geq \sum_{k=1}^{+\infty} \frac{k! \left(\mu^*\right)^{k-1}}{\left(\sum\limits_{i=1}^{\eta} \varphi_i\right)^k} = \sum_{k=1}^{+\infty} a_k, \tag{4.64}$$

and the ratio

$$\frac{a_{k+1}}{a_k} = \frac{(k+1)\mu^*}{\sum\limits_{i=1}^{\eta} \varphi_i} \to +\infty$$

as k increases. Thus, by the ratio test this series diverges, which by Eq. (4.64) implies Eq. (4.60) also diverges. Therefore, by (Iglehart, 1964, Theorem 3) extinction of all clonotypes in competition process $\mathcal{X}$ is certain at sufficiently late times.

Having shown that extinction is certain for the process, the time to extinction is the next step in the analysis. Consider the state $\mathbf{n} = (n_1, \ldots, n_\eta) \in \mathcal{S}$ and let $\tau_{\mathbf{n}}$ be the mean time to extinction of all clonotypes, if the initial state is $\mathbf{n}$. Equivalently, $\tau_{\mathbf{n}}$ is the mean time to reach the absorbing state from state $\mathbf{n}$. Using the uni-variate process shown in Figure 4.3 and (Iglehart, 1964, Theorem 4), it can be shown that $\tau_{\mathbf{n}}$ is finite for all $\mathbf{n}$ that have at least one non-zero entry if

$$\sum_{k=1}^{+\infty} \sigma_k < +\infty,$$

where the σ_i are the same as before.

Using Eq. (4.62) and Eq. (4.63) the following inequality is obtained

$$\sum_{k=1}^{+\infty} \sigma_k \leq \sum_{k=1}^{+\infty} \frac{\left(\sum_{i=1}^{\eta} \varphi_i\right)^k}{k!\,(\mu^*)^{k-1}} = \sum_{k=1}^{+\infty} b_k,$$

then the ratio

$$\frac{b_{k+1}}{b_k} = \frac{\sum_{i=1}^{\eta} \varphi_i}{(k+1)\mu^*} \to 0,$$

as k increases, and the series converges by the ratio test. Thus, by (Iglehart, 1964, Theorem 4) the mean time to extinction of all clonotypes from any non-zero initial state is finite. Using the expression for the mean time to extinction of a uni-variate process (Howard M & Karlin, 1998), an upper bound on the mean time to extinction for states in each level, τ'_m, can be found; that is, for all $\mathbf{n} \in L(m)$ the following relation holds $\tau_{\mathbf{n}} \leq \tau'_m$, and the value of τ'_m is given by

$$\tau'_m = \sum_{i=1}^{+\infty} \frac{1}{\lambda'_i \rho_i} + \sum_{j=1}^{m-1} \rho_j \sum_{k=j+1}^{+\infty} \frac{1}{\lambda'_k \rho_k},$$

with $\rho_1 = 1$, and $\rho_k = \prod_{i=0}^{k} \frac{\mu'_i}{\lambda'_i}$ for $k \geq 2$. This expression provides an upper bound for the mean time until total extinction that depends on the total number of cells of all clonotypes.

4.3.2 Mean time to first extinction event

In this section, the mean time to the first extinction event is calculated using a first step argument. Let $\mathbf{X}(t)$ and W_j be random variables describing the state of the system at time t and the waiting time from the $j-1$-th event to the j-th event, respectively. Denote by $T_{\mathbf{n}}$ the random variable that describes the time to the first extinction event, starting from state $\mathbf{n} \in \mathcal{A}^0$; that is

$$T_{\mathbf{n}} = \inf \{t \geq 0 : \ X_i(t) = 0 \text{ for some } i \mid \mathbf{X}(0) = \mathbf{n}\}.$$

Then, by the law of total probability the mean time to extinction starting in state $\mathbf{n}$, $\mathbb{E}[T_{\mathbf{n}}]$, can be written as

$$\hat{\tau}_{\mathbf{n}} \coloneqq \mathbb{E}[T_{\mathbf{n}}] = \mathbb{E}[W_1] + \sum_{\mathbf{m} \in \mathcal{S}} \mathbb{E}[T_{\mathbf{m}} | \mathbf{X}(0) = \mathbf{n}, \mathbf{X}(W_1) = \mathbf{m}] \, p_{\mathbf{nm}}, \tag{4.65}$$

where $p_{\mathbf{nm}}$ is the transition probability from $\mathbf{n}$ to $\mathbf{m}$ in the embedded Markov process (Allen, 2010); that is, $p_{\mathbf{nm}}$ can be written as

$$p_{\mathbf{nm}} = \begin{cases} \frac{\lambda_{\mathbf{n}}^{(i)}}{\Delta_{\mathbf{n}}}, & \text{if } \mathbf{m} = \mathbf{n}^{(+i)}, \\ \frac{\mu_{\mathbf{n}}^{(i)}}{\Delta_{\mathbf{n}}}, & \text{if } \mathbf{m} = \mathbf{n}^{(-i)}, \\ 0, & \text{otherwise}, \end{cases} \tag{4.66}$$

where $\Delta_{\mathbf{n}} = \sum_{i=1}^{\eta} \left(\lambda_{\mathbf{n}}^{(i)} + \mu_{\mathbf{n}}^{(i)} \right)$.

Now, from Eq. (4.65) the following set of difference equations can be obtained

$$\hat{\tau}_{\mathbf{n}} = \frac{1}{\Delta_{\mathbf{n}}} + \sum_{i=1}^{\eta} \frac{\lambda_{\mathbf{n}}^{(i)}}{\Delta_{\mathbf{n}}} \hat{\tau}_{\mathbf{n}^{(+i)}} + \sum_{i=1}^{\eta} \frac{\mu_{\mathbf{n}}^{(i)}}{\Delta_{\mathbf{n}}} \hat{\tau}_{\mathbf{n}^{(-i)}}. \tag{4.67}$$

Thus, the mean time to extinction from state $\mathbf{n}$ is defined in terms of the mean time to extinction from its adjacent states. This set of equations also has the boundary condition that $\hat{\tau}_{\mathbf{n}} = 0$ if $n_i = 0$ for any $1 \leq i \leq \eta$. Eq. (4.67) can be re-written as

$$-1 = -\Delta_{\mathbf{n}} \hat{\tau}_{\mathbf{n}} + \sum_{i=1}^{\eta} \lambda_{\mathbf{n}}^{(i)} \hat{\tau}_{\mathbf{n}^{(+i)}} + \sum_{i=1}^{\eta} \mu_{\mathbf{n}}^{(i)} \hat{\tau}_{\mathbf{n}^{(-i)}}, \tag{4.68}$$

from which it is easy to see that this system of equations can be represented as a matrix equation of the form $\mathbf{M}\boldsymbol{\tau} = \mathbf{1}$, which is solved numerically for $\boldsymbol{\tau}$ in Section 4.5, using the code presented in Appendix A.2 which exploiting the fact that the matrix of coefficients $\mathbf{M}$ is sparse.

4.3.3 Clonal size distribution at the first extinction event

The focus of this section is computing the probability of each clonotype to be the first one becoming extinct. Following a similar approach to that in Gómez-Corral & López-García (2012a,b), the infinite state space $\mathcal{S}$ is truncated. In particular, only states which have at most N_ε cells in total are considered; that is, a reflecting plane $\sum_{i=1}^{\eta} n_i = N_\varepsilon$ is introduced, and N_ε is defined as the minimum value such that

$$\sum_{k=\eta}^{N_\varepsilon} \mathbb{P}\left(\mathbf{X}(t) \in L(k) : \ \forall t \geq 0\right) \geq 1 - \varepsilon, \tag{4.69}$$

is satisfied; that is, N_ε is the level of the QSD for which $\bigcup_{k=\eta}^{N_\varepsilon} L(k)$ captures at least $1-\varepsilon$ of the mass of the probability distribution. To compute N_ε, a similar approach to that presented in Gómez-Corral & López-García (2012a,b) is used. Since three different approximations of the QSD have been introduced (see Sections 4.2.2 and 4.2.3), N_ε will be chosen to be the maximum between the values that satisfy Eq. (4.69) for all of the approximations.

Consider the subsets $\mathcal{A}^1 = \bigcup_{i=1}^{\eta} \mathcal{A}_i$ and $\mathcal{A}^0$, as defined in Eq. (4.26) and Eq. (4.32) respectively, and define the indicator function $\mathbf{1}_{\mathcal{A}^0}(\mathbf{n})$ as follows

$$\mathbf{1}_{\mathcal{A}^0}(\mathbf{n}) = \begin{cases} 1, & \text{if } \mathbf{n} \in \mathcal{A}^0, \\ 0, & \text{if } \mathbf{n} \in \mathcal{A}^1. \end{cases}$$

Then, consider a modified competition process, $\breve{\mathcal{X}} = \left\{\left(\breve{X}_1(t), \ldots, \breve{X}_\eta(t)\right) : t \geq 0\right\}$, on the state space $\mathcal{A}^0 \cup \mathcal{A}^1$, with birth and death rates given by

$$\begin{aligned} \breve{\lambda}_{\mathbf{n}}^{(i)} &= \mathbf{1}_{\mathcal{A}^0}(\mathbf{n}) \cdot \lambda_{\mathbf{n}}^{(i)}, \\ \breve{\mu}_{\mathbf{n}}^{(i)} &= \mathbf{1}_{\mathcal{A}^0}(\mathbf{n}) \cdot \mu_{\mathbf{n}}^{(i)}; \end{aligned}$$

that is, this process behaves exactly as the original competition process $\mathcal{X}$ until a clonotype becomes extinct, at which point the process $\breve{\mathcal{X}}$ comes to an end. Let the random vector $\breve{\mathbf{X}}(t) = \left(\breve{X}_1(t), \ldots, \breve{X}_\eta(t)\right)$ describe the population of all clonotypes being modelled at time t. First, a position function for states in $\mathcal{A}_i$ needs to be defined. To this end, the absorbing states are initially ordered as follows

$$\mathcal{A}_1 \prec \mathcal{A}_2 \prec \cdots \prec \mathcal{A}_{\eta-1} \prec \mathcal{A}_\eta. \tag{4.70}$$

Note that these sets can be bijectively projected onto a space of dimension $\eta - 1$ by removing the i-th element (which is the population of the clonotype that has become extinct). The projected state will be denoted by $\hat{\mathbf{n}}$ and the projected set by $\hat{\mathcal{A}}_i$. Given that $\hat{\mathcal{A}}_i$ has the same structure as $\mathcal{A}^0$, it can be separated into ordered levels as follows

$$L^1(\eta - 1) \prec L^1(\eta) \prec L^1(\eta + 1) \prec \cdots \prec L^1(N_\varepsilon - 1).$$

The last level that can be considered is $N_\varepsilon - 1$, since for a state to reach $L^1(N_\varepsilon)$ there must have been an extinction event from a state in $L^0(N_\varepsilon + 1)$, which is not part of the state space for the process with the reflecting boundary. The number of states in level k, L^1_k, is given by

$$L^1_k = |L^1(k)| = \binom{k-1}{\eta - 2},$$

and the total number of states in $\hat{\mathcal{A}}_i$ under the reflecting boundary is

$$|\hat{\mathcal{A}}_i| = \sum_{\ell=\eta-1}^{N_\varepsilon - 1} L^1_\ell = \sum_{\ell=\eta-1}^{N_\varepsilon - 1} \binom{\ell - 1}{\eta - 2} = \binom{N_\varepsilon - 1}{\eta - 1}. \tag{4.71}$$

Then, the position in $\hat{\mathcal{A}}_i$ of a state $\hat{\mathbf{n}} \in L^1(k)$ is given by

$$\mathrm{pos}_{\hat{\mathcal{A}}_i}(\hat{\mathbf{n}}, \eta - 1) = \mathrm{pos}_k(\hat{\mathbf{n}}, \eta - 1) + \sum_{i=\eta-1}^{k-1} L^1_i,$$

and by reversing the projection of $\mathcal{A}_i$ onto $\hat{\mathcal{A}}_i$ the following expression is obtained

$$\mathrm{pos}_{\mathcal{A}_i}(\mathbf{n}, \eta) = \mathrm{pos}_{\hat{\mathcal{A}}_i}(\hat{\mathbf{n}}, \eta - 1). \tag{4.72}$$

Then, from Eq. (4.70), Eq. (4.71), and Eq. (4.72) the position of state $\mathbf{n} \in L^1(k) \subset \mathcal{A}_i$ in $\mathcal{A}^1$ can be written as

$$\mathrm{pos}_{\mathcal{A}^1}(\mathbf{n}, \eta) = \mathrm{pos}_{\mathcal{A}_i}(\mathbf{n}, \eta) + (i - 1)|\hat{\mathcal{A}}_i|.$$

Now, the general position function can be defined as follows

$$\mathrm{pos}_{\mathcal{A}^0 \cup \mathcal{A}^1}(\mathbf{n}, \eta) = \begin{cases} \mathrm{pos}_{\mathcal{A}^0}(\mathbf{n}, \eta) & \text{if } \mathbf{n} \in \mathcal{A}^0, \\ \mathrm{pos}_{\mathcal{A}^1}(\mathbf{n}, \eta) + \binom{N}{\eta} & \text{if } \mathbf{n} \in \mathcal{A}^1. \end{cases} \tag{4.73}$$

Let $\mathcal{Y} = \{\mathbf{Y}(s) : \ s = 0, 1, 2, \ldots\}$ denote the embedded Markov process of $\breve{\mathcal{X}}$. Then, using the order of states as defined in Eq. (4.73), the transition matrix of $\mathcal{Y}$ is given by

$$\breve{\mathbf{P}} = \begin{bmatrix} \mathbf{P} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{bmatrix},$$

where $\mathbf{P}$ is a $|\mathcal{A}^0| \times |\mathcal{A}^0|$ matrix whose entries are the transition probabilities of states within $\mathcal{A}^0$, $\mathbf{R}$ is a $|\mathcal{A}^0| \times |\mathcal{A}^1|$ matrix whose entries are the transition probabilities from a state in $\mathcal{A}^0$ to a state in $\mathcal{A}^1$, $\mathbf{0}$ is the $|\mathcal{A}^1| \times |\mathcal{A}^0|$ zero matrix, and $\mathbf{I}$ is the identity matrix of size $|\mathcal{A}^1|$.

Let $\mathbf{U}^{(s)}$ be the $|\mathcal{A}^0| \times |\mathcal{A}^1|$ matrix of probabilities to reach every absorbing state in $\mathcal{A}^0$ from all states in $\mathcal{A}^1$ after at most s steps, where a step is defined as a birth or death event for any clonotype. That is, for a pair of states $\mathbf{n} \in \mathcal{A}^0$ and $\mathbf{m} \in \mathcal{A}^1$ the entry in row $\text{pos}_{\mathcal{A}^0}(\mathbf{n})$ and column $\text{pos}_{\mathcal{A}^1}(\mathbf{m})$ of $\mathbf{U}^{(s)}$ is given by

$$\mathbf{U}^{(s)}_{\text{pos}_{\mathcal{A}^0}(\mathbf{n}),\text{pos}_{\mathcal{A}^1}(\mathbf{m})} = \mathbb{P}\left\{\mathbf{Y}(s) = \mathbf{m} \mid \mathbf{Y}(0) = \mathbf{n}\right\}.$$

Then, $\mathbf{U}^{(s)}$ can be written as follows

$$\mathbf{U}^{(s)} = \left(\sum_{k=0}^{s-1} \mathbf{P}^k\right) \mathbf{R}.$$

Making use of the fundamental matrix $\mathbf{W} = (\mathbf{I} - \mathbf{P})^{-1}$ associated to $\mathbf{P}$ (Howard M & Karlin, 1998), it is easy to see that

$$\mathbf{U} = \lim_{s \to +\infty} \mathbf{U}^{(s)} = \mathbf{WR},$$

and thus, the matrix of probabilities to reach each absorbing state in $\mathcal{A}^1$ given any initial state in $\mathcal{A}^0$, $\mathbf{U}$, can be found by calculating

$$\mathbf{U} = (\mathbf{I} - \mathbf{P})^{-1}\mathbf{R}. \tag{4.74}$$

Note that the matrix $\mathbf{U}$ has a block structure and can be written as $\mathbf{U} = \begin{bmatrix}\mathbf{U}^1 & \mathbf{U}^2 & \cdots & \mathbf{U}^\eta\end{bmatrix}$, where $\mathbf{U}^i$ is the matrix of absorption probabilities from $\mathcal{A}^0$ to $\mathcal{A}_i$. That is, the entry in row ℓ and column j of $\mathbf{U}^i$, $(\mathbf{U}^i)_{\ell,j}$, is given by

$$\left(\mathbf{U}^i\right)_{\ell,j} = \lim_{s \to +\infty} \mathbb{P}\left\{\mathbf{Y}(s) = \mathbf{m} \mid \mathbf{Y}(0) = \mathbf{n}\right\},$$

where $\mathrm{pos}_{\mathcal{A}^0}(\mathbf{n}) = \ell$ and $\mathrm{pos}_{\mathcal{A}^1}(\mathbf{m}) = j$, for $1 \leq \ell \leq |\mathcal{A}^0|$ and $1 \leq j \leq |\mathcal{A}^1|$. Thus, each row of $\mathbf{U}$ adds up to 1 and represents the distribution of clonal sizes at the time of the first extinction event, with the initial state whose position is given by the row number. Furthermore, given the block structure of $\mathbf{U}$, each row of $\mathbf{U}^i$ contains the probabilities of clonal sizes for the surviving clonotypes when clonotype i is the first to become extinct.

Due to the computational cost, however, the matrix $\mathbf{I} - \mathbf{P}$ will not be directly inverted, but instead a linear level-reduction algorithm will be used on the matrix equation

$$(\mathbf{I} - \mathbf{P})\mathbf{U} = \mathbf{R}.$$

The matrix $\mathbf{I} - \mathbf{P}$ is of quasi-birth-and-death type and can be written as

$$\mathbf{I} - \mathbf{P} = \begin{bmatrix} \mathbf{I}_\eta & -\mathbf{B}_{\eta,\eta+1} & \cdots & \mathbf{0} \\ -\mathbf{B}_{\eta+1,\eta} & \mathbf{I}_{\eta+1} & \cdots & \mathbf{0} \\ \mathbf{0} & -\mathbf{B}_{\eta+2,\eta+1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & -\mathbf{B}_{N_\varepsilon-1,N_\varepsilon} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_{N_\varepsilon} \end{bmatrix}, \tag{4.75}$$

where $\mathbf{I}_k$ is the identity matrix of order $L_k^0 \times L_k^0$, $\mathbf{0}$ are zero matrices of the appropriate sizes, and the $\mathbf{B}_{k,k'}$ sub-matrices are defined as follows: Let $\mathbf{n}_k^i$ denote the state in $L^0(k)$ such that $\mathrm{pos}_k(\mathbf{n}_k^i, \eta) = i$, then

- for $\eta + 1 \leq k \leq N_\varepsilon$

$$(\mathbf{B}_{k,k-1})_{ij} = p_{\mathbf{n}_k^i \mathbf{n}_{k-1}^j}, \tag{4.76}$$

- for $\eta \leq k \leq N_\varepsilon - 1$

$$(\mathbf{B}_{k,k+1})_{ij} = p_{\mathbf{n}_k^i \mathbf{n}_{k+1}^i}, \tag{4.77}$$

where $p_{\mathbf{n}_k^i \mathbf{n}_{k-1}^j}$ and $p_{\mathbf{n}_k^i \mathbf{n}_{k+1}^i}$ are the transition probabilities given by Eq. (4.66).

The matrix $\mathbf{R}$ has the form $\mathbf{R} = \begin{bmatrix} \mathbf{R}^1 & \mathbf{R}^2 & \cdots & \mathbf{R}^\eta \end{bmatrix}$, where $\mathbf{R}^i$ is the matrix of transition probabilities from $\mathcal{A}^0$ to $\mathcal{A}_i$. The $\mathbf{R}^i$ matrices are diagonal by blocks and can be written as

$$\mathbf{R}^i = \begin{bmatrix} \mathbf{R}^i_{\eta,\eta-1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^i_{\eta+1,\eta} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}^i_{\eta+2,\eta+1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{R}^i_{N_\varepsilon,N_\varepsilon-1} \end{bmatrix}.$$

The $\mathbf{R}^i_{k,k-1}$ sub-matrices are defined as follows: First, let $\mathbf{n}^\ell_k$ to denote the state in $L^0(k)$ such that $\text{pos}_k(\mathbf{n}^\ell_k, \eta) = \ell$, and $\mathbf{m}^\ell_{i,k}$ denote the state in $L^1(k) \subset \mathcal{A}_i$, such that $\text{pos}_k(\hat{\mathbf{m}}^\ell_{i,k}, \eta - 1) = \ell$. Then, for $\eta \leq k \leq N_\varepsilon$, the entries of the matrix $\mathbf{R}^i_{k,k-1}$ are given by

$$(\mathbf{R}^i_{k,k-1})_{ab} = p_{\mathbf{n}^a_k \mathbf{m}^b_{i,k-1}}, \tag{4.78}$$

where $p_{\mathbf{n}^a_k \mathbf{m}^b_{i,k-1}}$ is the transition probability given by Eq. (4.66). The structure of $\mathbf{U}$ allows for the use of a linear level-reduction algorithm to solve Eq. (4.74) one $\mathbf{U}^i$ block at a time. An outline of the algorithm is given in Algorithm 4.2. In this way, and making use of the structure of $\mathbf{U}$, the probability of clonotype i being the first to become extinct starting in state $\mathbf{n}_0 \in \mathcal{A}^0$, denoted by $\mathcal{U}^i_{\mathbf{n}_0}$, can be calculated as follows

$$\mathcal{U}^i_{\mathbf{n}_0} = \sum_{j=1}^{|\mathcal{A}_i|} \left(\mathbf{U}^i\right)_{\text{pos}_{\mathcal{A}^0}(\mathbf{n}_0), j}, \tag{4.79}$$

where $(\mathbf{U}^i)_{\ell,j}$ denotes the entry of the matrix $\mathbf{U}^i$ in the ℓ-th row and j-th column.

Algorithm 4.2: Linear level-reduction algorithm to calculate the distribution of clonal sizes at the time of the first extinction event, $\mathbf{U}$.

$\mathbf{H}_{N_\varepsilon} = \mathbf{I}_{L^0_{N_\varepsilon}}$;
for $k = N_\varepsilon - 1, N_\varepsilon - 2, \ldots, \eta$ **do**
  $\mathbf{H}_k = \mathbf{I}_{L^0_k} - \mathbf{B}_{k,k+1}\mathbf{H}^{-1}_{k+1}\mathbf{B}_{k+1,k}$
end
for $i = 1, 2, \ldots, \eta$ **do**
  for $j = \eta - 1, \eta, \ldots, N_\varepsilon - 1$ **do**
    $\mathbf{K}^i_{N_\varepsilon,j} = \mathbf{R}^i_{N_\varepsilon,j}$;
    for $k = N_\varepsilon - 1, N_\varepsilon - 2, \ldots, \eta$ **do**
      $\mathbf{K}^i_{k,j} = \mathbf{B}_{k,k+1}\mathbf{H}^{-1}_{k+1}\mathbf{K}^i_{k+1,j} + \mathbf{R}^i_{k,j}$
    end
    $\mathbf{U}^i_{\eta,j} = \mathbf{H}^{-1}_\eta \mathbf{K}^i_{\eta,j}$;
    for $k = \eta + 1, \eta + 2, \ldots, N_\varepsilon$ **do**
      $\mathbf{U}^i_{k,j} = \mathbf{H}^{-1}_k(\mathbf{K}^i_{k,j} + \mathbf{B}_{k,k-1}\mathbf{U}^i_{k-1,j})$
    end
  end
end

4.3.4 Number of divisions before extinction of a clonotype

In this section another stochastic descriptor of the competition process is studied: the number of divisions of a naive T cell clonotype before its extinction. To calculate this random variable a truncated state space similar to that defined in Section 4.3.3 is considered. This descriptor serves as an indicator of how much proliferation takes place before the extinction of a T cell clonal family. Given a fiducial (fixed but arbitrary) initial state, $\mathbf{n} \in \mathcal{A}^0$, let $\mathcal{D}_{i,d}^{(s)}(\mathbf{n})$ denote the probability that clonotype i divided d times in at most s steps before becoming extinct if it had an initial size of $\mathbf{n}$; that is,

$$\mathcal{D}_{i,d}^{(s)}(\mathbf{n}) = \mathbb{P}\left(\begin{array}{c}\text{clone } i \text{ divided } d \text{ times in at most} \\ s \text{ steps before becoming extinct}\end{array} \middle| \, \mathbf{X}(0) = \mathbf{n}\right). \tag{4.80}$$

Since a cell of clonotype i can divide after the extinction of other clonotypes, the complete state space, $\mathcal{S}$, is considered with a reflecting boundary at $\sum_{i=1}^{\eta} n_i = N_\varepsilon$. To order the elements of $\mathcal{S}$, note that an element $\mathbf{n} \in \mathcal{S}$ can be bijectively mapped to $\mathcal{A}^0$ by adding $\mathbf{e} = (1, \ldots, 1)$ to $\mathbf{n}$. Then, the position of state $\mathbf{n}$ in the k-th level of $\mathcal{S}$, is given by

$$\mathrm{pos}_{\mathcal{S},k}(\mathbf{n}, \eta) = \mathrm{pos}_{k+\eta}(\mathbf{n} + \mathbf{e}, \eta).$$

Furthermore, this bijection also allows for the use of the levels defined in $\mathcal{A}^0$ (see Eq. (4.34)). Now, consider $\mathcal{D}_{i,d}^{(s)}$ as the vector built from entries $\mathcal{D}_{i,d}^{(s)}(\mathbf{n})$, for each $\mathbf{n} \in \mathcal{S}$. Making use of the level structure it can be divided into sub-vectors corresponding to the probabilities for states in each level of $\mathcal{S}$, *i.e.*, $\mathcal{D}_{i,d}^{(s)} = \left[\mathcal{D}_{i,d,0}^{(s)}, \mathcal{D}_{i,d,1}^{(s)}, \ldots, \mathcal{D}_{i,d,N_\varepsilon}^{(s)}\right]^\top$, where $\mathcal{D}_{i,d,k}^{(s)}$ is the vector of probabilities for $\mathbf{n} \in L(k)$.

Given that the subject of this stochastic descriptor is the probability of clone i dividing, the usual transition matrix is not used, but instead it is separated into the vector of probabilities for division in clonotype i, and the sub-stochastic matrix $\mathbf{P}^{(i)}$ of all transition probabilities except those for division in clonotype i;

that is,

$$\mathbf{P}^{(i)} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{C}^{(i)}_{1,0} & \mathbf{0} & \mathbf{C}^{(i)}_{1,2} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^{(i)}_{2,1} & \mathbf{0} & \mathbf{C}^{(i)}_{2,3} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}^{(i)}_{N_\varepsilon-1,N_\varepsilon} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}, \tag{4.81}$$

where $\mathbf{0}$ are zero matrices of the appropriate sizes. To define the $\mathbf{C}^{(i)}_{k,k'}$ sub-matrices, let $\mathbf{n}^\ell_k$ denote the state in $L(k)$ such that $\mathrm{pos}_{\mathcal{S},k}(\mathbf{n}^\ell_k, \eta) = \ell$, then the $\mathbf{C}_{k,k'}$ matrices are defined as

- for $1 \leq k \leq N_\varepsilon$

$$\left(\mathbf{C}^{(i)}_{k,k-1}\right)_{\ell j} = \begin{cases} p_{\mathbf{n}^\ell_k \mathbf{n}^j_{k-1}} & \text{if } \left(\mathbf{n}^\ell_k\right)_i > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.82}$$

- for $1 \leq k \leq N_\varepsilon - 1$

$$\left(\mathbf{C}^{(i)}_{k,k+1}\right)_{\ell j} = \begin{cases} p_{\mathbf{n}^\ell_k \mathbf{n}^j_{k+1}} & \text{if } \left(\mathbf{n}^\ell_k\right)_i > 0 \text{ and } \left(\mathbf{n}^\ell_k\right)_i = \left(\mathbf{n}^j_{k+1}\right)_i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.83}$$

Note that since the probability being calculated is the probability of division before extinction of clonotype i, all transition probabilities for a state in which clonotype i is already extinct are zero.

The distribution is calculated recursively on the number of divisions before extinction. Thus, first the vector $\mathbf{d}^{(i)}_0$ is defined as the vector containing the probability that in zero steps clonotype i will become extinct; that is,

$$\left(\mathbf{d}^{(i)}_0\right)_{\mathrm{pos}_{\mathcal{S},k}(\mathbf{n},\eta)} = \begin{cases} 1, & \text{if } n_i = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Then, the probability that clone i will not divide before becoming extinct in s or fewer steps is given by

$$\mathcal{D}^{(s)}_{i,0} = \left(\sum_{j=0}^{s} \left(\mathbf{P}^{(i)}\right)^j\right) \mathbf{d}^{(i)}_0.$$

Making use of the fundamental matrix associated to $\mathbf{P}^{(i)}$, $\mathbf{W}^{(i)} = \left(\mathbf{I} - \mathbf{P}^{(i)}\right)^{-1}$, the following expression for $\mathcal{D}_{i,0}$ is found

$$\mathcal{D}_{i,0} = \lim_{t \to +\infty} \mathcal{D}_{i,0}^{(s)} = \mathbf{W}^{(i)} \mathbf{d}_0^{(i)},$$

which leads to

$$\left(\mathbf{I} - \mathbf{P}^{(i)}\right) \mathcal{D}_{i,0} = \mathbf{d}_0^{(i)}. \tag{4.84}$$

Then, Eq. (4.84) allows for the probability of extinction of clonotype i after 0 divisions, $\mathcal{D}_{i,0}$, to be solved for:

$$\mathcal{D}_{i,0} = \left(\mathbf{I} - \mathbf{P}^{(i)}\right)^{-1} \mathbf{d}_0^{(i)}. \tag{4.85}$$

For $\ell > 0$ divisions, define the vector containing the probability of clonotype i becoming extinct after ℓ divisions given that it has divided once, as follows

$$\left(\mathbf{d}_\ell^{(i)}\right)_{\mathrm{pos}_{8,k}(\mathbf{n},\eta)} = p_{\mathbf{nn}^{(+i)}} \mathcal{D}_{i,\ell-1}\left(\mathbf{n}^{(+i)}\right). \tag{4.86}$$

Then, the vector of probabilities of clonotype i becoming extinct after ℓ divisions, $\mathcal{D}_{i,\ell}$, satisfies the recursive equation

$$\mathcal{D}_{i,\ell} = \left(\mathbf{I} - \mathbf{P}^{(i)}\right)^{-1} \mathbf{d}_\ell^{(i)}. \tag{4.87}$$

Finally, with the use of an adapted version of Algorithm 4.2, Eq. (4.85) and Eq. (4.87) can be solved recursively until the number of divisions reaches $\ell_{\mathbf{n},\delta}$; that is,

$$\sum_{d=0}^{\ell_{\mathbf{n},\delta}} \mathcal{D}_{i,d}(\mathbf{n}) \geq 1 - \delta,$$

with $\mathbf{n}$ the initial state under consideration. The values $\mathcal{D}_{i,d}(\mathbf{n})$ for $d = 1, 2, \ldots, \ell_{\mathbf{n},\delta}$ describe the probability distribution for the number of divisions of clonotype i before its extinction.

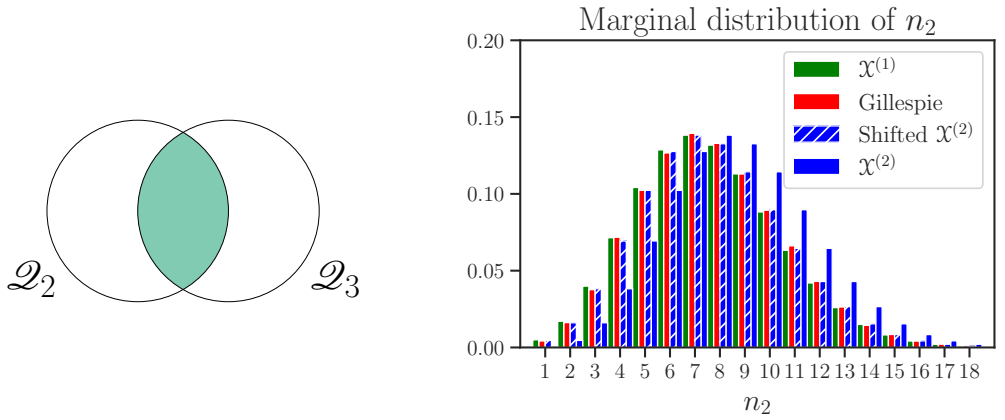
4.4 Numerical results in the case $\eta = 3$

In this section the multi-variate competition process (see Section 4.1), and the stochastic descriptors defined (see Sections 4.2 and 4.3) are used to study the

dynamics of two clonotypes homeostatically established in the periphery, which compete for self-pMHC stimuli with a third, and new, clonotype that has just exited the thymus. Thus, the two clonotypes homeostatically established will be in a state defined by the mean of their two-dimensional QSD (and in the absence of the newly arrived clonotype). In this case, there are only two p_{ij}^k probabilities for each clonotype. The probability $p_{i,0}$ that a self-pMHC is recognised only by clonotype i, and the probability $p_{i,1}$ that a self-pMHC is recognised by both clonotypes. Thus, by Eq. (4.17), only one probability value needs to be chosen to determine all others. The numerical results shown here were obtained using the code presented in Appendix A.

The new clonotype will be denoted as clonotype 1, and the established ones as clonotype 2 and clonotype 3, respectively. All three clonotypes are considered to have the same total homeostatic stimulus (Lythe *et al.*, 2016); that is, $\varphi_1 = \varphi_2 = \varphi_3 = 10$ divisions $\cdot$ year^{-1}, and before the introduction of clonotype 1 the self-pMHCs in $\mathscr{Q}_2 \cap \mathscr{Q}_3$ are hypothesised to be in the soft niche, and the remaining self-pMHCs in the hard niche. The per cell death rate of T cells has been estimated from experimental data (Borghans *et al.*, 2018), and is considered to be $\mu_1 = \mu_2 = \mu_3 = 1$ year^{-1}. Figure 4.4a shows the competition scenario described for the two established clonotypes, where the size of each $\mathscr{Q}_i$ circle represents the magnitude of φ_i, and the colour of each section represents the magnitude of ν_{ij}^k, with a darker colour indicating a greater value. In this competition scenario the values of $p_{2,1}$ and $\nu_{2,1}$ were chosen to be $1/3$ and 1, respectively.

Making use of the method described in Section 4.3.2, the mean time to extinction was calculated for all initial states with at most 10^2 total cells. For the choice of φ_i, and μ_i, the times were found to range between 80 and 125 years (see Table 4.2). Given these long extinction times, in a biological timescale, it is appropriate to approximate the QSD of this competition process. In Figure 4.4b the marginal distribution for clonotype 2 is plotted. Only the distribution for one of the established clonotypes is shown, since the competition between them is symmetric, and thus clonotypes 2 and 3 are identical before the introduction of clonotype 1. From this figure it can be observed that for the competition scenario considered, with most of the self-pMHCs in the hard niche, the approximation $\mathcal{X}^{(1)}$ better describes the behaviour of the system (see Gillespie simulations). However,



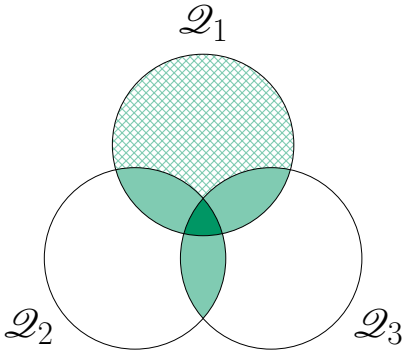
(a) **(b)**

Figure 4.4: Competition scenario for two established clonotypes and marginal distribution of the QSD. (a) Competition scenario considered for the two established clonotypes, where they compete for 1/3 of their homeostatic proliferation stimulus. The shaded region represents the subset of self-pMHCs considered to be in the soft niche, with a mean niche overlap value $\nu_{2,1} = \nu_{3,1} = 1$. (b) Marginal distribution of the QSD for clonotype 2, approximated using the processes $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$ defined in Section 4.2.2 and the code presented in Appendix A.1.1.

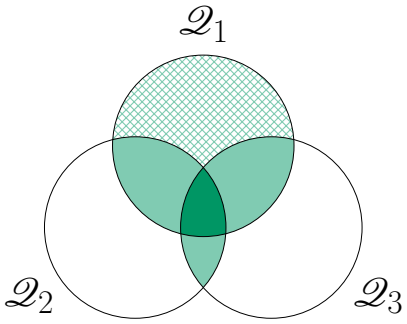
note that shifting $\mathcal{X}^{(2)}$ by one cell; that is, removing the immortal cell from each state, results again in a good approximation. This is to be expected for these small population sizes and taking into account the hypothesis of an immortal cell in this process.

Once clonotype 1 enters the periphery four different competition scenarios across a spectrum of symmetries are considered. These scenarios range from full symmetry to complete asymmetry. In the first scenario, Figure 4.5a, all clonotypes compete symmetrically; that is, all one-on-one competitions have the same probability. For the second scenario, Figure 4.5b, there is symmetric competition between clonotype 1 and clonotypes 2 and 3, but these competitions are greater than the competition between clonotypes 2 and 3, leaving the new clonotype at a disadvantage. The third scenario, Figure 4.5c, has clonotype 1 competing more for homeostatic stimuli with clonotype 3 and less with clonotype 2, giving clonotype 2 an advantage. Finally, the last scenario, Figure 4.5d, represents

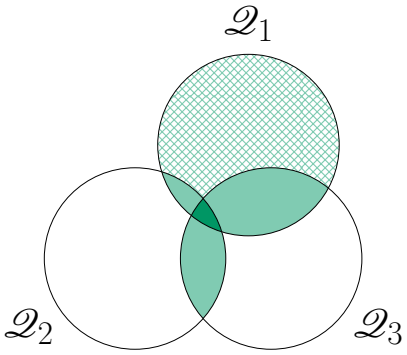
the case of extreme asymmetry in which clonotypes 1 and 3 compete completely for stimuli. In all scenarios, self-pMHCs recognised by more than one clonotype are considered to be in the soft niche, with the value of the mean niche overlap increasing as the number of clonotypes increases. For the self-pMHCs recognised by clonotype 1, both hard and soft niches are studied.



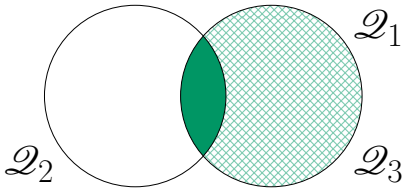
(a) Symmetric competition.



(b) Symmetric competition between clonotype 1 and clonotypes 2 and 3, but greater than the competition between clonotypes 2 and 3.



(c) Clonotype 1 has lower competition with clonotype 2, and greater competition with clonotype 3.



(d) Clonotype 1 competes completely with clonotype 3.

Figure 4.5: Competition scenarios when a new clonotype is introduced in a bi-variate system. Shaded areas represent sets of self-pMHCs that are in the soft niche, with a darker shade meaning a lager value of the mean niche overlap (ν_{ij}^k). Cross-hatched areas represent sections where both the soft and hard niche cases are considered.

The diagrams in Figure 4.5 were used as a guide to choose the probabilities shown in Table 4.1. Since the competition being considered is that between three different clonotypes ($\eta = 3$), the following set of relations for the p_{ij}^k probabilities are obtained from Eq. (4.17)

$$\begin{gathered} \varphi_1 p_{1,2}^1 = \varphi_2 p_{2,2}^1 = \varphi_3 p_{3,2}^1, \quad \varphi_1 p_{1,1}^1 = \varphi_2 p_{2,1}^1, \\ \varphi_1 p_{1,1}^2 = \varphi_3 p_{3,1}^1, \quad \varphi_2 p_{2,1}^2 = \varphi_3 p_{3,1}^2, \end{gathered} \tag{4.88}$$

which allows for all the probabilities to be determined using only the four values given in Table 4.1. In the case of a single clonotype with no direct competitors ($\eta = 1$), the mean time to extinction for mean niche overlap values greater than 10 is minimal (Stirk *et al.*, 2008). Therefore, the values chosen for the mean niche overlap in the competition scenarios were: 1 for self-pMHCs recognised by two clonotypes, and 10 for self-pMHCs recognised by three clonotypes. Since clonotypes 2 and 3 are considered to be homeostatically established in the periphery, their initial state will be the mean of the QSD for the competition between them rounded to the nearest integer, which with the chosen parameters is the state $(8, 8)$ (see Figure 4.4b).

Using the parameter values in Table 4.1, and the method defined in Section 4.3.2, the mean time until the first extinction event, $\hat{\tau}_{\mathbf{n}}$, was calculated for all starting states in the state space consisting of states with at most 10^2 total cells. The expected value and standard deviation of $\hat{\tau}_{\mathbf{n}}$ over all states in $\mathcal{A}^0$, $\mathbb{E}_{\mathcal{A}^0}[\hat{\tau}_{\mathbf{n}}]$ and $\sqrt{\mathbb{V}_{\mathcal{A}^0}[\hat{\tau}_{\mathbf{n}}]}$ respectively, can be found in Table 4.2. From this table it is easy to see that the introduction of clonotype 1 drastically reduces the mean time until the first extinction event occurs in the competition process. Even in competition scenario (a) in the hard niche case, which has the highest value of $\mathbb{E}_{\mathcal{A}^0}[\hat{\tau}_{\mathbf{n}}]$ of all the scenarios considered, the mean time until the first extinction is reduced to less than a fifth of its value before the introduction of the new clonotype. This strong perturbation of the time to extinction events indicates that the introduction of a new clonotype in the periphery has the potential to greatly disturb the dynamics of the competition process. This perturbation is further studied in the following sections using the other stochastic descriptors defined in Section 4.3.

	(a)	(b)	(c)	(d)	
$p^1_{1,1}$	$2/9$	$1/3$	$1/9$	0	
$p^2_{1,1}$	$2/9$	$1/3$	$5/9$	$2/3$	
$p^1_{1,2}$	$1/9$	$1/9$	$1/9$	$1/3$	
$p^2_{2,1}$	$2/9$	$2/9$	$2/9$	0	
$\nu^1_{i,0}$	0 clonotypes				
$\nu^k_{i,1}$	1 clonotype				(Stirk *et al.*, 2008)
$\nu^1_{i,2}$	10 clonotypes				(Stirk *et al.*, 2008)
$\nu^1_{1,0}$	$0, 1$ clonotypes				
φ_i	$1, 10, 10^2$ divisions $\cdot$ year^{-1}				(Lythe *et al.*, 2016)
μ_i	1 year^{-1}				(Borghans *et al.*, 2018)
$\langle n \rangle$	10 cells				

Table 4.1: Parameters for the competition scenarios shown in Figure 4.5. In this case ($\eta = 3$), the model has 11 parameters. For $\nu^1_{1,0}$, the mean niche overlap of self-pMHCs recognised only by clonotype 1, two values are given since the hard and soft niche cases are considered. The p^k_{ij} are the probabilities of clonotypes in I^k_{ij} competing for self-pMHC stimulus with clonotype i. Only four probabilities are required to determine all others using Eq. (4.88). The mean niche overlap, ν^k_{ij}, is the number of clonotypes, other than the three explicitly modelled, that recognise the self-pMHCs in $\mathscr{Q}^k_{ij}$. φ_i is the total homeostatic stimulus available to clonotype i, and μ_i is its per-cell death rate. $\langle n \rangle$ is the mean number of cells belonging to clonotypes not explicitly modelled.

4.4.1 Clonal distributions at the first extinction event

Using the method described in Section 4.3.3 the distribution of clonal sizes at the time of the first extinction event was calculated for the four different scenarios, considering the new clonotype in the hard and soft niche. Figure 4.6 shows the distribution of clonal sizes with initial state, $\mathbf{n}_0 = (4, 8, 8)$, as well as the probability for each clonotype to be the first becoming extinct, $\mathcal{U}^i_{\mathbf{n}_0}$. A triangle represents the initial state considered, $\mathbf{n}_0 = (4, 8, 8)$, and a diamond represents the mean of the resulting distribution of clonal sizes at the time of the first extinction event.

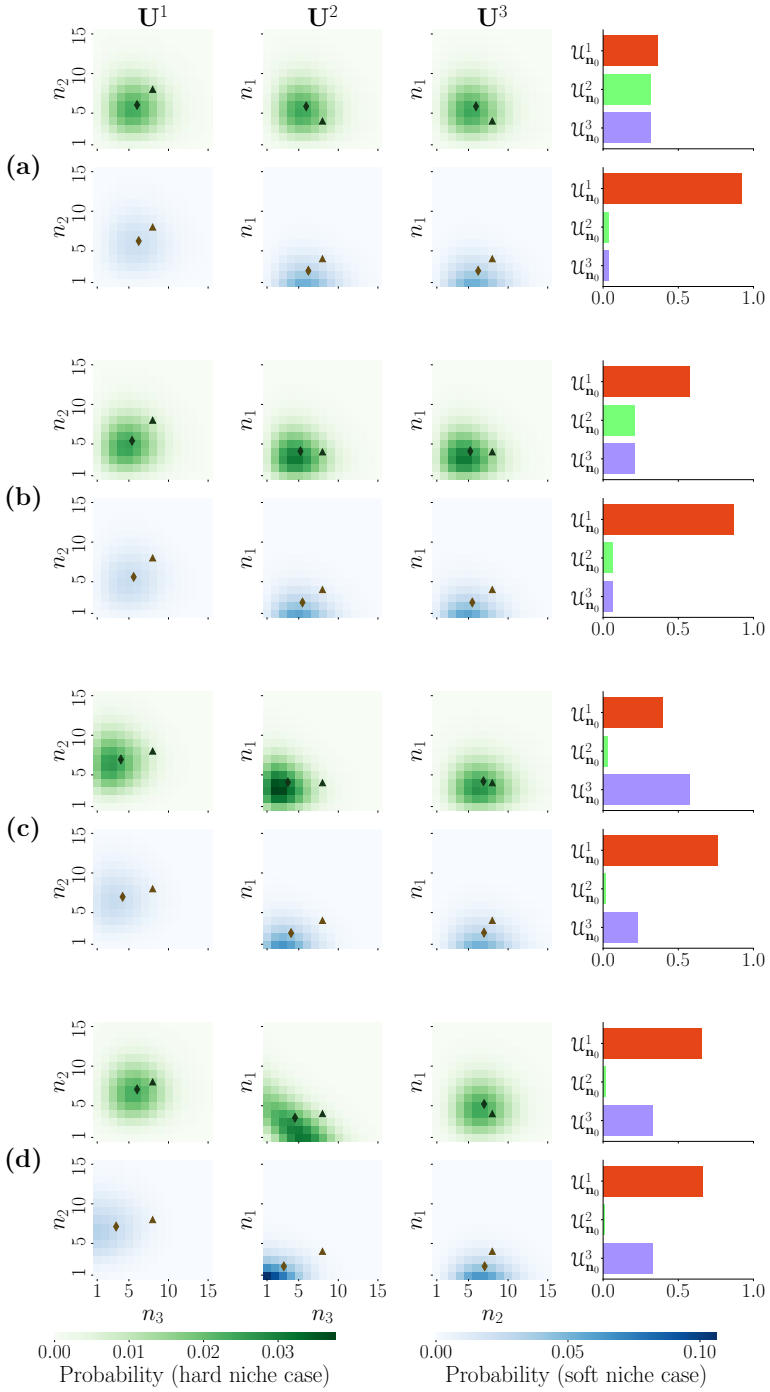


Figure 4.6: Distributions of clonal sizes at the time of the first extinction event. Each column, $\mathbf{U}^i$ for $i = 1, 2, 3$, represents the extinction of clonotype i in the hard and soft niche cases (clonotype 1), green and blue, respectively, with initial state $\mathbf{n}_0 = (4, 8, 8)$, and $\varphi_i = 10$ divisions $\cdot$ year^{-1} using the method described in Section 4.3.3 and the code in Appendix A.3. The fourth column shows the probability for each clonotype to be the first becoming extinct, $\mathcal{U}^i_{\mathbf{n_0}}$. A triangle represents the initial state and a diamond indicates the mean of the resulting distribution.

Two clonotype competition ($\eta = 2$)			
		$\mathbb{E}_{\mathcal{A}^0}[\hat{\tau}_{\mathbf{n}}]$ (years)	$\sqrt{\mathbb{V}_{\mathcal{A}^0}[\hat{\tau}_{\mathbf{n}}]}$ (years)
		122.65	5.69
Three clonotype competition ($\eta = 3$)			
Scenario	Niche ($\nu^1_{1,0}$)	$\mathbb{E}_{\mathcal{A}^0}[\hat{\tau}_{\mathbf{n}}]$ (years)	$\sqrt{\mathbb{V}_{\mathcal{A}^0}[\hat{\tau}_{\mathbf{n}}]}$ (years)
(a)	Hard	17.35	1.81
	Soft	4.94	1.25
(b)	Hard	7.59	1.21
	Soft	4.42	1.13
(c)	Hard	5.55	1.16
	Soft	3.96	1.04
(d)	Hard	3.89	1.11
	Soft	3.01	0.85

Table 4.2: Average and standard deviation (rounded to two decimal places) of the mean time to the first extinction event for the competition between two clonotypes (see Figure 4.4a). Four competition scenarios for three clonotypes were considered (see Figure 4.5), using the parameter values in Table 4.1. Numerical results obtained using the method defined in Section 4.3.2, and the code in Appendix A.2.

These results indicate that the probability of the new clonotype (clonotype 1) being the first to become extinct drastically increases when comparing the hard and soft niche cases in all the scenarios, except scenario (d). This shows that if the clonotype introduced is in the soft niche, this increases its probability of extinction, putting it at a large disadvantage against the other two homeostatically established clonotypes. In scenario (d) however, it is observed that not only does the probability of extinction of clonotype 1 change, but both $\mathcal{U}^1_{\mathbf{n}_0}$ and $\mathcal{U}^3_{\mathbf{n}_0}$ see a marginal increase. This different behaviour can be explained given that for (d) not only a new clonotype is being introduced in the system, but clonotype 3 changes from hard to soft niche, since this is a quality of the set of self-pMHCs and not of the clonotype itself. This implies that, while in scenario (d) clonotype 1 sees no significant change in its extinction probability between the hard and soft niche

cases, if it exits the thymus with more clonotypes that will also compete with it; that is, if there is a change to the niche case of other clonotypes, this will reduce their advantage in the competition for proliferation stimulus.

By comparing the probabilities of extinction for clonotype 1 in the four different scenarios, it is easy to see that in the hard niche case the most favourable scenario for its survival is scenario (a), of symmetric competition. This is the scenario in which it has less competition with the other two clonotypes overall. On the other hand, when considering the soft niche case, it is observed that the most favourable scenario is scenario (d), in which the new clonotype competes completely for stimuli with an established clonotype. One likely explanation of this seemingly counter-intuitive behaviour is that, in scenario (d) not only is clonotype 1 exerting pressure on the established clonotype but other clonotypes not explicitly modelled are doing so too, implying that the competitive pressure from clonotypes in $\mathcal{M}$ is shared between clonotypes 1 and 3.

Focusing on the cases when clonotype 1 is not the first to become extinct, it can be seen that in scenario (a) the expected population sizes have rebounded into an established state in which both surviving clonotypes have fewer cells than the mean of the QSD of two competing clonotypes. As a result the new clonotype expanded, while the established one contracted. In the soft niche, the size of the new clonotype does not bounce back to a homeostatic state, but instead both surviving clonotypes see a reduction in their cell number. In scenario (b) for the hard niche case, a move to a homeostatic state is observed again. However, there is little change in the size of the new clonotype, and a decrease in the size of the established clonotype to match the population of the surviving one. In the soft niche for this scenario a similar behaviour is observed, coupled with a reduction in the population size of clonotype 1.

In scenario (c) there is a break from the symmetry observed between $\mathbf{U}^2$ and $\mathbf{U}^3$ in the previous scenarios, since the competition considered is no longer symmetric. If clonotype 2 is the first to become extinct, the population of clonotype 3 (which has a greater competition with clonotype 1) is expected to decrease until it matches that of clonotype 1. On the other hand, if clonotype 3 becomes extinct first there is very little change in the population of clonotype 2 (which has a lower competition with clonotype 1), and only a minor change in the size of clonotype 1.

Lastly, in scenario (d) an interesting change in the shape of the clonal size distribution is observed when clonotype 2 becomes extinct first. Since clonotypes 1 and 3 directly compete for all proliferation stimuli, they behave as a single population. Thus, the distribution is no longer centred around a point, but around a line where the sum of both populations is constant. However given the uneven initial state, the distribution has more density on the end that has clonotype 3 surviving with more cells than clonotype 1. If clonotype 3 is the first to become extinct, as observed before, the population of the larger surviving clonotype decreases to match that of the new clonotype. In the soft niche there is an interesting behaviour when clonotype 2 is the first to become extinct. Since both remaining clonotypes are in the soft niche, the distribution has most of its density accumulated around the state $(1, 1)$, implying that even if these two clonotypes survive they can very easily become extinct due to their small population sizes.

Now, focusing on the $\mathbf{U}^1$ distributions, it can be seen that even when the new clonotype is the first to become extinct, it has a negative effect on the populations of the established clonotypes, reducing their average sizes in all cases. To better understand the effect the new clonotype has on the homeostatically established clonotypes, the probability for each clonotype to become extinct as a function of the initial number of cells in the new clonotype is calculated and plotted in Figure 4.7.

The first thing to note is that when comparing the hard and soft niche cases (for clonotype 1) in each scenario, the probability of clonotype 1 becoming extinct first is always higher in the soft niche than in the hard niche case. For clonotypes 2 and 3 the opposite behaviour is observed, with these clonotypes being more likely to become extinct first in the hard niche case. The only exception to this (by a minimal margin) is clonotype 3 in scenario (d), since in this case clonotype 3 has the added disadvantage of being in the soft niche, which increases its probability of becoming extinct. Another property observed is that the probability of extinction very quickly becomes saturated. That is, it shows little sensitivity to changes in the number of initial cells in the new clonotype after a certain low threshold. The threshold level depends on the competition scenario considered.

The largest difference between hard and soft niche cases is seen in scenario (a) for clonotype 1, where the probability of clonotype 1 being the first to become

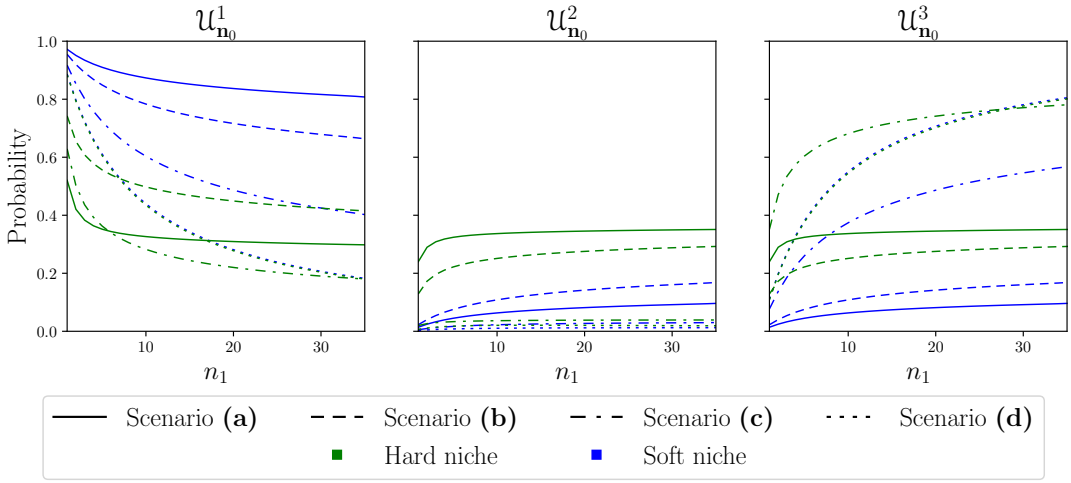


Figure 4.7: Probability of extinction for each clonotype, $\mathcal{U}^i_{\mathbf{n}_0}$, in the four scenarios with hard and soft niche cases (for clonotype 1) as a function of the initial number of cells in clonotype 1 for $\varphi_i = 10$ divisions $\cdot$ year^{-1}, calculated using the method described in Section 4.3.3 and the code in Appendix A.3. The initial number of cells of the other two clonotypes is the mean of the QSD of their two-dimensional competition process, namely $(n_2, n_3) = (8, 8)$.

extinct is almost halved when comparing the soft to the hard niche. This is due to the fact that in this case the value of $p^1_{1,1}$ is highest, meaning that the proportion of self-pMHCs shared with clonotypes 2 and 3 is the lowest. Thus, a change in the mean niche overlap has a very strong effect on clonotype 1, since it changes from a scenario of low competition to one of complete competition. This is the same reason why there is such little change in scenario (d), since clonotype 1 is already competing for all of its proliferation stimulus and a change in the mean niche overlap has a much weaker effect.

From this figure it can be determined that in the hard niche case there are different optimal competition scenarios for the survival of the new clonotype. For initial numbers of cells fewer than six, the lowest probability of extinction is given by the symmetric competition described in scenario (a). For values between 6 and 30 the optimal competition is the asymmetrical competition of scenario (b). Finally, for values above 30 complete competition with an established clonotype gives clonotype 1 the lowest probability of becoming extinct first. Interestingly,

this behaviour is not observed in the soft niche case, where there is only one optimal competition scenario for all possible initial numbers of cells, namely complete competition with another clonotype (scenario (d)).

So far only the effects of different competition scenarios have been studied. However, the effect of increasing and decreasing the total amount of stimulus available, φ, by an order of magnitude was also considered. The distribution of clonal sizes at the time of the first extinction event for the initial state $\mathbf{n}_0 = (4, 8, 8)$ with $\varphi_i = 1$ divisions $\cdot$ year^{-1} is plotted in Figure 4.8. The first thing to note from this figure is that there is minimal change on the $\mathcal{U}^i_{\mathbf{n}_0}$ probabilities, not only between scenarios but also between the hard and soft niche cases, suggesting that in low stimulus conditions the effects of the competition scenario and niche are greatly reduced. Even more so, the means of the $\mathbf{U}^i$ are also seen to be consistent across all scenarios and niche cases, further supporting the hypothesis that with low stimulus the effects of the scenario and niche are minimal.

In Figure 4.9 the probabilities of extinction, $\mathcal{U}^i_{\mathbf{n}_0}$, have been plotted as functions of the initial number of cells in clonotype 1, n_1, for the cases when $\varphi_i = 1, 10^2$ divisions $\cdot$ year^{-1}. For $\varphi_i = 1$ divisions $\cdot$ year^{-1}, due to the scarcity of stimulus, there is little effect from the competition scenario, but in the hard niche the probabilities of extinction for clonotype 1 are smaller than in the soft niche, and the opposite for the established clonotypes. The strongest effect on the extinction probabilities in this case comes from the initial number of cells. However, in contrast to Figure 4.7 the probabilities do not become saturated as quickly, meaning that in this case each cell of clonotype 1 has a weaker effect on the sizes of the established clonotypes.

For $\varphi_i = 10^2$ divisions $\cdot$ year^{-1} the opposite behaviour is observed, where the probabilities become saturated so quickly that they appear as almost completely horizontal lines, with the notable exception of scenario (d). The first thing to note is that for clonotype 2 all the probabilities are minimal except for scenario (a) in the hard niche case. In this case, given the overabundance of stimulus, the competition between clonotypes has no effect and all clonotypes are equally likely to be the first to become extinct. In scenario (b), for both hard and soft niche cases, the increased competition causes the extinction probability to accumulate on clonotype 1. The asymmetric competition of scenario (c) accumulates most

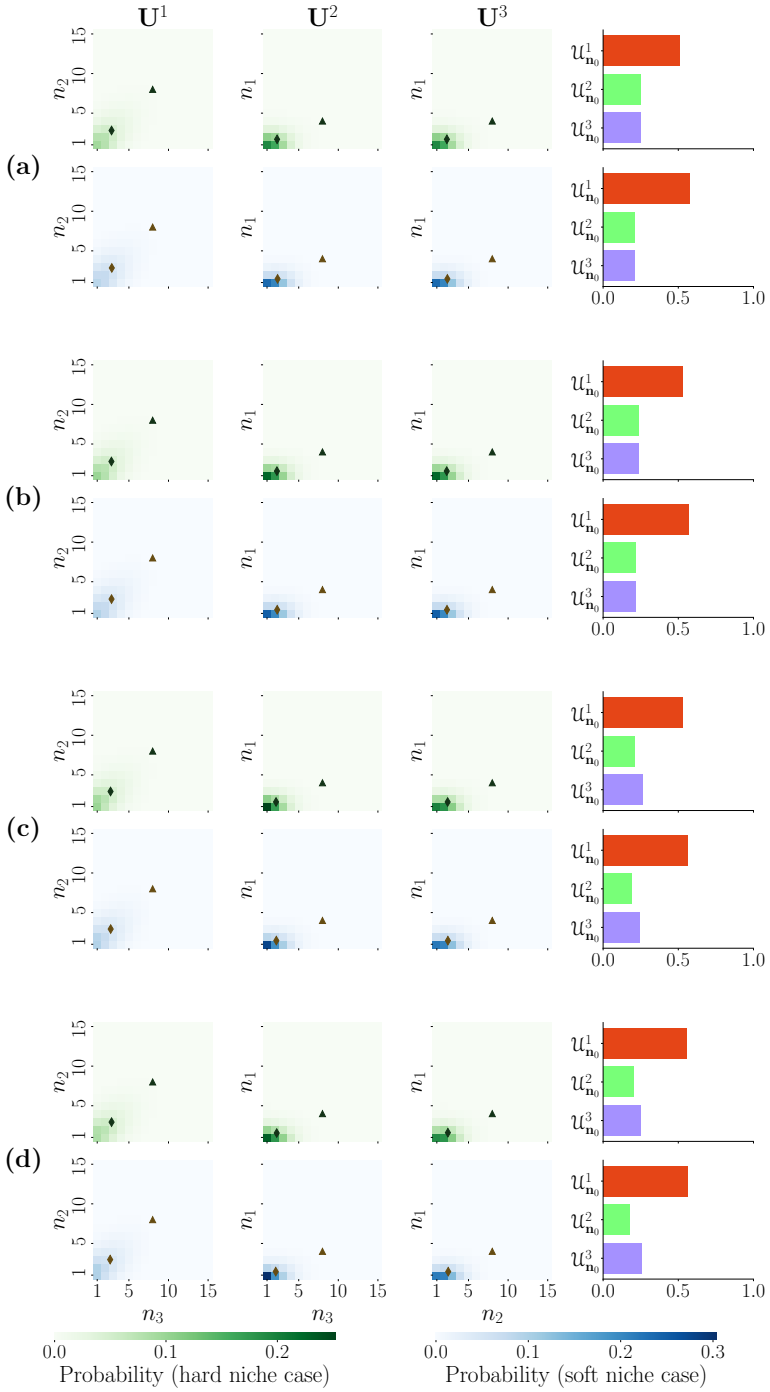


Figure 4.8: Distributions of clonal sizes at the time of the first extinction event for the initial state $\mathbf{n}_0 = (4, 8, 8)$, and $\varphi_i = 1$ divisions $\cdot$ year^{-1} calculated using the method described in Section 4.3.3 and the code in Appendix A.3. The fourth column shows the probability for each clonotype to be the first becoming extinct, $\mathcal{U}^i_{\mathbf{n_0}}$. A triangle represents the initial state and a diamond indicates the mean of the resulting distribution.

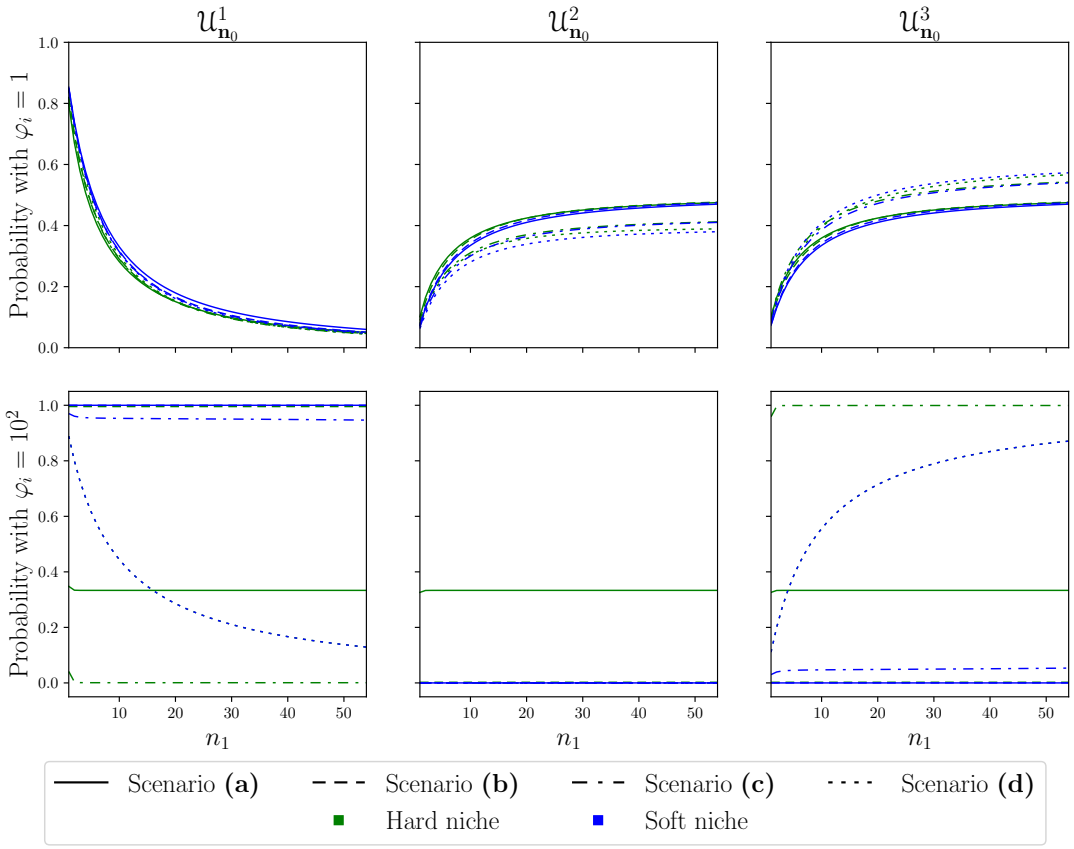


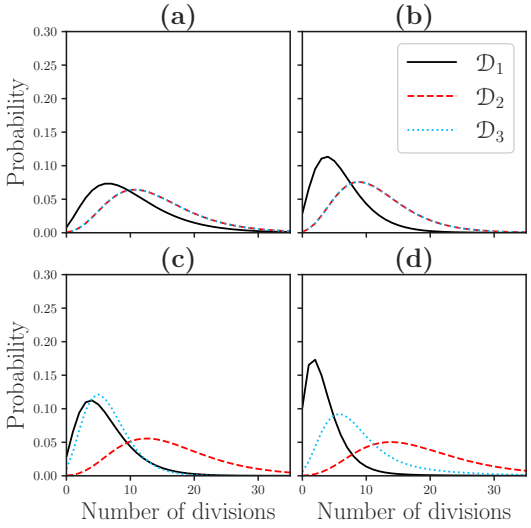
Figure 4.9: Probability of extinction for each clonotype in the four scenarios for the hard and soft niche cases (for clonotype 1) as a function of the initial number of cells in clonotype 1 for $\varphi_i = 1, 10^2$ divisions $\cdot$ year^{-1} (top and bottom row, respectively). The probabilities were calculated using the method described in Section 4.3.3 and the code in Appendix A.3. The initial number of cells in the other two clonotypes is the same as in Figure 4.7, namely $(n_2, n_3) = (8, 8)$.

of the probability of extinction on clonotype 1 and 3 in the soft and hard niche cases, respectively. Finally, in scenario (d) the behaviour is similar to that of $\varphi_i = 1, 10$ divisions $\cdot$ year^{-1}, where the probabilities of clonotypes 1 and 3 are coupled and the difference between hard and soft niche is minimal. This behaviour stems from the fact that there is complete competition for stimulus, and regardless of its abundance, the competitive exclusion principle (Hardin, 1960) states that one population must become extinct. With every other aspect of the populations being

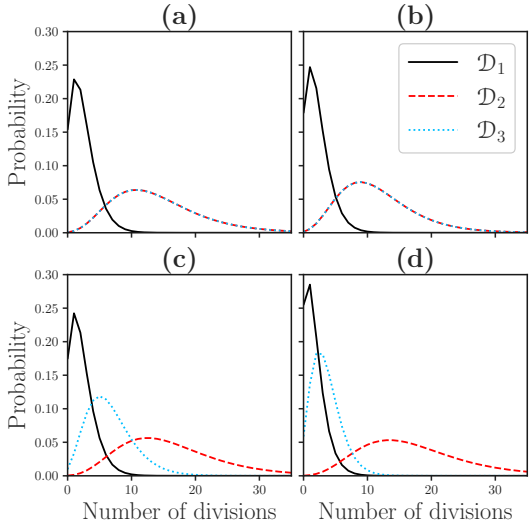
equal, this means that initial conditions completely determine which clonotype is most likely to become extinct.

4.4.2 Number of divisions before extinction of a clonotype

Now, the probability distribution of the number of divisions before extinction of each clonotype, $\mathcal{D}_i$, is calculated making use of the method described in Section 4.3.4. Figure 4.10 shows these distributions for the four scenarios in the hard and soft niche cases (for clonotype 1) for the initial state $\mathbf{n}_0 = (4, 8, 8)$.



(A) Probability distribution of the number of divisions before extinction in the hard niche case (for clonotype 1).



(B) Probability distribution of the number of divisions before extinction in the soft niche case (for clonotype 1).

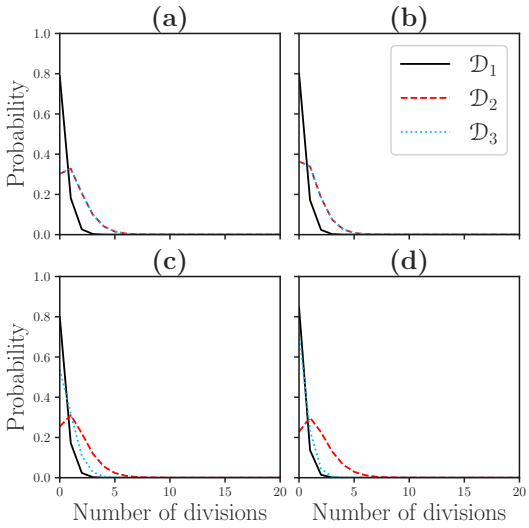
Figure 4.10: Probability distribution of the number of divisions before extinction for the initial state $\mathbf{n}_0 = (4, 8, 8)$, and $\varphi_i = 10$ divisions $\cdot$ year^{-1} calculated using the method described in Section 4.3.4 and the code in Appendix A.4. (A) shows the distributions when clonotype 1 is in the hard niche case and (B) in the soft niche case.

For scenarios (a) and (b), in both the hard and soft niche cases (for clonotype 1), the division distributions for clonotypes 2 and 3 are very similar, since they are not only competing in the same way but have the same initial number of cells. In

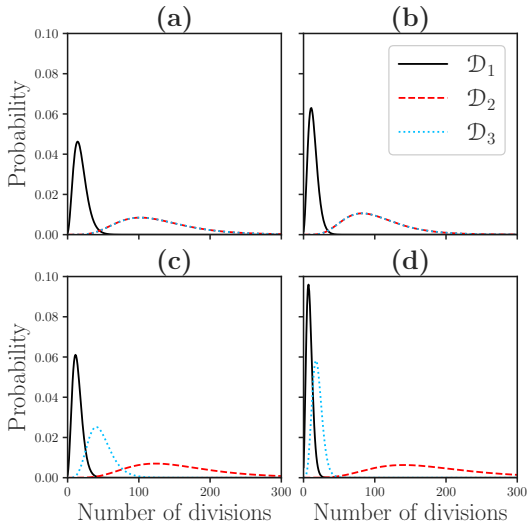
scenario (c), the distribution for clonotype 3 is shifted to the right when compared to clonotype 2, due to the increase in competition for clonotype 3, but again there is no notable difference between the hard and soft niche cases. There is a break in these similarities between the hard and soft niche cases in scenario (d). Here, the distribution of divisions of clonotype 3 is narrower and centred around smaller values in the soft niche case when compared to the hard niche case. This is due to the fact that the new clonotype being in the soft niche changes all self-pMHCs recognised by clonotype 3 from the hard to the soft niche, greatly reducing its probability of dividing a greater number of times before becoming extinct.

Focusing on the new clonotype (clonotype 1), it can be seen that regardless of the niche considered in scenarios (a) and (b), the median of the number of divisions is lower than the other modelled clonotypes, with this difference being greater in the soft niche case. In scenario (a), this behaviour is caused by the smaller population size of the new clonotype (since every other aspect of the competition is symmetric). The behaviour is amplified in scenario (b) given the increased competition experienced by the new clonotype. In scenario (c), in the hard niche case the distribution of divisions of the new clonotype is very similar to that of clonotype 3. This is caused by the change in the distribution for clonotype 3 due to its increased competition. However, this similarity is undone in the soft niche case where the distribution for the new clonotype is narrower and has a smaller median. This agrees with the previous observation in Figure 4.7: the niche in which the new clonotype is has a greater impact on its fate than its competition scenario. Finally, in scenario (d) in the hard niche case the distribution already has a low median and is rather narrow. This behaviour is stronger in the soft niche case. In the hard niche case, scenario (d) has the lowest median and the most narrow distribution for the new clonotype overall, followed by the distributions of scenarios (b), (c), and (a) in that order. This agrees with the probabilities of extinction for the initial state $\mathbf{n} = (4, 8, 8)$ in Figure 4.7, where it can be seen that these probabilities decrease in the same order. In the soft niche case there is a direct relation between the probabilities of extinction and distribution of divisions for scenarios (a), (b), and (c), but not for scenario (d). This can be explained by the fact that the probabilities of extinction are changing due to an increased

chance of clonotype 3 becoming extinct, and not a direct change to the extinction of the new clonotype.



(A) Probability distribution of the number of divisions before extinction for $\varphi_i = 1$ divisions $\cdot$ year^{-1}.



(B) Probability distribution of the number of divisions before extinction for $\varphi_i = 10^2$ divisions $\cdot$ year^{-1}.

Figure 4.11: Probability distribution of the number of divisions before extinction for the initial state $\mathbf{n}_0 = (4, 8, 8)$ in the soft niche case calculated using the method described in Section 4.3.4 and the code in Appendix A.4. (A) shows the distributions for $\varphi_i = 1$ divisions $\cdot$ year^{-1} and (B) for $\varphi_i = 10^2$ divisions $\cdot$ year^{-1}.

In Figure 4.11 the distribution of divisions in the soft niche case for the initial state $\mathbf{n}_0 = (4, 8, 8)$ is plotted for $\varphi_i = 1$ divisions $\cdot$ year^{-1} (Figure 4.11A), and for $\varphi_i = 10^2$ divisions $\cdot$ year^{-1} (Figure 4.11B). Note that in the case $\varphi_i = 1$ divisions$\cdot$year^{-1} the distributions remain mostly unchanged between all scenarios, in agreement with what was observed about the $\mathcal{U}^i_{\mathbf{n}_0}$ probabilities in Figure 4.8. Then, these distributions corroborate the hypothesis that in low stimulus conditions the effects of the competition scenario and the niche are greatly diminished, and it is the initial conditions that have the strongest effect on proliferative capacity, and the probability of extinction of a clonotype. For the case $\varphi_i = 10^2$ divisions$\cdot$year^{-1} a different behaviour is observed, where the effects of the competition scenario are markedly present, but the support of the distribution is much larger than that for

the other values of φ_i, with some distributions having non-negligible values for over 250 divisions before extinction. However, the difference between clonotypes is much stronger, reflecting the quick saturation of the extinction probabilities observed in Figure 4.9.

Altogether, then, by comparing these distributions to that for $\varphi_i = 10$ divisions$\cdot$ year^{-1} (Figure 4.10B) it is easy to see that the choice of φ_i is critical for the competition model. Since very low values will produce models in which there is little to no proliferation and the competitive dynamics is not captured, while larger values generate models that do not properly describe the effects of competition, due to the fact that excess stimulus naturally changes the effects of competition.

4.5 Discussion

Maintaining the diversity of the T cell repertoire is essential for the immune system to mount a strong and effective immune response (Kedzierska *et al.*, 2006; Nikolich-Žugich *et al.*, 2004; Qi *et al.*, 2014b; Yager *et al.*, 2008). Promoting survival of significantly different clonotypes maximises this diversity by allowing TCRs with similar self-pMHC recognition profiles to become extinct. However, it has been observed that several clonotypes often overlap on their recognition profile (*e.g.*, Duan *et al.*, 2015) as a compromise between TCR diversity and coverage over the space of foreign antigens. This motivates the multi-variate representation developed in this chapter, which extends and generalises that presented in Stirk *et al.* (2010) to three or more clonotypes.

It was shown that the proliferation of competing clonotypes decreases as the overlap in their recognition profiles increases. That is, the effects of competitive exclusion can be seen as the distribution of the number of divisions moves to the left when there is an increase in the competition for self-pMHCs. Furthermore, it was proved that extinction of all clonotypes is certain, and the mean time to this absorption event is bounded by a value that depends on the total number of cells of all clonotypes. A feature of the proposed competition process, prior to the first extinction event, is that the system is driven to a state where all clones have low numbers of cells, with the specific number of cells depending on their recognition overlap. This agrees with the biological understanding that in a homeostatic state

each clonotype consists of only a few cells, with the exact number varying from clonotype to clonotype (Casrouge *et al.*, 2000; Lythe *et al.*, 2016).

The multi-variate competition model presented here shows that the introduction of a new clonotype to the periphery perturbs the homeostatically established clonotypes. This is observed in the decreased mean number of cells after the first extinction event when compared to the initial state of the process. Even in the scenario with the least competition for stimulus (scenario (a)) there is a perturbation of the established clonotypes. More than this, the probability of extinction as a function of the initial number of cells in the new clonotype very quickly becomes saturated (see Figure 4.7).

From the distribution of divisions before extinction (see Figure 4.10), it can be concluded that in the soft niche case the new clonotype is expected to experience very little proliferation before becoming extinct. Even in scenario (a), where competition with other modelled clonotypes is the lowest, the distribution of divisions is very narrow and centred around a value less that five. This major change outlines an important feature of the soft niche: even when considering a mean niche overlap of only ten clonotypes the probability of extinction is greatly increased and the distribution of divisions is made narrower and its mode moves to the left. This major change in behaviour between the hard and soft niche cases can be interpreted in two ways: (*i*) for the soft niche assumption to be correct in the naive T cell repertoire, the mean niche overlap value must be a small number of clonotypes. Otherwise, as observed in the case $\varphi_i = 10^2$ divisions $\cdot$ year^{-1} (see Figure 4.9), a new clonotype entering the periphery would have a very low probability of becoming established, and the naive T cell repertoire would become mostly static; or (*ii*) if the mean niche overlap is not small then cells in the soft niche must have a very fast turnover, and make up for most of the thymic output to counteract this low probability of proliferation and establishment. These two interpretations are made under the assumption that the thymic output is mostly homogeneous; that is, most of the cells exiting the thymus are either on the hard or soft niche. By relaxing this assumption and considering a more heterogeneous thymic output, the naive T cell repertoire can be thought of as being divided into two major compartments: one compartment comprised of clonotypes in the hard niche, which is deeply established (with low extinction

probabilities) and is expected to remain mostly constant through life, and another compartment of clonotypes in the soft niche, which is constantly changing and is more dynamic through life (as clonotypes appear from the thymus and become extinct due to clonal competition). This type of heterogeneity would allow for the maintenance of a T cell repertoire that has the capability of constantly producing new clonotypes (in the soft niche), while maintaining other clonotypes (in the hard niche) throughout life.

So far only the possible effects of a distribution of soft and hard niche clonotypes exiting the thymus have been discussed, with the assumption that thymic output is constant throughout the ageing process (Goronzy *et al.*, 2007; Qi *et al.*, 2014b; Zhang *et al.*, 2021). However, given the known changes in thymic output through life, such as thymic involution (Aspinall & Andrew, 2000; Lynch *et al.*, 2009), and other age-related changes (Srinivasan *et al.*, 2021), it is plausible that the composition of thymic emigrants (in terms of hard and soft niche, or clonal size) changes as well during the lifespan of a host. Therefore, if instead of a constant mixture of hard and soft niche clonotypes, a mixture that evolves from mostly hard niche clonotypes during foetal stages to mostly soft niche clonotypes during adulthood is considered, then the model predicts the early establishment of clonotypes that exited the thymus early in life, and then a declining supply of mostly short-lived clonotypes later in life. This decline in the production of soft niche clonotypes could be justified by the fact that during the initial development of the T cell repertoire, fewer clonotypes are expected to be present in the periphery, and therefore smaller values of the mean niche overlap ν_{ij}^{k}. Furthermore, this behaviour is compatible with the analysis by Gaimann *et al.* (2020), in which T cell clonotype sizes were found to follow a power-law distribution, where clonotypes generated during the early foetal stages (characterised by no nucleotide insertions during V(D)J recombination (Park *et al.*, 2020)) were found to be the most enriched in the periphery.

As shown in Figures 4.6 and 4.8, the methods to calculate the distribution of clonal sizes at the time of the first extinction event defined in Section 4.3.3 permit a very detailed study of the competition process at the time of the first extinction, by providing not only the probabilities of each clonotype being the first to become extinct but also the distribution of population sizes conditioned

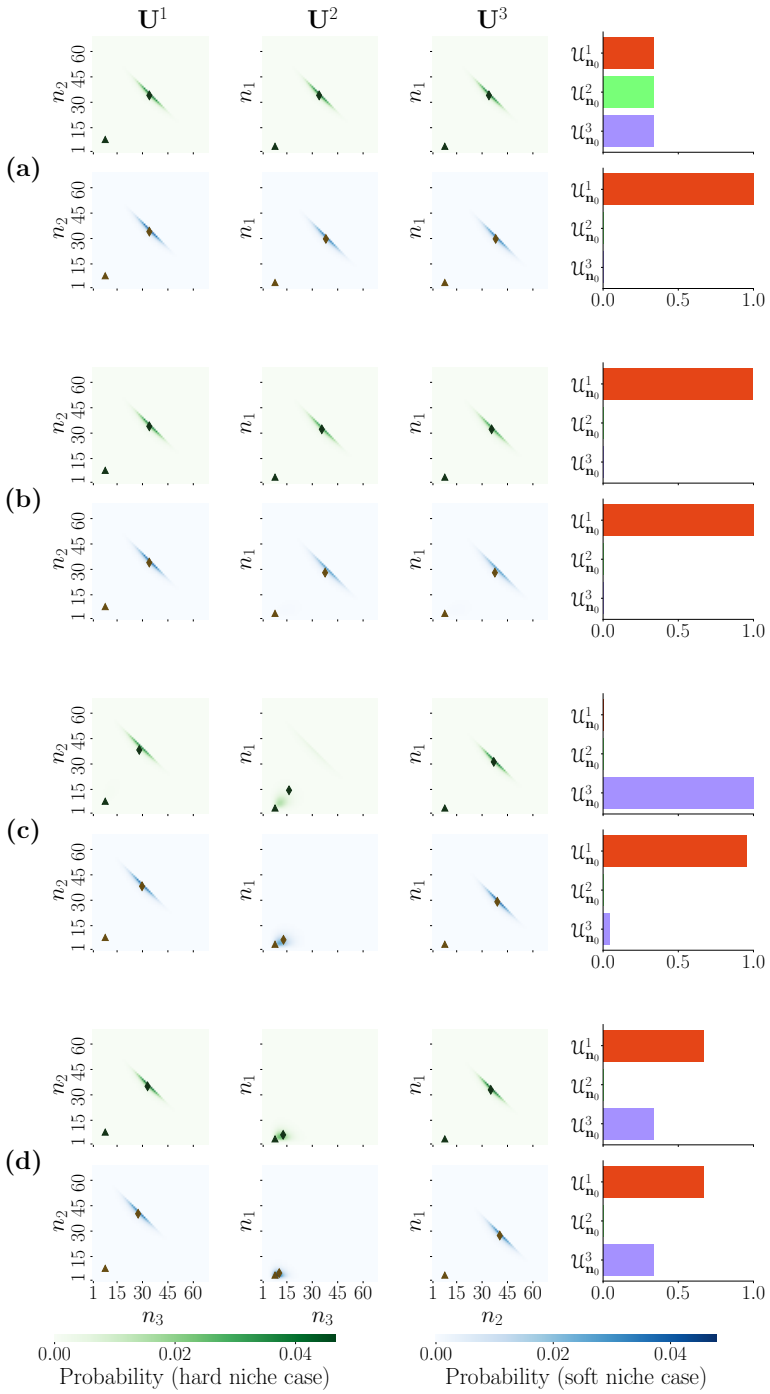


Figure 4.12: Distributions of clonal sizes at the time of the first extinction event for the initial state $\mathbf{n}_0 = (4, 8, 8)$, and $\varphi_i = 10^2$ divisions $\cdot$ year^{-1} calculated using the method described in Section 4.3.3 and the code in Appendix A.3. The fourth column shows the probability for each clonotype to be the first becoming extinct, $\mathcal{U}^i_{\mathbf{n_0}}$. A triangle represents the initial state and a diamond indicates the mean of the resulting distribution.

on the extinction of each clonotype. The method to calculate the distribution of divisions before extinction defined in Section 4.3.4 provides better understanding of proliferation of clonotypes in the competition process before extinction events. Finally, the approximations for the QSD defined in Section 4.2 provide the tools for the study of the system conditioned on non-extinction, which allows for the study of perturbations of already established clonotypes as discussed in Section 4.5. And all these methods are based on the same matrix analytic analysis of first step arguments.

However, a strong limitation of these matrix analytic methods is the computational cost of the linear level-reduction algorithm. While it is far less computationally expensive than directly solving a system of equations, such as that presented in Eq. (4.31), or than directly inverting the matrices in Equations (4.74) and (4.85), the size of the sub-matrices used in the level-reduction algorithm grows at such a rate that it becomes a limiting factor of the computation. See for example Figure 4.12, here the distribution of clonal sizes has been calculated for the competition scenarios defined in Section 4.5 with $\varphi_i = 10^2$ divisions $\cdot$ year^{-1}. While the values of $\mathcal{U}^i_{\mathbf{n}_0}$ can be obtained with the linear level-reduction algorithm, the $\mathbf{U}^i$ distributions accumulate most of their mass at the reflecting plane that truncates the state space. However, due to the computational cost of solving the matrix equation for the next level, it is not possible at the time to consider a larger state space.

This limitation applies to all the methods that rely on linear level-reduction algorithms to solve matrix equations. In the case of the QSD, however, a different approximation which does not rely on matrix analytic methods was introduced in Section 4.2.3. This linear noise approximation provides an approximation of the stochastic competition process in the form of a deterministic limit plus noise (van Kampen, 2007). In order to compare the methods defined in Section 4.2.2, and Section 4.2.3 the Hellinger distance between the distributions (Oosterhoff & van Zwet, 2012) was calculated using the code presented in Appendix A.1.3, and the results presented in Table 4.3. First, note that the two approximating processes result in distributions that are relatively different to one another, with a Hellinger distance of 0.293 on their most similar scenario. This is not unexpected however, since both approximating processes consider different modifications of

Hellinger distance between approximations of the QSD			
Competition scenario	$\mathcal{X}^{(1)}$ – $\mathcal{X}^{(2)}$	LNA – $\mathcal{X}^{(1)}$	LNA – $\mathcal{X}^{(2)}$
	Hard niche		
(a)	0.2755	0.0836	0.2604
(b)	0.3109	0.1025	0.2551
(c)	0.3129	0.133	0.2332
(d)	0.2973	0.1387	0.2258
	Soft niche		
(a)	0.293	0.3076	0.2118
(b)	0.3025	0.3069	0.2052
(c)	0.3121	0.3272	0.1741
(d)	0.7071	0.7071	0.1685

Table 4.3: Comparison of the linear noise approximation and the approximating processes $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$ using using the Hellinger distance. The competition scenarios and parameters chosen in Section 4.4 were used (see Table 4.1), and the Hellinger distance was calculated with the code presented in Appendix A.1.3. In the hard niche case the linear noise approximation is more similar to the $\mathcal{X}^{(1)}$ process than the $\mathcal{X}^{(2)}$ for all competition scenarios. On the other hand, in the soft niche case the reverse is observed, where the linear noise approximation is closer to the $\mathcal{X}^{(2)}$ process under the Hellinger distance.

the population dynamics, one introducing an immortal individual and the other removing the absorbing state (see Section 4.2.2).

From these results, it is easy to see that in the hard niche case the linear noise approximation results in a distribution very similar to the $\mathcal{X}^{(1)}$ approximating process. Thinking about the basis of this approximating process it should not be unexpected that the linear noise approximation will resemble it, since the modification of this model is the removal of the absorbing states. This change to the state space forces the model to wander stochastically around its QSD, which is qualitatively similar to the way a deterministic process behaves. Interestingly, this

similarity is lost in the soft niche case, where the distance to both approximating processes is large (there is a decrease with respect to $\mathcal{X}^{(2)}$ but the distances still remain high), suggesting that the stochastic nature of the approximating processes is most important in the soft niche case, which as seen in Section 4.4 is the case with higher probabilities of extinction and the lower mean times to extinction.

In general, the estimation of the QSD using the auxiliary process $\mathcal{X}^{(1)}$ approximates the distribution from below (resulting in a lower bound for the QSD), while process $\mathcal{X}^{(2)}$ approximates it from above (resulting in an upper bound for the QSD). The linear noise approximation in general estimates the mean of the QSD between these two processes, and is computationally less expensive than the matrix analytic methods necessary for these approximations. However, given the fact that the noise obtained from the linear noise approximation is Gaussian (Elf & Ehrenberg, 2003; van Kampen, 2007), depending on the type of results expected (an approximation of the mean or a more accurate approximation of the distribution) the auxiliary process approximation might be the ideal option, regardless of the computational cost.

Chapter 5

Perturbation of the distribution of T cell clonotypes by viral infection

In the previous chapter the dynamics of naive T cell populations was studied, specifically in the context of a new clonotype entering the periphery and competing with already established clonotypes. However, even greater perturbations of the balance observed during homeostasis come from the expansion of T cell populations during infection (Busch & Pamer, 1999; Butz & Bevan, 1998; Lawrence *et al.*, 2005), which can have long lasting effects on individual immunity to future infections (Gil *et al.*, 2015; Lanfermeijer *et al.*, 2020; Yang *et al.*, 2022). In particular, this chapter will focus on heterologous cross-reactive influenza A virus (IAV) infections. That is to say, a history of two influenza infections of distinct sub-types, which elicit cross-reactive immune responses and confer some measure of immunity to strains that have not been encountered by an individual before (Duan *et al.*, 2015; Hillaire *et al.*, 2013; McMichael *et al.*, 1983; Sridhar *et al.*, 2013).

In this context it is not well known specifically how the history of infection impacts future exposures to IAV. A mathematical model has been proposed to better understand the age distribution of cases, and severity of future pandemics based on cross-reactive immunity acquired during childhood (Gostic *et al.*, 2016). However, this model only interprets already known population level information of antibody cross-reactivity, but not how cross-reactive T cell responses behave on a cellular level. A better understanding of this dynamics would prove useful in vaccine development to target specifically the desired cross-reactive responses that

would confer immunity to a wider variety of pathogens, which have been observed but have not been replicated (Elong Ngono & Shresta, 2019; Mateus *et al.*, 2021; Moris *et al.*, 2011; Webster & Askonas, 1980).

The bipartite recognition network introduced in Chapter 4 provides a good platform to introduce viral infection into the model. While the initial definition explicitly considers only self-pMHCs used for homeostatic maintenance, the network encodes only TCR-epitope interactions, meaning that it can easily be extended to include virus derived peptides (VDPs) and simulate an infection. These VDPs can be arbitrarily separated into subsets, used to represent different qualities of the epitopes. Since the main focus of this chapter is cross-reactivity in the context of heterologous infection, the subsets are used to represent the specific peptides presented in the context of MHC during different infections. However, this is not the only possible interpretation that these sets can be given. For example, if multiple infections are not the desired focus of study the sets can be used to represent different levels of affinity, avidity, or immunogenicity.

The aim of this chapter is to propose a mathematical model to study the effects and dynamics of cross-reactive immune responses. To this end, in collaboration with the Paul Thomas laboratory at St. Jude Children's Research Hospital, data from heterologous influenza infection experiments on mice, carried out and designed by Jessica Gaevert, was studied in order to better understand the nature of the cross-reactive responses being observed.

To develop a stochastic model for heterologous viral infection, first, the bipartite recognition network is generalised to a k-partite recognition network in Section 5.1.1. In Section 5.1.2 the naive, effector, and memory T cell phenotypes are introduced as cellular compartments in the model, as well as the differentiation pathway that will connect them to each other during homeostasis and infection. In Section 5.1.3 the stochastic model is formally defined, and an important result on the nature of the competition process is presented, namely the certainty of extinction of the T cell population for sufficiently late times for primary infections, but not in all cases of challenge infections. Due to the large population sizes observed during infection, the van Kampen Large N expansion (van Kampen, 2007) is used to find a deterministic limit of the stochastic process in Section 5.2. The experimental data provided by the Paul Thomas laboratory at St. Jude Children's

hospital is presented, and summarised in Section 5.3.1, and the limitations of the deterministic model are discussed in Section 5.3.2. Finally, the expansion and contraction behaviour of the T cell populations in the data are analysed, and the results presented in Section 5.4.

5.1 Stochastic model with viral infection

In this section the stochastic model presented in Chapter 4 is extended to include viral infection and the cellular dynamics associated with it. To do this, first virus derived peptides (VDPs) are introduced as an extension of the bipartite recognition network (van Stipdonk *et al.*, 2001). Then, the extended recognition network is used to define a stochastic process that models the populations of naive, effector, and memory cells during homeostasis and acute infection.

5.1.1 k-partite network of TCR-peptide recognition for infection

In Chapter 4 a bipartite network was used to represent the recognition of self-pMHCs by T cells, and a stochastic model was defined to describe naive T cells in homeostasis. In order to extend this model to include the presentation of VDPs, first the bipartite recognition network must be extended to a k-partite network of T cell clonotypes and $k-1$ different types of peptides. An example of a k-partite network with T cell clonotype nodes, self-pMHC nodes, and three different sets of VDPs is shown in Figure 5.1.

In general the network can be extended to have as many parts as desired, and they can be used to represent different types of peptides. For example, if the aim is to study the affinity of peptides, be it self-peptides or virus derived peptides, they can be classified into being high or low affinity. Since the aim of this chapter is to study the dynamics of cross-reactivity in the context of heterologous infections, the recognition network considered will consist of T cell clonotypes, self-pMHCs, and VDPs divided into those presented during a first infection, and those presented during a second infection. However, given that for some infections conserved epitope presentation across variants is a common occurrence, in particular for

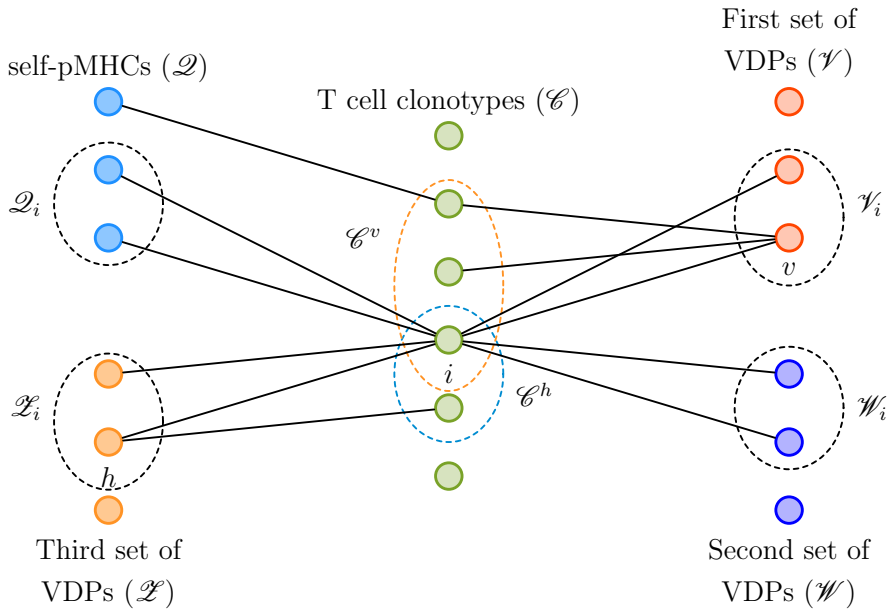


Figure 5.1: Example of a 5-partite recognition network. The 5 types of nodes in the network are: T cell clonotypes nodes ($\mathscr{C}$), self-pMHCs ($\mathscr{Q}$), and three sets of different VDPs ($\mathscr{V}$, $\mathscr{W}$, and $\mathscr{Z}$). An edge between a clonotype node and a peptide node represents the ability of that clonotype to receive stimulus from that peptide. For each clonotype i, there are sets $\mathscr{Q}_i$, $\mathscr{V}_i$, $\mathscr{W}_i$, and $\mathscr{Z}_i$ of peptides of each type it can recognise. For every peptide q there is set $\mathscr{C}^q$ of clonotypes that can recognise it.

influenza A viruses (Eickhoff *et al.*, 2019; Ekiert *et al.*, 2009; Tan *et al.*, 2011), the sets of VDPs presented during infection do not need to be disjoint.

5.1.2 T cell phenotypes and differentiation pathway

To model the dynamics of T cell expansion during infection, and contraction back into homeostatic levels, it is not enough to include VDPs in the model (Kaech & Cui, 2012). The three most functionally relevant T cell phenotypes, namely naive, effector, and memory T cells must be considered (Ahmed & Gray, 1996; Bevan & Fink, 2001; Doherty & Christensen, 2000). Naive T cells are already modelled using the stochastic process defined in Chapter 4, therefore only effector and memory cells need to be added to the model.

To include the three different T cell phenotypes, the stochastic model will consist of three compartments (naive, effector, and memory) for each clonotype. Since these compartments are related to each other by a biological differentiation process, a differentiation pathway is also necessary to be able to model transitions between the compartments. While the differentiation process is not yet fully understood, some hypotheses have been already proposed for the mechanisms with which T cells differentiate from one compartment to another (Kaech & Cui, 2012).

Based on the hypotheses presented by Kaech & Cui (2012), the differentiation pathway shown in Figure 5.2 is considered for the stochastic competition process. In this differentiation pathway, N, E, and M represent the naive, effector, and memory compartments, respectively. Red arrows represent the death events for all three compartments (occurring with rates μ_N, μ_E, and μ_M), black arrows represent homeostatic proliferation events for the naive and memory compartments (with rates λ_N, and λ_M, respectively), blue arrows represent VDP mediated differentiation events (with rates α_N, and α_M) and division events (with rate λ_E). Finally, the purple wavy arrow represents the differentiation event from the effector to the memory compartment (occurring with rate ψ_E). This last arrow is different from the others, since in the model this differentiation event only occurs after the clearance of the infection.

Now, since transitions between compartments represent differentiation into a new phenotype, the available sources of stimuli for each compartment must also be specified. As discussed in Chapter 4, naive cells receive homeostatic proliferation stimuli from self-pMHCs, and to differentiate into the effector compartment they must receive stimuli from VDPs they are able to recognise (Croft *et al.*, 2019; Luciani *et al.*, 2013; van Stipdonk *et al.*, 2001). Effector cells, on the other hand, require constant VDP stimulation to divide and cannot receive survival stimulus from self-pMHCs (Huppa *et al.*, 2003). The exact mechanism for differentiation from effector to memory phenotype is not yet fully understood. It has been hypothesised that memory is generated by selecting for cells that have not been overly stimulated by the foreign peptides, and thus, have not terminally differentiated into effector cells (Badovinac *et al.*, 2005; Joshi *et al.*, 2007; Sarkar *et al.*, 2008), or by simple asymmetric division of naive cells (Chang

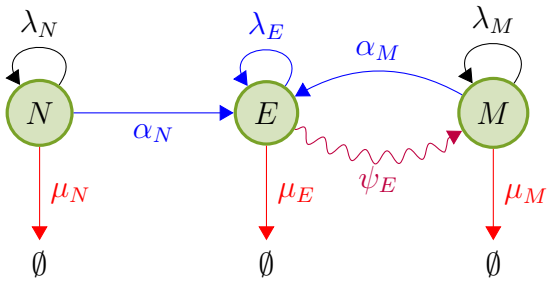


Figure 5.2: Differentiation pathway for T cell phenotypes during an immune response to a viral infection. The compartments considered are naive (N), effector (E), and memory (M). Cells in N are homeostatically maintained with rate λ_N, die with rate μ_N, and differentiate into the E with rate α_N. Cells in E divide with rate λ_E, die with rate μ_E, and differentiate into M with rate ψ_E only once the infection has been cleared (this condition is represented by a wavy line). Finally, cells in M are homeostatically maintained with rate λ_M, die with rate μ_M, and differentiate back into E with rate α_M.

et al., 2007). However, it is known that only between 5–10% of the effector population present at the time the infection is cleared from the system goes on to differentiate to the memory phenotype (Ahmed & Gray, 1996). Finally, memory cells are independently maintained from naive cells during homeostasis via cytokine signalling (Harty & Badovinac, 2008), and they can become activated and differentiate back into the effector compartment via VDP stimulation (Lauvau & Soudja, 2015). These restrictions on pMHC stimulation are summarised in Figure 5.3 for a recognition network comprised of self-pMHCs, and two different infections.

5.1.3 Stochastic competition process for T cell clonotypes during infection

To model the dynamics of η T cell clonotypes during homeostasis and infection, the following continuous-time Markov process is used

$$\mathcal{X} = \{(X_1(t), \ldots, X_\eta(t), Y_1(t), \ldots, Y_\eta(t), Z_1(t), \ldots, Z_\eta(t)) : t \geq 0\}\,, \tag{5.1}$$

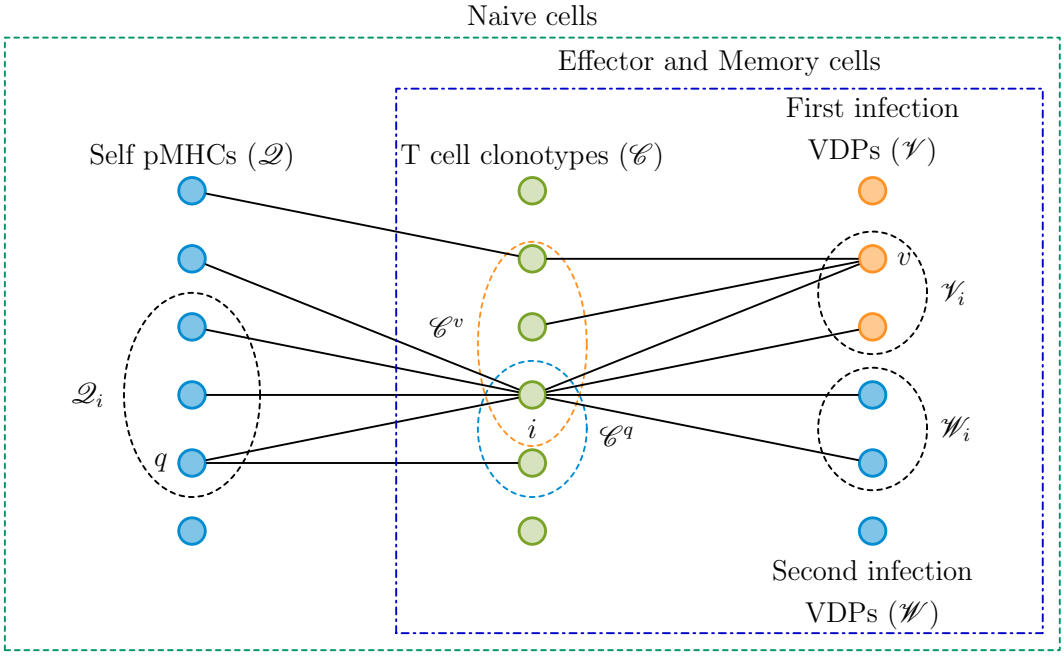


Figure 5.3: Available sources of pMHC stimulus for each T cell phenotype in a recognition network with two infections. Naive cells are able to receive stimulus from both self-pMHCs for homeostatic maintenance, and VDPs for differentiation into effector phenotype. Effector cells can only receive stimulus from VDPs in order to proliferate during an infection, and naturally die out once the infection is cleared. Finally, memory cells can only receive stimulus from VDPs in order to differentiate back to the effector compartment, since they are homeostatically maintained by cytokine signalling (Harty & Badovinac, 2008), and therefore are not modelled to receive stimulus from self-pMHCs.

where, for every clonotype $1 \leq i \leq \eta$, $X_i(t)$ represents the number of naive cells at time t, $Y_i(t)$ the number of effector cells at time t, $Z_i(t)$ the number of memory cells at time t, and the random vector

$$\mathbf{X}(t) = (X_1(t), \ldots, X_\eta(t), Y_1(t), \ldots, Y_\eta(t), Z_1(t), \ldots, Z_\eta(t)) \tag{5.2}$$

describes the populations of all naive, effector, and memory T cells at time t. Since the number of cells present at time t can only be a positive integer or 0, the state space of this Markov process is

$$\mathcal{S} = \{(n_1, \ldots, n_\eta, e_1, \ldots, e_\eta, m_1, \ldots, m_\eta) : n_i, e_i, m_i \geq 0, \forall i\} = \mathbb{N}_0^{3\eta}. \tag{5.3}$$

To define the birth, death, and differentiation rates of the different phenotypes, the following four sets must be considered: the set $\mathscr{C}$ of η different clonotypes, each with three compartments called naive, effector, and memory, the set $\mathscr{Q}$ of all self-pMHCs that the naive populations in $\mathscr{C}$ can receive stimulus from, and two sets $\mathscr{V}$, and $\mathscr{W}$ of virus derived peptides (VDPs) which are presented during two subsequent infections and the clonotypes in $\mathscr{C}$ are able to recognise. Since the homeostatic proliferation rate of naive cells is not affected by the dynamics of infection, given that the naive phenotype is the only one able to receive homeostatic proliferation stimulus from self-pMHCs, the birth rate of naive cells, $\lambda_N^{(i)}(\mathbf{n})$, is the same as in Chapter 4, and therefore it is given by Eq. (4.20). Similarly, the death rate of naive cells, $\mu_N^{(i)}(\mathbf{n})$, also remains unchanged, and it is given by Eq. (4.3).

The death rates of effector and memory cells are assumed to be linear, similarly to that of naive cells. That is, when the competition process is in state $\mathbf{n} = (n_1, \ldots, n_\eta, e_1, \ldots, e_\eta, m_1, \ldots, m_\eta)$, the death rates of effector and memory cells for clonotype i are given, respectively, by

$$\mu_E^{(i)}(\mathbf{n}) = \mu_{Ei}\, e_i, \tag{5.4}$$

$$\mu_M^{(i)}(\mathbf{n}) = \mu_{Mi}\, m_i, \tag{5.5}$$

where μ_{Ei}, and μ_{Mi} are the per-cell death rates for effector and memory T cells of clonotype i. Since homeostatic division of memory cells is mediated by cytokine signalling, which is independent of TCR specificity (Surh & Sprent, 2008; Tan *et al.*, 2002), the birth rate of memory T cells is also assumed to be linear. That is, for clonotype i the birth rate of memory cells is given by

$$\lambda_M^{(i)}(\mathbf{n}) = \lambda_{Mi}\, m_i. \tag{5.6}$$

The last remaining transition rate that does not depend on TCR-VDP interactions is the differentiation rate from effector to memory. As previously discussed, this rate is such that after the infection has been cleared only a fraction β_i of the remaining effector cells of clonotype i differentiate into the memory phenotype (Ahmed & Gray, 1996; Kaech & Cui, 2012). Furthermore, since proliferation of effector cells is dependent on VDP interactions, once the infection has been cleared the only possible events for effector cells are death, or differentiation into

memory. Therefore, the per-cell differentiation rate for clonotype i from effector to memory is ψ_{Ei} such that

$$\frac{\psi_{Ei}}{\psi_{Ei} + \mu_{Ei}} := \beta_i. \tag{5.7}$$

Then, the differentiation rate for clonotype i from effector to memory is given by

$$\psi_E^{(i)}(\mathbf{n}) = \psi_{Ei}\, e_i = \frac{\beta_i}{1 - \beta_i}\, \mu_{Ei}\, e_i. \tag{5.8}$$

Since differentiation of naive and memory cells into the effector phenotype, and proliferation of the effector population all depend on TCR-VDP interactions, it is useful to first define the per-cell rate provided by a VDP $v \in \mathscr{V}$ to all T cells that are able to recognise it. Note that the expression of the per-cell stimulus rate is analogous for both infections, $\mathscr{V}$ and $\mathscr{W}$, thus, only one needs to be considered to find said expression. Since all phenotypes are able to receive stimuli from VDPs, the per-cell stimulus rate from VDP v can be written as

$$\delta_v(\mathbf{n}) = \frac{\gamma_v}{n_v + e_v + m_v}, \tag{5.9}$$

where γ_v is the constant rate of stimulus for VDP v, n_v is the total number of naive cells that can recognise v, e_v the total number of effector cells that can recognise v, and m_v is the total number of memory cells that can recognise v. That is, defining $\mathscr{C}^v$ to be the set of clonotypes that can recognise v, then

$$n_v = \sum_{i \in \mathscr{C}^v} n_i, \tag{5.10}$$

$$e_v = \sum_{i \in \mathscr{C}^v} e_i, \tag{5.11}$$

$$m_v = \sum_{i \in \mathscr{C}^v} m_i. \tag{5.12}$$

Given that every clonotype will interact differently with each VDP, a dimensionless parameter, $\kappa_v(i) \in [0, 1]$, is introduced to represent the avidity of clonotype i to VDP v. That is, the strength of the response generated when v is recognised by i (La Gruta *et al.*, 2004; van den Boorn *et al.*, 2006). Then, the total per-cell stimulus rate from all available VDPs in $\mathscr{V}$ to clonotype i is given by

$$\delta_{\mathscr{V}}^{(i)}(\mathbf{n}) = \sum_{v \in \mathscr{V}_i} \kappa_v(i) \delta_v(\mathbf{n}). \tag{5.13}$$

Using this expression for the per-cell stimulus available, the differentiation rates from naive and memory to effector are given by

$$\alpha_N^{(i)}(\mathbf{n}) = \alpha_N \; n_i \; \delta_\gamma^{(i)}(\mathbf{n}), \tag{5.14}$$

$$\alpha_M^{(i)}(\mathbf{n}) = \alpha_M \; m_i \; \delta_\gamma^{(i)}(\mathbf{n}), \tag{5.15}$$

where α_N and α_M are dimensionless constants, which encode the differences between signalling pathways for naive and memory T cells (Adachi & Davis, 2011; Farber, 2009). Similarly, the birth rate of effector cells is given by

$$\lambda_E^{(i)}(\mathbf{n}) = \lambda_E \; e_i \; \delta_\gamma^{(i)}(\mathbf{n}), \tag{5.16}$$

where λ_E is a dimensionless constant encoding the difference of the effector T cell signalling pathway (Gerriets & Rathmell, 2012).

Consider two states in the state space $\mathcal{S}$, $\mathbf{n}$ as defined before, and $\tilde{\mathbf{n}} = (\tilde{n}_1, \ldots, \tilde{n}_\eta, \tilde{e}_1, \ldots, \tilde{e}_\eta, \tilde{m}_1, \ldots, \tilde{m}_\eta)$. Then the infinitesimal transition probability from state $\mathbf{n}$ to $\tilde{\mathbf{n}}$ is defined as

$$p_{\mathbf{n}\tilde{\mathbf{n}}}(\Delta t) = \mathbb{P}\left\{ \begin{array}{l|l} X_i(t+\Delta t) = \tilde{n}_i, & X_i(t) = n_i, \\ Y_i(t+\Delta t) = \tilde{e}_i, & Y_i(t) = e_i, \text{ for } i = 1, 2, \ldots, \eta \\ Z_i(t+\Delta t) = \tilde{m}_i & Z_i(t) = m_i \end{array} \right\}, \tag{5.17}$$

which, as $\Delta t \to 0$, satisfies

$$p_{\mathbf{n}\tilde{\mathbf{n}}}(\Delta t) = \begin{cases} \lambda_N^{(i)}(\mathbf{n})\Delta t + o(\Delta t) & \tilde{n}_i = n_i + 1 \\ \lambda_E^{(i)}(\mathbf{n})\Delta t + o(\Delta t) & \tilde{e}_i = e_i + 1 \\ \lambda_M^{(i)}(\mathbf{n})\Delta t + o(\Delta t) & \tilde{m}_i = m_i + 1 \\ \mu_N^{(i)}(\mathbf{n})\Delta t + o(\Delta t) & \tilde{n}_i = n_i - 1 \\ \mu_E^{(i)}(\mathbf{n})\Delta t + o(\Delta t) & \tilde{e}_i = e_i - 1 \\ \mu_M^{(i)}(\mathbf{n})\Delta t + o(\Delta t) & \tilde{m}_i = m_i - 1 \\ \alpha_N^{(i)}(\mathbf{n})\Delta t + o(\Delta t) & \tilde{n}_i = n_i - 1, \tilde{e}_i = e_i + 1 \\ \alpha_M^{(i)}(\mathbf{n})\Delta t + o(\Delta t) & \tilde{m}_i = m_i - 1, \tilde{e}_i = e_i + 1 \\ \psi_E^{(i)}(\mathbf{n})\Delta t + o(\Delta t) & \tilde{e}_i = e_i - 1, \tilde{m}_i = m_i + 1 \\ \begin{array}{l} 1 - \sum_{i=1}^{\eta}(\lambda_N^{(i)}(\mathbf{n}) + \lambda_E^{(i)}(\mathbf{n}) + \lambda_M^{(i)}(\mathbf{n}))\Delta t \\ - \sum_{i=1}^{\eta}(\mu_N^{(i)}(\mathbf{n}) + \mu_E^{(i)}(\mathbf{n}) + \mu_M^{(i)}(\mathbf{n}))\Delta t \\ - \sum_{i=1}^{\eta}(\alpha_N^{(i)}(\mathbf{n}) + \alpha_M^{(i)}(\mathbf{n}) + \psi_E^{(i)}(\mathbf{n}))\Delta t \end{array} & \mathbf{n} = \tilde{\mathbf{n}} \\ o(\Delta t) & \text{otherwise} \end{cases}$$

5.1.4 Certainty of extinction

Using the method described in Iglehart (1964), similarly to Section 4.3.1, it can be shown that the competition process will reach the state $(0,\ldots,0)$ with probability 1. For this, first the state space $\mathcal{S}$ must be partitioned into levels, $L(k)$, defined as follows

$$L(k) = \left\{ (n_1,\ldots,n_\eta,e_1,\ldots,e_\eta,m_1,\ldots,m_\eta) \in \mathcal{S} : \sum_{i=1}^{\eta} n_i + e_i + m_i = k \right\},$$

for $k = 0,1,\ldots$, so that $\mathcal{S} = \bigcup_{k=0}^{+\infty} L(k)$. Now, by defining the following birth and death rates

$$\begin{aligned} \lambda'_k &= \max_{\mathbf{n}\in L(k)} \left\{ \sum_{i=1}^{\eta} \lambda_N^{(i)}(\mathbf{n}) + \lambda_E^{(i)}(\mathbf{n}) + \lambda_M^{(i)}(\mathbf{n}) \right\}, \\ \mu'_k &= \min_{\mathbf{n}\in L(k)} \left\{ \sum_{i=1}^{\eta} \mu_N^{(i)}(\mathbf{n}) + \mu_E^{(i)}(\mathbf{n}) + \mu_M^{(i)}(\mathbf{n}) \right\}, \end{aligned} \tag{5.18}$$

a uni-variate birth and death process can be defined on the state space $\mathcal{S}' = \{L(k) : k = 0,1,2,\ldots\}$. Where each level is considered to be a macro-state, and transitions occur between adjacent states. Given that the birth rate of this process is the maximum of the total birth rates, and the death rate is the minimum of the total death rates, it is easy to see that the uni-variate process moves towards the absorbing state $(0,\ldots,0)$ at a slower rate than the original competition process. Therefore, if this process reaches the absorbing state in finite time, so will the original competition process.

According to Reference (Iglehart, 1964, Theorem 3), a sufficient condition for absorption at $L(0)$ to be certain is that the sum

$$\sum_{k=1}^{+\infty} \frac{1}{\lambda'_k \sigma_k}, \tag{5.19}$$

where $\sigma_k = \frac{\lambda'_1\lambda'_2\ldots\lambda'_{k-1}}{\mu'_2\mu'_3\ldots\mu'_k}$ for $k \geq 2$ and $\sigma_1 = 1$, is divergent. Note that the division rate of effector cells, $\lambda_E^{(i)}(\mathbf{n})$, is bounded by

$$\begin{aligned} \lambda_E^{(i)}(\mathbf{n}) &= \lambda_E \sum_{v\in\mathscr{V}_i} \frac{e_i\ \kappa_v(i)}{n_v + e_v + m_v}\gamma_v \\ &\leq \lambda_E \sum_{v\in\mathscr{V}_i} \gamma_v \end{aligned}$$

$$\leq \lambda_E \, |\mathscr{V}_i| \max_{v \in \mathscr{V}_i} \{\gamma_v\}$$
$$\leq \lambda_E \, |\mathscr{V}| \max_{v \in \mathscr{V}} \{\gamma_v\} .$$

From this result, and the bound for $\lambda_N^i(\mathbf{n})$ found in Eq. (4.61), it can be seen that

$$\lambda'_k = \max_{\mathbf{n} \in L(k)} \left\{ \sum_{i=1}^{\eta} \lambda_N^{(i)}(\mathbf{n}) + \lambda_E^{(i)}(\mathbf{n}) + \lambda_M^{(i)}(\mathbf{n}) \right\} \leq \sum_{i=1}^{\eta} \varphi_i + \lambda_E \, |\mathscr{V}| \max_{v \in \mathscr{V}} \{\gamma_v\} + k\lambda_M^*,$$

where $\lambda_M^* = \max_{i=1,2,\ldots,\eta} \{\lambda_{Mi}\}$. It is also easy to see that

$$\mu'_k = \min_{\mathbf{n} \in L(k)} \left\{ \sum_{i=1}^{\eta} \mu_N^{(i)}(\mathbf{n}) + \mu_E^{(i)}(\mathbf{n}) + \mu_M^{(i)}(\mathbf{n}) \right\} = k\mu^*,$$

where $\mu^* = \min_{i=1,2,\ldots,\eta} \{\mu_{Ni}, \mu_{Ei}, \mu_{Mi}\}$. Then, replacing these values in Eq. (5.19) the following inequality is obtained

$$\sum_{k=1}^{+\infty} \frac{1}{\lambda'_k \sigma_k} \geq \sum_{k=1}^{+\infty} \frac{k!(\mu^*)^{k-1}}{\left(\sum_{i=1}^{\eta} \varphi_i + \lambda_E \, |\mathscr{V}| \max_{v \in \mathscr{V}} \{\gamma_v\} + k\lambda_M^* \right)^k} = \sum_{k=1}^{+\infty} a_k. \tag{5.20}$$

Then, the divergence of Eq. (5.19) can be checked by considering the limit as $k \to +\infty$ of following ratio

$$\frac{a_{k+1}}{a_k} = \frac{(k+1)\mu^*}{\sum_{i=1}^{\eta} \varphi_i + \lambda_E \, |\mathscr{V}| \max_{v \in \mathscr{V}} \{\gamma_v\} + k\lambda_M^*}. \tag{5.21}$$

In the case where there are no cells of memory phenotype; that is, $m_i = 0$ for all i, and the term $k\lambda_M^*$ can be considered to be zero, this limit tends to $+\infty$, and therefore $\mathcal{X}$ reaches the absorbing state $(0, \ldots, 0)$ with probability 1. This means that prior to the generation of memory, the extinction of naive and effector phenotype cells is certain for sufficiently late times. In the case where there are memory phenotype cells present, however, the limit tends to $\frac{\mu^*}{\lambda_M^*}$. This means that once a population of memory cells is generated, according Theorem 3 of (Iglehart, 1964), extinction of every phenotype for all clonotypes is only certain if the smallest per-cell death rate is greater that the largest per-cell birth rate of memory cells.

5.2 Linear noise approximation of the competition process

Given the fact that during infection the populations of effector cells grow by several orders of magnitude (Busch & Pamer, 1999), analysis of the stochastic model becomes intractable due to the large state space that must be considered. Instead, van Kampen's large N expansion (van Kampen, 2007) is used in this section to find a linear noise approximation of the stochastic process which can then be used to study the dynamics of the populations. First, the following step operators are defined

$$\mathbb{S}_N^{\pm i} f(\mathbf{n}) = f(n_1, \ldots, n_{i-1}, n_i \pm 1, n_{i+1}, \ldots, n_\eta, e_1, \ldots, e_\eta, m_1, \ldots, m_\eta), \tag{5.22a}$$

$$\mathbb{S}_E^{\pm i} f(\mathbf{n}) = f(n_1, \ldots, n_\eta, e_1, \ldots, e_{i-1}, e_i \pm 1, e_{i+1}, \ldots, e_\eta, m_1, \ldots, m_\eta), \tag{5.22b}$$

$$\mathbb{S}_M^{\pm i} f(\mathbf{n}) = f(n_1, \ldots, n_\eta, e_1, \ldots, e_\eta, m_1, \ldots, m_{i-1}, m_i \pm 1, m_{i+1}, \ldots, m_\eta). \tag{5.22c}$$

Using these step operators the Kolmogorov equation of the competition process can be written as

$$\begin{aligned}
\frac{\mathrm{d}p_{\mathbf{n}}(t)}{\mathrm{d}t} = \sum_{i=1}^{\eta} & \left(\mathbb{S}_N^{-i} - 1\right)\left[\lambda_N^{(i)}(\mathbf{n})p_{\mathbf{n}}(t)\right] + \left(\mathbb{S}_N^{+i} - 1\right)\left[\mu_N^{(i)}(\mathbf{n})p_{\mathbf{n}}(t)\right] \\
& + \left(\mathbb{S}_E^{-i} - 1\right)\left[\lambda_E^{(i)}(\mathbf{n})p_{\mathbf{n}}(t)\right] + \left(\mathbb{S}_E^{+i} - 1\right)\left[\mu_E^{(i)}(\mathbf{n})p_{\mathbf{n}}(t)\right] \\
& + \left(\mathbb{S}_M^{-i} - 1\right)\left[\lambda_M^{(i)}(\mathbf{n})p_{\mathbf{n}}(t)\right] + \left(\mathbb{S}_M^{+i} - 1\right)\left[\mu_M^{(i)}(\mathbf{n})p_{\mathbf{n}}(t)\right] \\
& + \left(\mathbb{S}_N^{+i}\mathbb{S}_E^{-i} - 1\right)\left[\alpha_N^{(i)}(\mathbf{n})p_{\mathbf{n}}(t)\right] + \left(\mathbb{S}_E^{-i} - 1\right)\left[\alpha_M^{(i)}(\mathbf{n})p_{\mathbf{n}}(t)\right] \\
& + \left(\mathbb{S}_E^{+i}\mathbb{S}_M^{-i} - 1\right)\left[\psi_E^{(i)}(\mathbf{n})p_{\mathbf{n}}(t)\right].
\end{aligned} \tag{5.23}$$

Now, the variables $x_i(t)$, $y_i(t)$, $z_i(t)$, $\xi_i(t)$, $\theta_i(t)$, and $\zeta_i(t)$ are defined as follows

$$X_i(t) = \Omega x_i(t) + \Omega^{1/2}\xi_i(t), \tag{5.24a}$$

$$Y_i(t) = \Omega y_i(t) + \Omega^{1/2}\theta_i(t), \tag{5.24b}$$

$$Z_i(t) = \Omega z_i(t) + \Omega^{1/2}\zeta_i(t), \tag{5.24c}$$

where the parameter Ω represents the size of the system, and the fluctuations $\xi_i(t)$, $\theta_i(t)$, and $\zeta_i(t)$ are of order $\Omega^{1/2}$. By considering these variables the probability density $p_{\mathbf{n}}(t)$ becomes instead the density $\Pi(\xi, \theta, \zeta; t)$; that is, $p_{\mathbf{n}}(t) = \Pi(\xi, \theta, \zeta; t)$,

with $\xi = (\xi_1, \ldots, \xi_\eta)$, $\theta = (\theta_1, \ldots, \theta_\eta)$, and $\zeta = (\zeta_1, \ldots, \zeta_\eta)$. Then, the Kolmogorov equation of this probability density is

$$\begin{aligned}
&\frac{\partial \Pi}{\partial t} - \Omega^{1/2} \sum_{i=1}^{\eta} \frac{\mathrm{d}x_i}{\mathrm{d}t}\frac{\partial \Pi}{\partial \xi_i} + \frac{\mathrm{d}y_i}{\mathrm{d}t}\frac{\partial \Pi}{\partial \theta_i} + \frac{\mathrm{d}z_i}{\mathrm{d}t}\frac{\partial \Pi}{\partial \zeta_i} \\
&\quad = \sum_{i=1}^{\eta} \lambda_N^{(i)}(\mathbf{n}) \left[-\Omega^{-1/2}\frac{\partial}{\partial \xi_i} + \frac{1}{2}\Omega^{-1}\frac{\partial^2}{\partial \xi_i^2} \right] \Pi + \mu_N^{(i)}(\mathbf{n}) \left[\Omega^{-1/2}\frac{\partial}{\partial \xi_i} + \frac{1}{2}\Omega^{-1}\frac{\partial^2}{\partial \xi_i^2} \right] \Pi \\
&\qquad + \lambda_E^{(i)}(\mathbf{n}) \left[-\Omega^{-1/2}\frac{\partial}{\partial \theta_i} + \frac{1}{2}\Omega^{-1}\frac{\partial^2}{\partial \theta_i^2} \right] \Pi + \mu_E^{(i)}(\mathbf{n}) \left[\Omega^{-1/2}\frac{\partial}{\partial \theta_i} + \frac{1}{2}\Omega^{-1}\frac{\partial^2}{\partial \theta_i^2} \right] \Pi \\
&\qquad + \lambda_M^{(i)}(\mathbf{n}) \left[-\Omega^{-1/2}\frac{\partial}{\partial \zeta_i} + \frac{1}{2}\Omega^{-1}\frac{\partial^2}{\partial \zeta_i^2} \right] \Pi + \mu_M^{(i)}(\mathbf{n}) \left[\Omega^{-1/2}\frac{\partial}{\partial \zeta_i} + \frac{1}{2}\Omega^{-1}\frac{\partial^2}{\partial \zeta_i^2} \right] \Pi \\
&\qquad + \alpha_N^{(i)}(\mathbf{n}) \left[\Omega^{-1/2}\left(\frac{\partial}{\partial \xi_i} - \frac{\partial}{\partial \theta_i} \right) + \frac{1}{2}\Omega^{-1}\left(\frac{\partial}{\partial \xi_i} - \frac{\partial}{\partial \theta_i} \right)^2 \right] \Pi \\
&\qquad + \alpha_M^{(i)}(\mathbf{n}) \left[-\Omega^{-1/2}\frac{\partial}{\partial \theta_i} + \frac{1}{2}\Omega^{-1}\frac{\partial^2}{\partial \theta_i^2} \right] \Pi \\
&\qquad + \psi_E^{(i)}(\mathbf{n}) \left[\Omega^{-1/2}\left(\frac{\partial}{\partial \theta_i} - \frac{\partial}{\partial \zeta_i} \right) + \frac{1}{2}\Omega^{-1}\left(\frac{\partial}{\partial \theta_i} - \frac{\partial}{\partial \zeta_i} \right)^2 \right] \Pi.
\end{aligned} \tag{5.25}$$

To collect the terms of order $\Omega^{1/2}$ from Eq. (5.25) to obtain the deterministic approximation, it is necessary to first find explicit expressions for the transition rates using the variables introduced in Eq. (5.24). Since the dynamics of homeostatic proliferation of naive cells remains unchanged from that presented in Chapter 4, the results of the previous van Kampen expansion can be used again (see Eq. (4.49a)). Substituting the variables proposed in Eq. (5.24) in Equations (4.3), (5.4), and (5.5) results in

$$\begin{aligned}
\mu_N^{(i)}(\mathbf{n}) &= \mu_N\left(\Omega x_i + \Omega^{1/2}\xi_i\right), \\
\mu_E^{(i)}(\mathbf{n}) &= \mu_E\left(\Omega y_i + \Omega^{1/2}\theta_i\right), \\
\mu_M^{(i)}(\mathbf{n}) &= \mu_M\left(\Omega z_i + \Omega^{1/2}\zeta_i\right).
\end{aligned}$$

Similarly, substituting for Eq. (5.24) in Equations (5.6), and (5.8) results in the following expressions

$$\begin{aligned}
\lambda_M^{(i)}(\mathbf{n}) &= \lambda_M\left(\Omega z_i + \Omega^{1/2}\zeta_i\right), \\
\psi_E^{(i)}(\mathbf{n}) &= \frac{\beta}{1-\beta}\mu_E\left(\Omega y_i + \Omega^{1/2}\theta_i\right).
\end{aligned}$$

Finally, since $\alpha_N^{(i)}(\mathbf{n})$, $\alpha_M^{(i)}(\mathbf{n})$, and $\lambda_E^{(i)}(\mathbf{n})$ all depend on $\delta_{\mathscr{V}}^{(i)}(\mathbf{n})$, they can be written as

$$\begin{aligned}\alpha_N^{(i)}(\mathbf{n}) &= \alpha_N \left(\Omega x_i + \Omega^{1/2}\xi_i\right) \delta_{\mathscr{V}}^{(i)}(\mathbf{n}),\\ \lambda_E^{(i)}(\mathbf{n}) &= \lambda_E \left(\Omega y_i + \Omega^{1/2}\theta_i\right) \delta_{\mathscr{V}}^{(i)}(\mathbf{n}),\\ \alpha_M^{(i)}(\mathbf{n}) &= \alpha_M \left(\Omega z_i + \Omega^{1/2}\zeta_i\right) \delta_{\mathscr{V}}^{(i)}(\mathbf{n}).\end{aligned}$$

Next, define $x_v = \sum_{i\in\mathscr{C}^v} x_i$, $y_v = \sum_{i\in\mathscr{C}^v} y_i$, and $z_v = \sum_{i\in\mathscr{C}^v} z_i$. Then, $\delta_{\mathscr{V}}^{(i)}(\mathbf{n})$ is given by

$$\delta_{\mathscr{V}}^{(i)}(\mathbf{n}) = \sum_{v\in\mathscr{V}_i} \frac{\tilde{\gamma}_v}{x_v + y_v + z_v}\left(1 - \frac{\xi_v + \theta_v + \zeta_v}{x_v + y_v + z_v}\Omega^{-1/2}\right),$$

where $\gamma_v = \Omega\tilde{\gamma}_v$. Finally, replacing these results in Eq. (5.25), and collecting all the terms of order $\Omega^{1/2}$ results in the following deterministic approximation of the stochastic competition process

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = x_i\left(\tilde{\varphi}_i \sum_{j=0}^{\eta}\sum_{k=1}^{\binom{\eta-1}{j}} \frac{p_{ij}^k}{x_{ij}^k} - \mu_N - \alpha_N \sum_{v\in\mathscr{V}_i} \frac{\tilde{\gamma}_v}{x_v + y_v + z_v}\right), \tag{5.26}$$

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = \left(\lambda_E\, y_i + \alpha_N\, x_i + \alpha_M\, z_i\right) \sum_{v\in\mathscr{V}_i} \frac{\tilde{\gamma}_v}{x_v + y_v + z_v} - \left(1 - \frac{\beta}{1-\beta}\right)\mu_E\, y_i, \tag{5.27}$$

$$\frac{\mathrm{d}z_i}{\mathrm{d}t} = z_i\left(\lambda_M - \mu_M\right) + \frac{\beta}{1-\beta}\mu_E\, y_i. \tag{5.28}$$

In the following section, novel experimental data on cross-reactivity to influenza A viruses will be presented, and used to try to parametrise this deterministic approximation of the model.

5.3 Heterologous influenza A virus murine infection models

The data presented and analysed in this section is the result of murine experiments of influenza A virus (IAV) infection performed by Jessica Gaevert at the Paul Thomas laboratory in the Immunology department of St. Jude Children's Research Hospital. These experiments focus on the quantification of T cell cross-reactivity to variants of IAV in the context of heterologous infection; that is, two infections with different variants of the virus, in this case occurring sequentially.

5.3.1 Description of the experiment

The experiments used a prime-challenge model, in which mice are first infected with a variant of IAV (called a priming or primary infection), and once they have cleared the infection they are infected again with any of the variants (called a challenge infection). In the experiments three different variants of IAV are considered, with the only difference between them being a single amino acid change in the immunodominant NP_{366} epitope (see Table 5.1). This is done to mimic genetic drift of influenza A viruses. H3N2 and H1N1 background IAVs are used on the primary and challenge infections respectively, to be able to examine the cross-reactive T cell responses during the challenge infection without recalling a B cell response. This experimental design allows for the identification of cross-reactive $CD8^+$ T cells using tetramers with each NP_{366} epitope and multi-parameter flow-cytometry. Given the three IAV variants there are 9 possible permutations (with replacement) for the possible prime-challenge combinations, all of which are considered in these experiments. The variant with an unchanged NP_{366} epitope is called wild type (WT), and the two other derive their names from their difference to the WT peptide. That is, the T8A variant replaces the T amino acid in the 8^{th} position with an A amino acid, and the N3A variant replaces the N in the 3^{rd} position with an A.

Influenza A virus variants	
Name	NP_{366} epitope
WT	ASNENMETM
T8A	ASNENMEAM
N3A	ASAENMETM

Table 5.1: Influenza A virus variants, and their respective NP_{366} epitope, used in the heterologous infection experiment. Red denotes the single amino acid change when compared to the wild type (WT) variant, and the names of the variants are derived from their difference with the WT peptide. The T8A variant replaces the T amino acid in the 8^{th} position with an A, and the N3A variant replaces the N amino acid in the 3^{rd} position with an A.

For all the possible prime-challenge combinations the experiment was performed as follows: on day 0, mice were intranasally infected with an H3N2 background primary variant. On day 10, at the peak of the primary infection, lungs and spleen were harvested from mice. On day 70, when the infection has been cleared and immune memory has been generated, lungs and spleen were harvested. On day 80, mice are infected with an H1N1 background challenge variant. Finally, on day 90 lung and spleen are harvested from mice at the peak of the challenge infection. This timeline is summarised in Figure 5.4. Since the data is only sampled during harvesting timepoints, these names (primary, memory, and challenge) will be used to refer to the timepoints in the data.

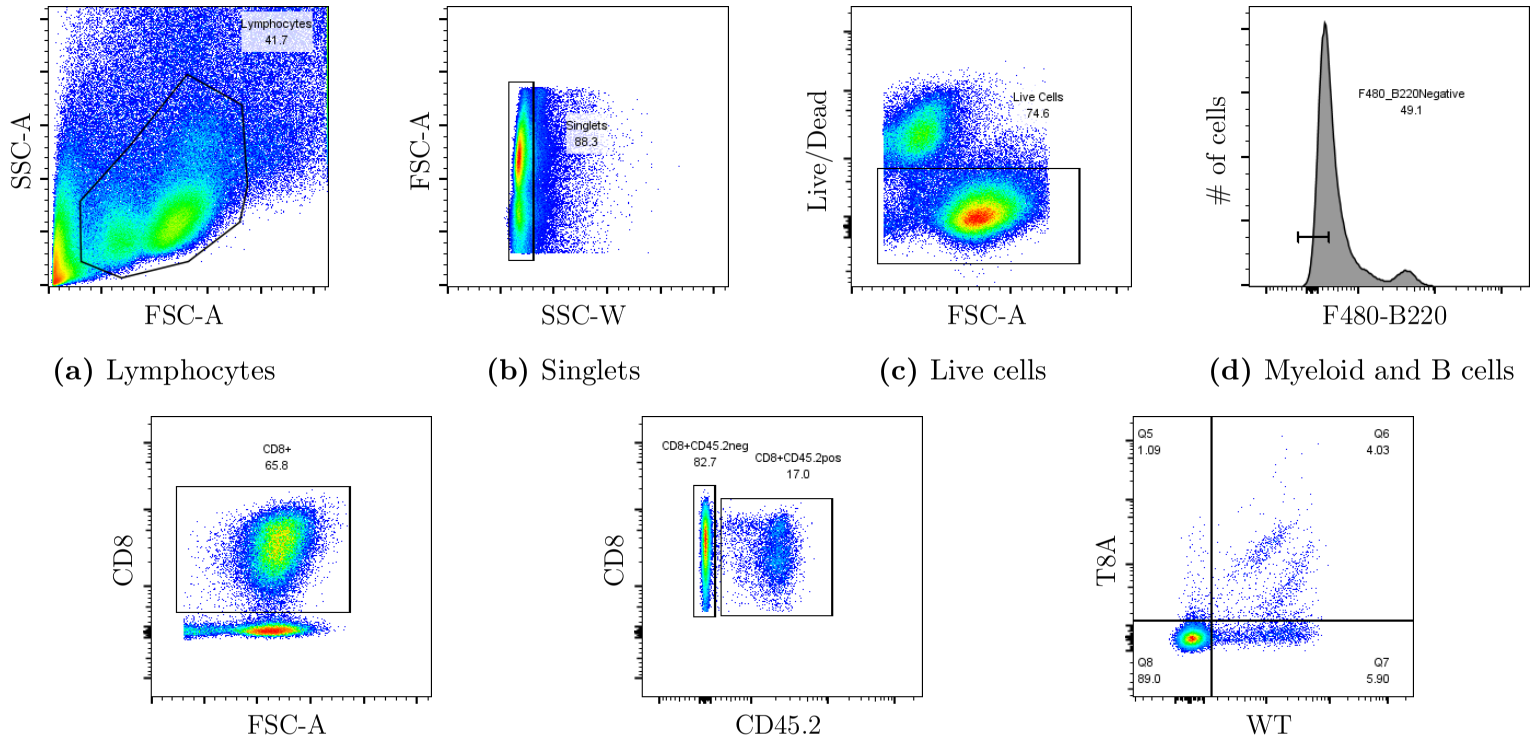
Figure 5.4: Timeline of infection and harvesting for the prime-challenge experiment of IAV in mice. Red arrows represent infection timepoints, and blue arrows represent harvest timepoints. The primary infection makes used of an H3N2 background virus, and the challenge infection uses an H1N1 background virus to prevent a recall response of B cells.

For all timepoints, the analysis of the samples using flow-cytometry was conducted in the same way. First, the samples are prepared by staining with tetramers for live/dead, myeloid and B cells, $CD8^+$ cells, circulating cells, and the three NP_{366} epitopes (WT, T8A, N3A). Then, using a flow-cytometer the stained samples were analysed and gated as follows: first, lymphocytes are gated by forward and side scatter area (Figure 5.5a), and singlets (single cells that are not attached to another cell) are gated by forward scatter width and side scatter area (Figure 5.5b). Next, living cells are gated using a live/dead tetramer (Figure 5.5c), and myeloid and B cells are excluded using the myeloid and B cell

tetramer (Figure 5.5d). Finally, $CD8^+$ cells are gated from the remaining cells (Figure 5.5e), which are then separated into circulating and resident using the CD45 marker (Figure 5.5f). After this, the WT, T8A, and N3A tetramers are used to identify cells that can recognise the NP_{366} epitope (Figure 5.5g). These epitope positive populations are then separated into cells that are positive for only one of the tetramers (single positive cells), cells that can recognise two of the tetramers (double positive cells), and cells that can recognise all tetramers (triple positive cells). Cells that are unable to recognise any of the tetramers are called triple negative cells.

It is important to note that when doing flow-cytometry experiments the complete sample is not run through the flow-cytometer. A sub-sample of only 10^5 cells is taken and analysed with the flow-cytometer. From this sub-sample the frequency of each population is calculated, which is then multiplied by the total number of cells in the sample to approximate the number of cells in each of the gates. Given that the site of infection for IAV is the lungs, lung resident cells will be the main focus of the analysis of the data, since it is this population that is most prevalent at the site of infection (Gebhardt *et al.*, 2009; Jiang *et al.*, 2012; Masopust & Soerens, 2019; Topham *et al.*, 1997).

Figure 5.6 is a summary of the frequencies of lung resident tetramer-positive cells. Each circle of the Venn diagrams represents positivity to a given tetramer, and its area is proportional to the frequency of that population. It is important to note that the size of each circle only represents its frequency in comparison to the other frequencies of the timepoint, since the areas are normalised to the largest population in the timepoint. Another population of importance during infection are spleen circulating cells, since these are cells that were present in the lymph nodes and are migrating to the site of infection (Ely *et al.*, 2006; Kohlmeier *et al.*, 2007). Figure 5.7 is a summary of the frequencies for spleen circulating epitope positive cells.



(a) Lymphocytes **(b)** Singlets **(c)** Live cells **(d)** Myeloid and B cells

(e) $CD8^+$ cells **(f)** Circulating cells **(g)** Epitope-specific cells

Figure 5.5: Gating strategy used to identify $CD8^+$ cells from the samples taken during the harvesting timepoints. Using flow-cytometry, first lymphocytes are identified (a), followed by singlets; that is, cells that did not become attached to another cell during the flow-cytometry experiment (b). Then, the data is gated for living cells (c), and all myeloid and B cells are excluded (d). Following this, the sample is gated for $CD8^+$ cells (e), which are then gated into circulating and resident compartments (f). The resulting populations of circulating and resident cells are then finally gated according to their positivity to the NP_{366} epitope (g).

Figure 5.6: Summary of the flow-cytometry frequencies of tetramer-positive cells for lung resident CD8$^+$ cells. Each circle of the Venn diagrams represents positivity to a given variant's tetramer, with its area proportional to the mean frequency of cells with that recognition profile. Note that diagrams are not to scale with one another, since the areas are normalised to the total frequency of the largest populations for each timepoint. For the numerical values see Appendix B.

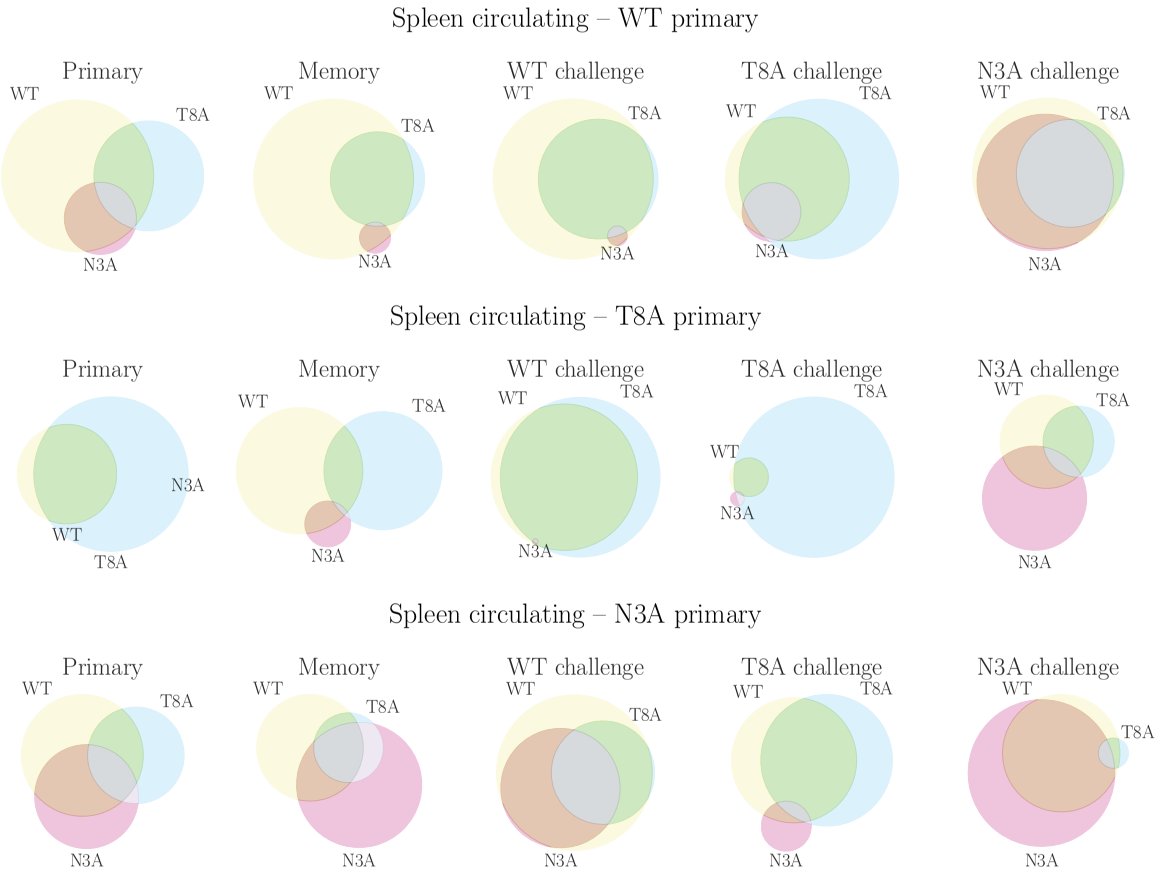


Figure 5.7: Summary of the flow-cytometry frequencies of tetramer-positive cells for spleen circulating $CD8^+$ cells. Each circle of the Venn diagrams represents positivity to a given variant's tetramer, with its area proportional to the mean frequency of cells with that recognition profile. Note that diagrams are not to scale with one another, since the areas are normalised to the total frequency of the largest populations for each timepoint. For the numerical values see Appendix B.

Note that given the nature of the data, it is not possible to identify individual clonotypes, but instead clonotype families (defined by their tetramer-positivity). Furthermore, since only positivity for three specific tetramers is measured, activation from presentation of other immunogenic epitopes (Staneková & Varečková, 2010), and generalised activation of non-specific cells (Bangs *et al.*, 2006; Chapman *et al.*, 2005) are not detected. Thus, a bipartite recognition network that represents the epitope recognition profiles of the tetramer-positive cells presented in the data must not only consider the WT, T8A, and N3A epitopes, but also an extra *pseudo*-epitope representing the mean stimulus available to all clonotypes from the presentation of other epitopes, and generalised activation. This bipartite recognition network is shown in Figure 5.8

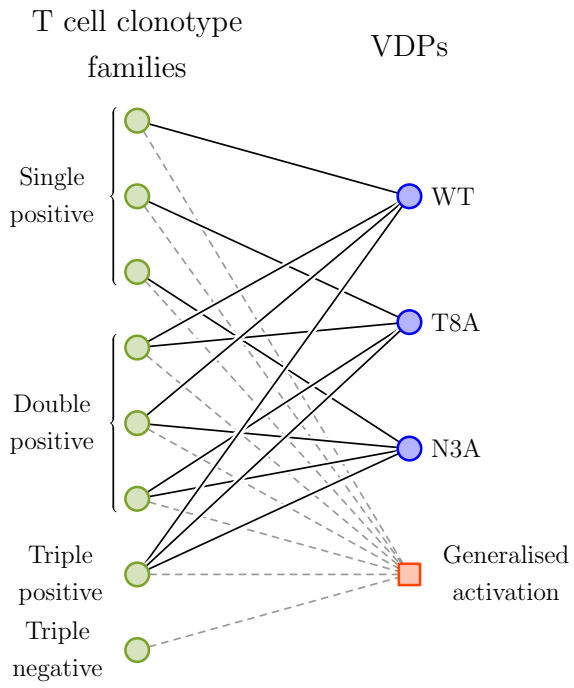


Figure 5.8: Bipartite recognition network for the experimental data. Clonotype families are considered instead of clonotypes, and they are defined in terms of their tetramer-positivity. In addition to the three epitopes that are experimentally measured, a pseudo-epitope representing the mean stimulus available to all clonotypes from the presentation of other epitopes, and generalised activation is considered. This is represented by a square node, and grey dashed lines.

5.3.2 Structural identifiability of the deterministic model

In order to determine the viability of parametrising the model presented in Section 5.2 with the data in Section 5.3.1, structural identifiability analysis was performed (Hong *et al.*, 2020). In particular, the SIAN structural identifiability toolbox (Hong *et al.*, 2019) was used to determine the identifiability of the parameters in the deterministic model given the data. For this analysis two different versions of the data were considered: first, the original data as described in Section 5.3.1; that is, populations of T cells which are either single, double, or triple positive. Second, the same data with the addition of a distinction between the naive, effector, and memory phenotypes.

Focusing first on the original data, the results of identifiability analysis using the SIAN structural identifiability toolbox (Hong *et al.*, 2019) showed that the only parameters that would be locally identifiable from the model with the data would be: β, the fraction of effector cells that differentiate to the memory phenotype after the infection is cleared, μ_E, the per-cell death rate of effector cells, and μ_N, the per-cell death rate of naive cells. Furthermore, using this data no parameters were globally identifiable. When considering the data with the added distinction of the phenotypes of cells, the same parameters that were locally identifiable with the original data were now globally identifiable. However, no other parameters were identifiable. These results are summarised in Table 5.2.

As discussed before, the original aim of this chapter was to use the experimental data provided by the Paul Thomas laboratory to parametrise the model presented in Section 5.2 and gain a better understanding of cross-reactivity, specifically by using the concept of avidity. However, the results of the structural identifiability analysis show that with the given data the parameter of greatest interest, $\kappa_v(i)$ the avidity of clonotype i for peptide v (see Eq. (5.13)), cannot be estimated. Thus, instead of continuing with this parametrisation, statistical analysis of the data is presented in Section 5.4 to better understand the behaviour of the cross-reactive immune responses observed in the experiments.

Identifiable parameters for the deterministic model		
Distinction between phenotypes	Globally identifiable parameters	Locally identifiable parameters
No	–	β, μ_N, μ_E
Yes	β, μ_N, μ_E	–

Table 5.2: Identifiable parameters of the model described in Section 5.2 with the data described in Section 5.3.1. The SIAN structural identifiability toolbox (Hong *et al.*, 2019) was used to find the identifiable parameters in the case where the phenotypes of naive, effector, and memory cells were separated in the data, as well as the case in which they were not. In both cases the same parameters are identifiable (β, μ_N, and μ_E), with the distinction that in the case with no separation of the phenotypes the parameters are locally identifiable, while they are globally identifiable in the other case.

5.4 Statistical analysis of the experimental data

Given the lack of identifiable parameters of interest, statistical analysis of the data was performed, and the results of this analysis are presented in this section (using the code presented in Appendix C). First, the populations of tetramer-specific T cells (single, double, and triple positive) were studied to understand how the populations of these cells are affected by the different prime-challenge infections. Following this, the populations of T cells positive for a given IAV tetramer (regardless of whether they were single, double, or triple positive) were studied to search for effects that cross-reactivity between the three IAV variants has on the total populations of tetramer-specific cells. Finally, by considering the mean populations at the primary and challenge timepoints the contraction of IAV specific T cells from primary to memory, and expansion from memory to challenge were characterised by the rates of change of the populations between timepoints, and the statistical properties of these slopes were studied.

5.4.1 Analysis of tetramer-specific T cells

To understand the relationships between populations of IAV specific T cells that are single positive, double positive, and triple positive, first the correlations between these populations at the primary, memory, and challenge timepoints were calculated between samples from mice with the same infection history. For this, the Spearman rank correlation coefficient (Kendall, 1948), and the Bonferroni adjustment to correct for multiple comparisons (Bland & Altman, 1995) were used. Due to some populations being below the level of detection for all samples, not all the correlation coefficients could be calculated for all timepoints. The results presented here were obtained using the code described in Appendix C.2.

First, the correlations on lung resident cells from WT-primed mice were considered (see the top panel in Figure 5.9). At the primary timepoint there are no significant correlations between the populations; that is, no correlations with a significance level better than 0.05 for the likelihood of the null hypothesis, suggesting that during this infection the immune response is diverse and not biased towards any epitope-specific cells in particular. Now, at the memory timepoint there is a strong positive correlation between T8A single positive cells and WT-N3A double positive cells. This suggests that following a WT primary infection, due to the expansion of the double positive population, the repertoire could be biased towards a cross-reactive response to a homologous or an N3A challenge, and a specific response to a T8A challenge. This change in the way the populations are correlated indicates that, for WT infections, contraction into memory is likely to be the mechanism by which the immune repertoire is being biased for future infections, since it is at the memory timepoint that significant correlations between the populations are first observed.

At the challenge timepoint a different behaviour is observed for each of the possible challenge infections. For the homologous WT challenge there is a correlation between triple negative cells and WT-T8A double positive cells, suggesting that when the immune response grows (marked by a decrease in the frequency of triple negative cells) it becomes more cross-reactive towards the T8A epitope. In the case of the T8A heterologous challenge, an increased frequency of WT-T8A positive cells results in an increased frequency of WT single positive cells, suggesting that

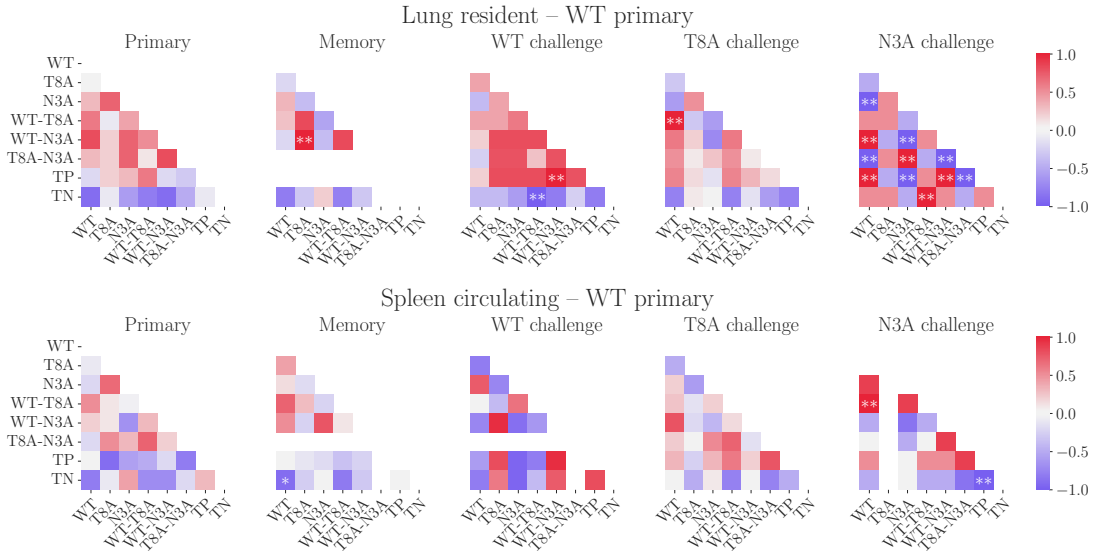


Figure 5.9: Spearman rank correlation coefficients between epitope positive and negative populations for lung resident and spleen circulating cells from mice primed with WT virus. Asterisks indicate a significant correlation after applying the Bonferroni adjustment ($* \rightarrow p < 0.05$ and $** \rightarrow p < 0.01$).

part of the WT-specific response is being recalled from the primary infection. The most significant correlations are found is in the N3A heterologous challenge, where several of the populations are significantly correlated to each other. In this case the correlations suggest that if the response is widely cross-reactive with increased triple positive cells, it will have a decreased ability to target N3A epitopes specifically, as evidenced by the negative correlations to N3A single positive and T8A-N3A double positive cells.

In comparison to the lung resident cells, for spleen circulating cells of the same mice there are fewer significant correlations in the data, see the bottom panel in Figure 5.9. Notably, during both the WT and T8A challenges there are no significant correlations between the cell populations. Meaning that during these infections the bias caused by the primary infection is present only at the site of infection and not in the spleen. In the case of the N3A challenge however, there is a negative correlation between triple negative and triple positive cells. This implies that as the response grows in size, as demonstrated by the decrease in the frequency

of triple negative cells, the frequency of triple positive cells in the spleen increases.

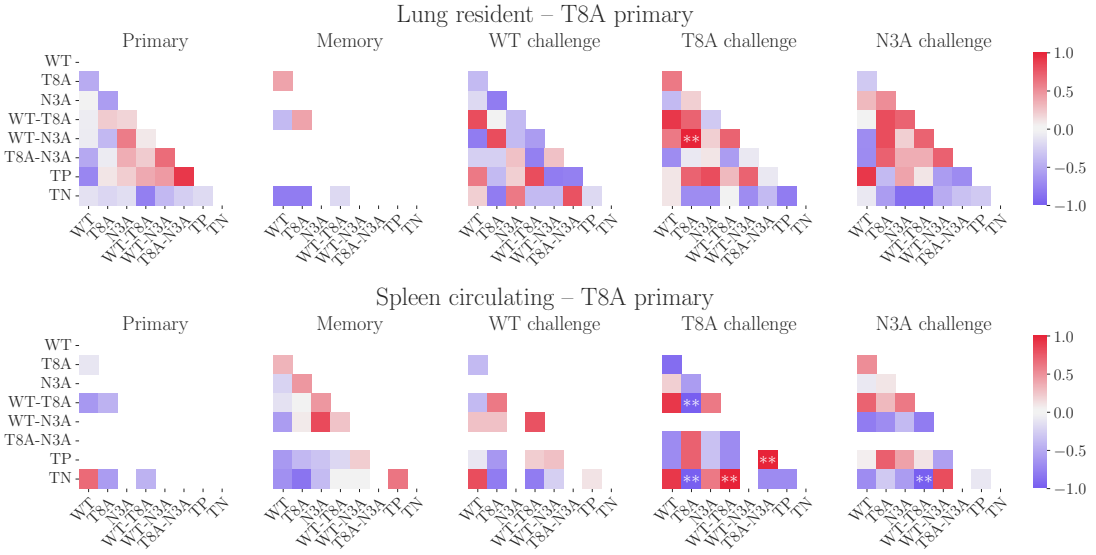


Figure 5.10: Spearman rank correlation coefficients between epitope positive and negative populations for lung resident and spleen circulating cells from mice primed with T8A virus. Asterisks indicate a significant correlation after applying the Bonferroni adjustment ($* \rightarrow p < 0.05$ and $** \rightarrow p < 0.01$).

Compared to WT-primed mice, data from T8A-primed mice showed very few significant correlations (see Figure 5.10). For lung resident cells, only the homologous challenge shows a correlation between T8A single positive cells and WT-N3A double positive cells. This suggests that responses to homologous T8A infection that are highly T8A-specific will have increased cross-reactivity with WT and N3A through the WT-N3A double positive population. Spleen circulating cells during a homologous challenge, on the other hand, show an increase in T8A single positive cells and a decrease in WT-T8A double positive cells as the magnitude of the response increases, showing a preference for epitope-specific cells outside the site of infection. Further confirming this, there is also a negative correlation between T8A and WT-T8A positive cells. Finally, in the N3A challenge there is an increase of WT-T8A double positive cells as the magnitude of the response increases, which is not seen in the lung resident cells, meaning that at the site of infection the total magnitude of the response has no effect on the cross-reactivity

with WT and T8A epitopes, while on the spleen the opposite is true.

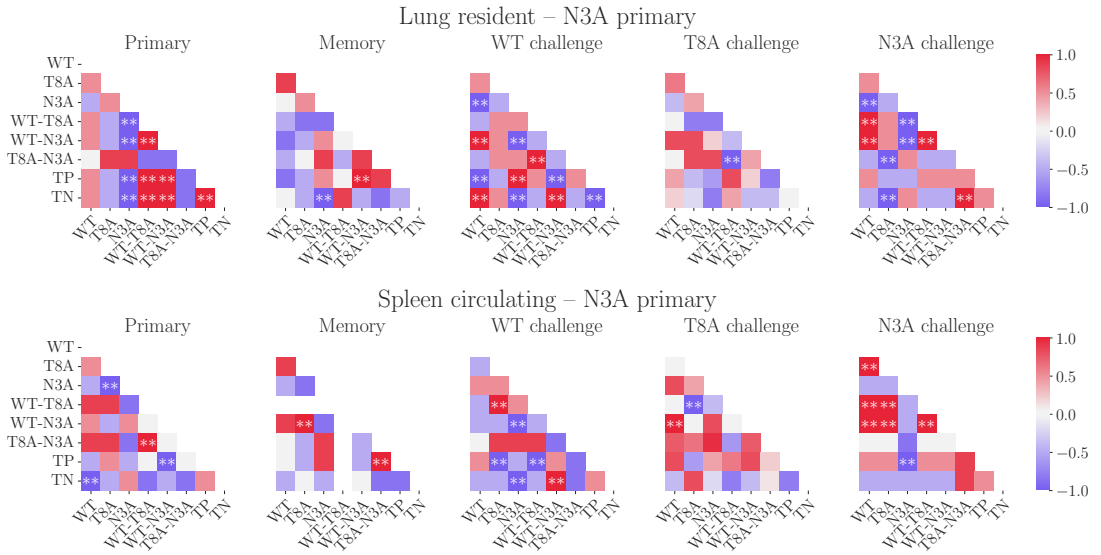


Figure 5.11: Spearman rank correlation coefficients between epitope positive and negative populations for lung resident and spleen circulating cells from mice primed with N3A virus. Asterisks indicate a significant correlation after applying the Bonferroni adjustment ($* \rightarrow p < 0.05$ and $** \rightarrow p < 0.01$).

Now, in the case of the N3A primary infection a starkly different behaviour is observed (see Figure 5.11). For both the WT and T8A-primed mice note that the primary timepoint showed no correlations, suggesting an immune response that is not biased towards any specific epitope positive population. N3A-primed mice, on the other hand, show a bias towards N3A-specific cells at the primary timepoint, as evidenced by the negative correlation with the frequency of tetramer-negative cells. More than this, WT-T8A, WT-N3A, and triple positive cells all are positively correlated to the triple negative population, implying that as the magnitude of the response increases, the frequency of WT double positive cells, and thus, its cross-reactive capability, decreases. Comparing this to the WT primary infection, it can be said that an N3A primary infection is biasing the immune response to be specific for the N3A epitope from the primary timepoint, as opposed to the memory timepoint. At the memory timepoint, the bias to N3A single positive cells in the lungs is maintained (shown by the negative correlation with triple

negative cells), while the spleen shows a positive correlation between cross-reactive populations.

During the challenge infections there is again a difference between N3A-primed mice, and WT and T8A-primed mice. During the WT challenge, the correlations in the lungs show that as the magnitude of the response increases there is an increase in the frequency of N3A single positive and triple positive cells, and a decrease of both WT single positive and WT-N3A double positive cells. These observations suggest that, after being primed with the N3A variant the immune response to a WT challenge is still being pulled towards the N3A epitope, and WT-specific populations are not being expanded in the same as in a WT-primed N3A-challenged mouse. Interestingly, during the T8A challenge few correlations between the populations are found to be significant. In the lungs only WT-T8A double positive cells are negatively correlated to T8A-N3A, suggesting that the response is mostly cross-reactive to the T8A and N3A variants, and not the WT one. Finally, on the N3A challenge there is an increase in the frequency of T8A single positive cells with the magnitude of the response. This suggests that homologous N3A infections can generate cross-reactive responses even in the absence of other variants.

Next, in order to better understand the relationship between the populations of tetramer-positive cells, analysis of variance (ANOVA) (Manly & Alberto, 2016, Chapter 4) was used to determine whether the difference between the mean of the populations for tetramer-positive cells were significantly different, clustering them by their priming infection for lung resident and spleen circulating cells. When significant differences were found, Tukey's HSD test was used to identify which population means were different (Tukey, 1949). In Figures 5.12, and 5.13 the results of this analysis are shown. These results were obtained using the code presented in Appendix C.3. In these figures each node represents each of the possible epitope positive cells, and using Tukey's HSD edges are added to the network when two populations are found to be statistically different in their means. When there is an edge between two populations, the one with a larger node is the one with a greater mean. In all plots the triple negative population has been removed. This is done because the majority of the T cell repertoire will be negative for the IAV epitopes being studied during both infection and

homeostasis, therefore all other populations will always be significantly different in their mean to the triple negative population.

First, consider the lung resident cells, shown in Figure 5.12. By looking at the primary infection timepoint it can be seen that for WT and T8A epitopes there is only one population with a significant difference to others, other than the triple negative population, namely the respective single positive population for each epitope. In the case of the N3A epitope, no significant differences can be identified, meaning that the immune response to a primary infection with this variant is more cross-reactive than that of the other two, since the means of all populations do not behave significantly differently.

At the memory time point, it is observed that WT single positive cells are still significantly different to other populations in WT-primed mice, and interestingly in T8A-primed mice, WT single positive are also significantly different to other populations. However, since these are lung resident cells they do not represent the entirety of the memory compartment, which might explain why there is not a bias towards population of T8A single positive cells after a T8A infection.

Where the most interesting results are found is at the challenge timepoints. In these timepoints, the effects and importance of the order of infection start to become clear. The first thing to note, is that the challenge infections do not behave symmetrically. This means that cross-reactivity is not symmetric, in the sense that inverting the order of the primary and challenge infections generates significantly different immune responses. Focussing first on the WT challenge, it is observed that after homologous infection the response shows no significant differences, meaning that this response showed no particular bias towards any of the epitope positive populations. In contrast to this is the response of T8A-primed mice, in which the WT-T8A double positive population is significantly different to the others, showing that when primed with T8A, a challenge infection with WT will retain some of its preference for the T8A epitope. Finally, for N3A-primed mice the N3A single positive population remains significantly different, even after being challenged with the WT variant.

In the T8A challenge timepoint there is no change in the response of WT-primed mice; that is, there is no population that is significantly different to the others. In T8A-primed mice there is an interesting change in the nature of the

Figure 5.12: Results of ANOVA and Tukey's HSD on lung resident $CD8^+$ cells for all timepoints and all prime-challenge combinations. Each circle represents one of the populations, and a line between two circles represents a significant difference between their means. The size of the circle represents which population is larger in the comparison.

response. In this case there are two populations that are significantly different to others, namely T8A and N3A single positive cells. That means that T8A-primed mice maintain some specificity to T8A after experiencing a homologous challenge, but they also generate a response to an epitope they have not encountered before. Similarly to WT-primed mice, N3A-primed mice also show no change from memory to the WT challenge infection, however in this case this means that the response is still maintaining specificity to the original epitope instead of having a generally even cross-reactive response. Finally, in the N3A challenge timepoint both WT and T8A-primed mice showed no significant differences other than those of the triple negative population, and the N3A-primed mice still maintain their N3A-specific population over cross-reactive populations.

In Figure 5.13 results analogous to the ones discussed above for spleen circulating cells are shown. Since this is not the site of infection the focus will be mostly on the memory timepoint. However, some interesting results on the other timepoints can be observed. First, when primed with the WT variant, the difference of the WT single positive population is present in the spleen during the priming infection, but not during challenge infections. Suggesting that after a single infection the memory generated outside of the site of infection is mostly specific to the WT epitope, but it becomes more cross-reactive after a challenge infection, as evidenced by the lack of significant differences between the population means. Second, during the priming timepoint both T8A and N3A variants present no significant differences. This behaviour extends to the challenge timepoints for N3A-primed mice, but not for those primed with the T8A variant. In this case the populations show an increased frequency of WT-T8A cross-reactive cells in the WT challenge, and the T8A and N3A-specific populations in the T8A and N3A challenges, respectively.

In the memory timepoint, it can be seen that for WT-primed mice the difference of WT single positive cells is maintained in the spleen, further suggesting that the memory generated by this priming infection is biased towards WT single positive cells. In the case of T8A-primed mice, on top of the difference of WT single positive cells observed in the lungs, the T8A single positive population also show a significant difference. This helps explain why there is still a bias to T8A cross-reactive responses after heterologous infection but not a significant difference

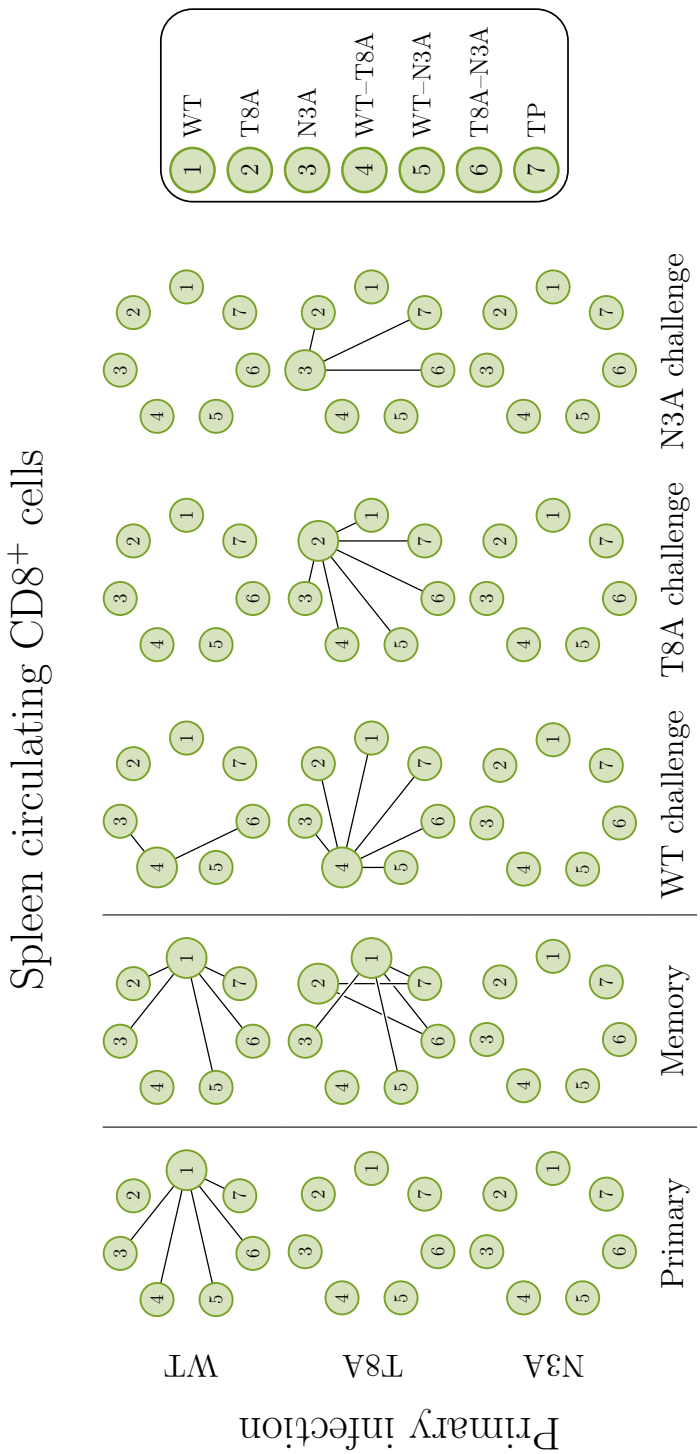


Figure 5.13: Results of ANOVA and Tukey's HSD on spleen circulating CD8$^+$ cells for all timepoints and all prime-challenge combinations. Each circle represents one of the populations, and a line between two circles represents a significant difference between their means. The size of the circle represents which population is larger in the comparison.

of T8A-positive cells in the lungs, because this subset of the memory repertoire is being maintained by circulating cells and not at the site of infection. Finally, N3A-primed mice show no significant differences in the spleen during the memory timepoint, suggesting that the memory repertoire that biases future responses is maintained in the lungs and not in the spleen.

5.4.2 Analysis of epitope-specific T cells

Next, to further understand the effect of the order of infection the populations were grouped by epitope positivity and compared between primary and challenge infections. That is, all cells that are positive for one peptide in a given primary infection are considered as a population, regardless of whether they are single, double, or triple positive, and they are compared with the other primary infections. ANOVA is used to determine whether the difference between the means of these populations were significantly different based on their priming infection for lung resident and spleen circulating cells. Note that, by definition double positive cells will belong to two tetramer-positive populations, and triple positive cells to all three, thus, some cells are being double or triple counted in this analysis. In Figures 5.14, 5.15, and 5.16 the results of these analyses are shown. Similarly to previous ANOVA results, each node represents a population, which in this case is characterised by the primary infection. These results were obtained using the code presented in Appendix C.4.

In Figure 5.14 the results of the analysis for WT-positive cells are presented. In this case, there are overall no significant differences, except for spleen circulating cells in the WT challenge timepoint. This lack of significant differences suggests that WT-positive cells are present at the site of infection regardless of the primary infection the mice are subject to. In the spleen, however, circulating cells in the WT challenge timepoint for N3A-primed mice show increased numbers in WT-positive cells when compared to the other priming infections. This leads to the hypothesis that N3A-positive cells are more cross-reactive with the WT variant than the T8A variant.

For T8A-positive cells, shown in Figure 5.15, a different behaviour is observed. In this case, spleen circulating T8A cells are present in increased numbers in

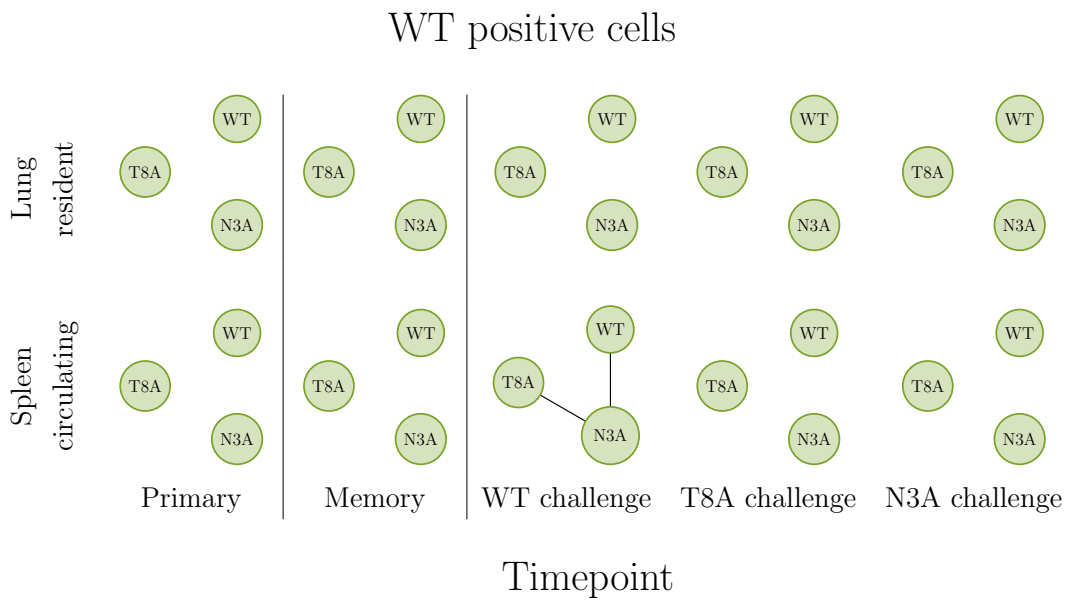


Figure 5.14: Results of ANOVA and Tukey's HSD on WT-positive cells for all timepoints and all prime-challenge combinations. Each circle represents one of the primary infections, and a line between two circles represents a significant difference between their means. The size of the circle represents which populations is larger in the comparison.

T8A-primed mice when compared to WT-primed mice. This suggests that T8A-specific cells less cross-reactive and more biased towards T8A-specific responses than the others, a hypothesis which is further supported by the responses to the T8A challenge infection. In this case, mice that were not primed with T8A have decreased T8A-positive populations when compared to T8A-primed mice, while the two other challenge infections show no statistically significant differences between the primary infections.

Finally, the population of N3A-positive cells was analysed in the same way, and the results are shown in Figure 5.16. At the site of infection, for both the priming and memory timepoints, N3A-primed mice have larger populations of N3A cells when compared to the other primary infections. In the spleen a different behaviour is observed on the primary timepoint, in this case there is an increase on WT-primed mice compared to T8A-primed mice, further suggesting that WT and N3A are more cross-reactive between each other than with T8A. At the memory

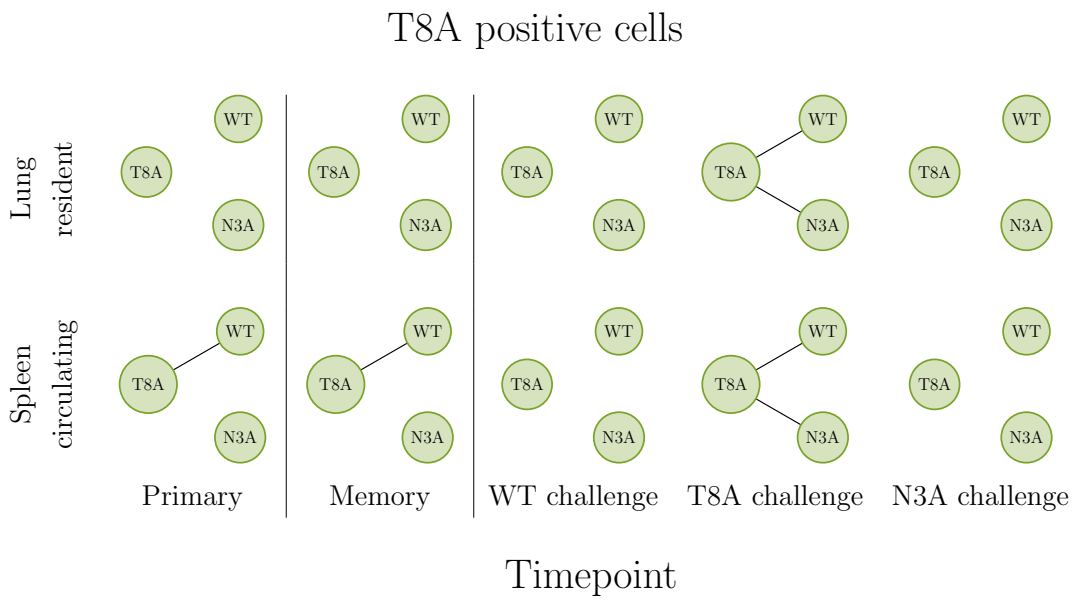


Figure 5.15: Results of ANOVA and Tukey's HSD on T8A-positive cells for all timepoints and all prime-challenge combinations. Each circle represents one of the primary infections, and a line between two circles represents a significant difference between their means. The size of the circle represents which populations is larger in the comparison.

timepoint, for spleen circulating cells there is an increase in N3A-positive cells in N3A-primed mice when compared to WT-primed mice, but there is no significant difference between N3A and T8A-primed mice.

Focussing on the challenge timepoints, the cross-reactive behaviour of the N3A variant can be better understood. When challenging with the WT variant it can be seen that N3A-primed mice will have a stronger N3A response, meaning that the immune response is being recalled even after infection with a different strain. This pattern is lost on the T8A challenge, where T8A-primed mice have a larger N3A-positive population when compared to WT-primed mice. This behaviour further supports the previous suggestion that WT and N3A-positive cells are more cross-reactive with the N3A and T8A variants, respectively, than with the T8A variant. The biological reason for this is hypothesised to be that the T8A and N3A peptides have drifted enough from each other that when there is a heterologous infection, an immune response to it behaves closer to a primary infection than

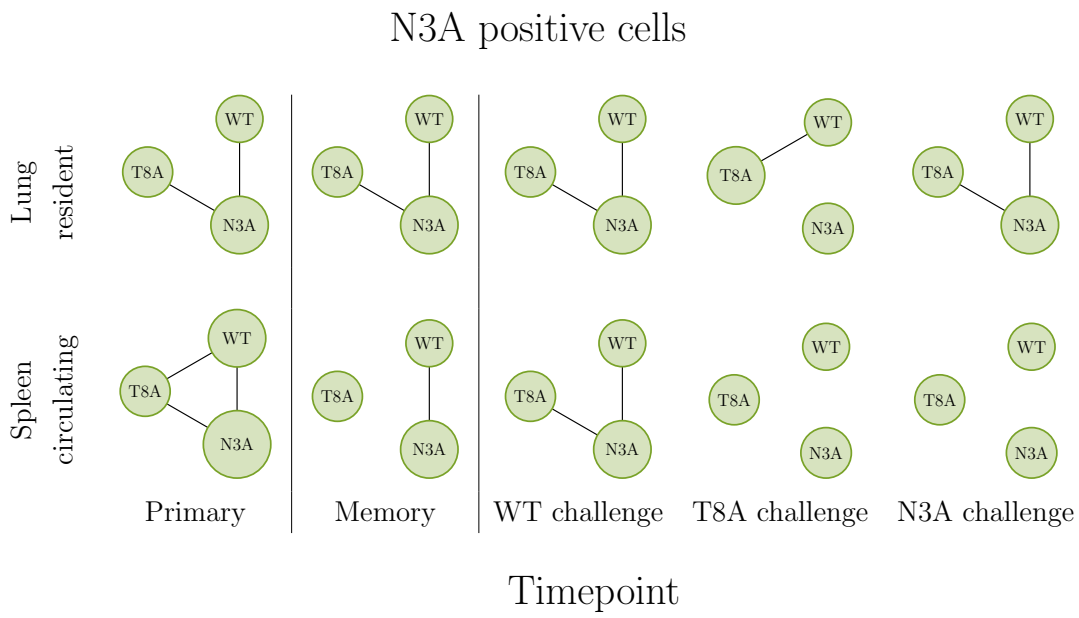


Figure 5.16: Results of ANOVA and Tukey's HSD on N3A-positive cells for all timepoints and all prime-challenge combinations. Each circle represents one of the primary infections, and a line between two circles represents a significant difference between their means. The size of the circle represents which populations is larger in the comparison.

a challenge infection. Meaning that, little cross-reactive immunity is conferred between the two variants in comparison to the WT variant.

5.4.3 Contraction after primary infection

So far only the populations of epitope-specific T cells at a given timepoint have been considered, but given the data available the dynamics between timepoints can also be studied. However, given that each data point is independent from the others, due to the fact that they are taken from different mice, they cannot be clustered by time. Therefore, instead of considering the contraction between a specific sample in the primary timepoint and a sample in the memory timepoint, the mean value of the population is calculated at the primary timepoint for every variant, and the slope of contraction is calculated for the log values of each sample in the memory timepoint. This results in a set of linear decay rates from the

primary to the memory timepoint clustered by primary infection. An example of the points used to calculate the decay rates is shown in Figure 5.17. Here, the contraction of the total number of tetramer-positive cells is plotted for all primary infections. It is important to note that due to the low copy numbers of T cells in homeostasis (Jameson, 2002) some population frequencies will be below the level of detection of flow-cytometry, and therefore will be represented as the population not being present in the sample; that is, the population will have a frequency of 0. With this in mind, in order to prevent biasing the results to a faster contraction rate, any samples below the level of detection have not been included in the analysis.

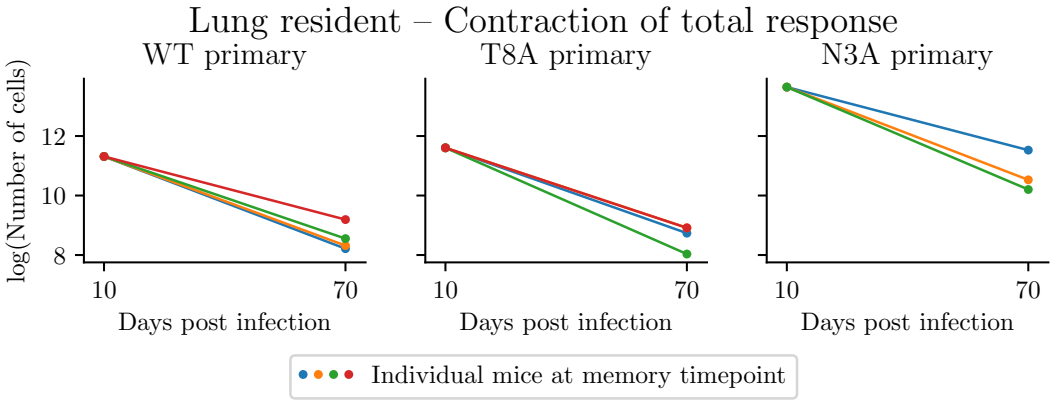


Figure 5.17: Example of contraction from primary infection to memory of the total population of lung resident epitope-specific cells. Each colour represents a different sample. For every epitope specificity (single, double, and triple positive), as well as the total number of cells of any specificity, the slopes of contraction are calculated for all primary infections to be compared.

In order to compare the decay rates between primary infections, they are clustered by primary infection and ANOVA is used together with Tukey's HSD test to identify significant differences. The results of this analysis (using the code presented in Appendix C.5) for every tetramer positivity, as well as the total immune response, are shown in Figure 5.18. To interpret these results, it is important to keep in mind that what is being compared are the signed exponents of the exponential decay from primary infection to memory. Therefore, when an

exponent is greater this means that the population is contracting at a slower rate than one with a smaller exponent.

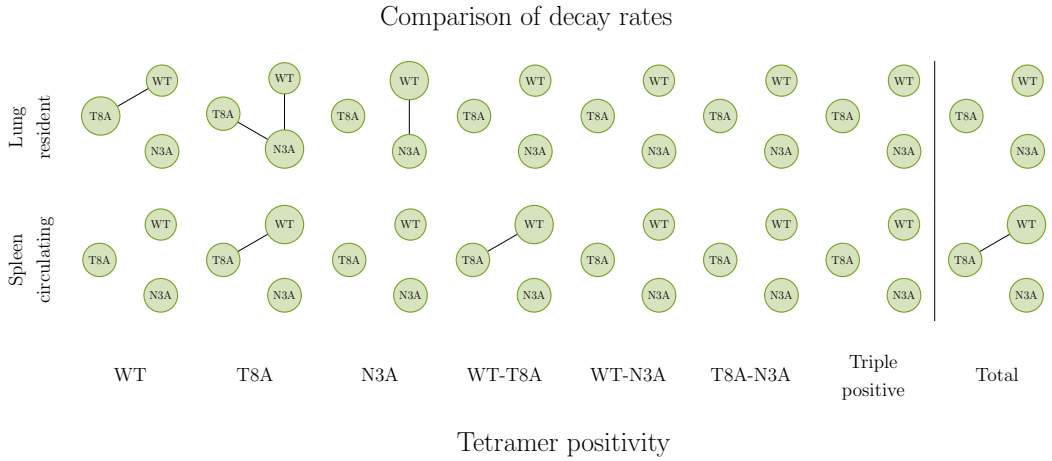


Figure 5.18: Results of ANOVA and Tukey's HSD on decay rates for each epitope positive population. Each node represents the slopes for a primary infection, and a line between nodes represents a significant difference between the means of the slopes. When an edge connects two nodes the larger node is the one that is larger in the comparison. Note that in this context a larger node means that the population is contracting at a slower rate.

From the results shown in Figure 5.18, it is easy to see that the dynamics of contraction into memory is different between the site of infection and the spleen. Not only are the populations of tetramer-specific cells behaving differently between tissues, but when comparing the total population of epitope positive cells different behaviours are observed. In the spleen, mice primed with the WT variant show a slower contraction into memory than T8A-primed mice, while in the lungs there are no significant differences observed in the decay rates. This suggests that for primary infections, the decay of immune cells does not behave significantly differently between the possible variants the mice were infected with. In terms of conferred immunity at the memory timepoint, considering the same hypothesis on generation of memory assumed in Section 5.1.2; that is, memory is generated from the effector population after the infection is cleared, all variants generate memory proportionally to the magnitude of the response, and the primary infection does not inherently change the behaviour of the overall response.

Focussing now on the different epitope specificities, it is easy to see that at the site of infection single positive cells for the epitope of the primary infection decay at a faster rate than others. This seemingly contradictory behaviour is explained by the fact that during the primary response it is the matching single positive cells which make up the majority of single positive cells, thus, in order for the population to contract back to the numbers measured at the memory timepoint (which are within an order of magnitude form one another) it must decay at a faster rate than the others. Comparing this to the results on the memory timepoint of the ANOVA on epitope-specific cells (see Figure 5.12), the data suggests that after this rapid contraction into memory both WT and N3A single positive cells are given precedence in the memory compartment and maintained in larger numbers than the other single positive cells. In the case of T8A single positive cells, not only are cells decaying faster in T8A-primed mice, but also in WT-primed ones.

For double and triple positive cells no significant differences are found in lung resident cells, suggesting that these populations either do not take an active part in the primary challenge immune response, or they do and they all behave similarly. Going back to the results of the epitope-specific ANOVA (see Figure 5.12), it can be seen that during primary infections there is no significant difference that would indicate an increase in the activation of double or triple positive populations. In the case of spleen circulating cells, there is a significant difference for WT-T8A double positive cells. The analysis indicates that this population of cells is decaying at a faster rate on T8A-primed mice compared to WT-primed ones. Given that the results presented in Figure 5.13 showed that there were no significant differences between cell populations during the primary infection, this faster decay rate suggests that the population of spleen circulation cells observed at the memory timepoint does not come exclusively from spleen circulating precursors, otherwise a difference in the population means would be observed.

5.4.4 Expansion during challenge infection

Studying the decay rates after the primary infection allows for the study of contraction into memory, as discussed above. However, to understand the effect of the primary infection on future challenge infections it is necessary to study

the dynamics of expansion during the challenge infection. Similarly to the study of decay rates, the slopes of expansion are calculated by considering the mean value of each population at the challenge timepoint and the individual memory values, and computing the slope of the log values for each prime-challenge infection combination. An example of the data points used to calculate the expansion rates of the total immune response is shown in Figure 5.19.

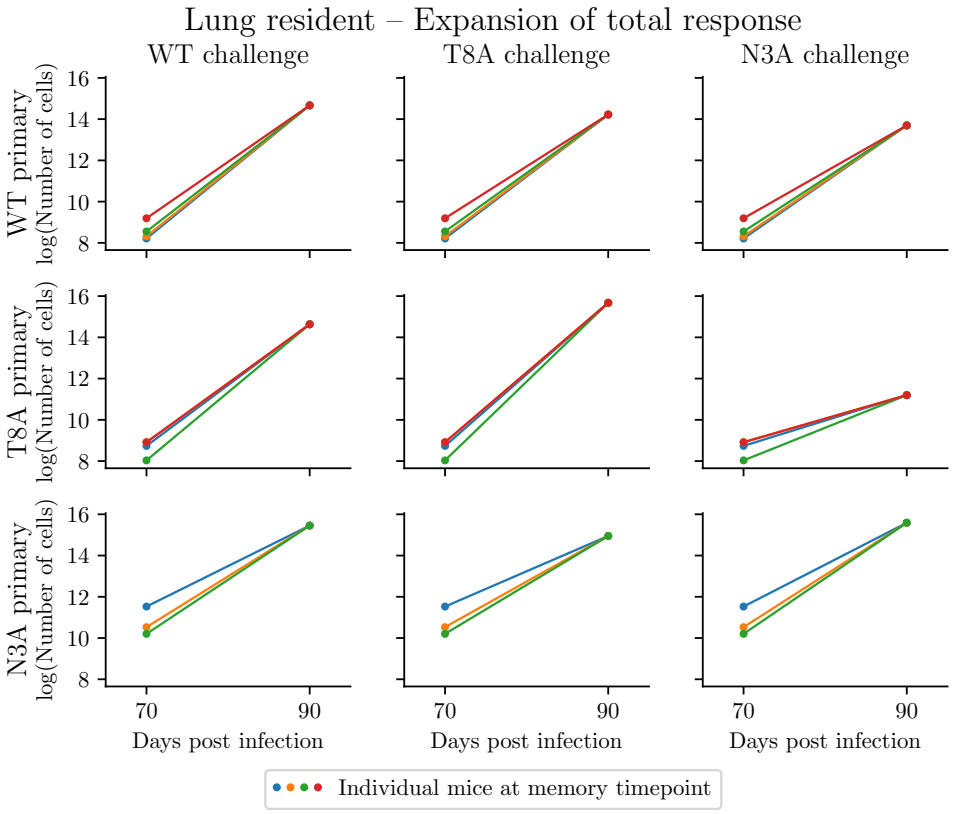


Figure 5.19: Example of expansion from memory to challenge infection of the total population of lung resident epitope-specific cells. Each colour represents a different sample. For every epitope specificity (single, double, and triple positive), as well as the total number of cells of any specificity, the slopes of expansion are calculated for all prime-challenge infection pairs.

ANOVA and Tukey's HSD test are used on the resulting set of expansion rates to identify differences between the dynamics of each prime-challenge infection pair. In this case the rates are clustered first by primary infection, and then

by challenge infection, so that for each primary infection the dynamics of all its following challenge infections can be compared. First, consider the population of lung resident cells, since they are the population that drives the immune response (Gebhardt *et al.*, 2009; Jiang *et al.*, 2012; Masopust & Soerens, 2019). The results of the statistical analysis (using the code presented in Appendix C.5) for this population of cells are shown in Figure 5.20. It is important to note that, unlike the results presented in Section 5.4.3, here each node in the figure represents a challenge infection, and an infection having a greater mean expansion rate means that when infected with that variant the population expanded faster than with other variants.

Focussing first on the total response, it is observed that the T8A variant responds generally better to homologous infection, evidenced by the fact that when primed and challenged with T8A the expansion rate is greater than when challenged with other variants. This suggests that this variant has a propensity to generate responses focussed on its own NP_{366} epitope. More than this, the expansion during a WT challenge is significantly different to that of an N3A challenge, further supporting the hypothesis that cross-reactivity is not a symmetric relationship, as discussed previously in Section 5.4.2. Comparing the expansion rates allows for an even greater understanding of this asymmetry, as it shows that not only is the combination of infections a relevant factor, but their order matters as well. For the T8A primary infection in particular, the analysis suggests that the total expansion of the immune response is greater with a homologous challenge, followed by a WT challenge, and finally an N3A challenge elicits the slowest expansion. In terms of cross-reactivity this means that for the T8A variant, the WT and N3A variants can be placed in a spectrum, with WT being more cross-reactive to T8A than N3A is.

In the case of WT-primed mice, the analysis shows that only the expansion rate for the WT homologous challenge is significantly different to that of the N3A challenge infection. Meanwhile, there is no significant difference found between the T8A challenge and the other challenges. This means that, similarly to the T8A-primed mice, the total response to a homologous infection will be stronger than an N3A challenge infection, but no significant difference is found to the response to a T8A challenge. Comparing these results to those of the T8A-primed mice, it can be said that WT-primed mice are more widely cross-reactive, since they only show

Figure 5.20: Results of ANOVA and Tukey's HSD on expansion rates for each epitope positive population of lung resident CD8$^+$ cells. Each node represents the slopes for a challenge infection, and a line between nodes represents a significant difference between the means of the slopes. When an edge connects two nodes the larger node is the one that is larger in the comparison. Note that in this context a larger node means that the population is expanding at a faster rate.

a significant difference between WT and N3A challenges. Finally, considering the total response of N3A-primed mice no significant differences between the expansion rates at the challenge timepoint are found. This is interesting, as it suggests that regardless of the challenge infection these mice mounted similar immune responses, at least in magnitude, if not in composition. This can be then interpreted as the N3A variant being the most widely cross-reactive with the other variants.

In order to better understand the composition of the immune responses being generated by the different prime-challenge combinations, the different tetramer-positive populations have to be considered. First, note that both triple positive, and T8A-N3A double positive cells have no significant differences between their expansion rates in any prime-challenge combination. This suggests that these two types of cells are not taking an active part in the response, since if they were, T8A-N3A double positive cells would be expected to have a greater expansion rate in T8A or N3A challenge infections. This is further supported by the results of Section 5.4.1, where it was shown that neither of these populations is significantly greater in any of the challenge timepoints.

Now, consider first the population of WT single positive cells. It is observed that they exhibit a similar behaviour for WT and T8A-primed mice, where a WT challenge will cause a faster expansion of the populations, followed by a T8A challenge, and finally an N3A challenge with the smallest of the expansion rates. Only N3A-primed mice display a different behaviour, where there are no significant differences between the challenge infections, which is consistent with what has already been noted on the total immune response. For T8A single positive cells, the results show that in general the population has a higher expansion rate when the challenge is with the T8A variant, but the primary infection has an effect on how the other challenges compare. On WT-primed mice it is only the T8A challenge that shows an increased expansion rate, while a WT challenge shows no significant differences. On T8A-primed mice both WT and T8A challenges cause the population to expand at a faster rate than an N3A infection. Finally, in N3A-primed mice it is T8A and N3A challenges which cause the population of cells to expand faster. For N3A single positive cells only WT-primed mice

show any significant differences, in this case with both T8A and N3A infections resulting in a faster expansion rate for these cells.

As in previous analyses, the expansion rates of spleen circulating cells are also considered. The results of the analysis of these cells are shown in Figure 5.21. Similarly to lung circulating cells, when looking at the expansion rates of the total response each primary infection behaves differently. In this case it is WT-primed mice that show no significant differences between challenge infections, suggesting that when primed with the WT variant, recall responses are mainly in the form of lung resident cells, and there is less migration from the spleen to the site of infection. For T8A-primed mice, there is an increase in the rate of expansion for WT and T8A challenge infections, indicating that part of the immune response is being recalled from the spleen and lymph nodes for WT and T8A infections, but not for the N3A challenge, further supporting the hypothesis that these two variants have drifted enough that during a challenge infection a portion of the immune response is being generated for the first time at the site of infection, and is not being recalled from memory.

5.5 Discussion

Cross-reactive immune responses in the context of heterologous infections are still not yet fully understood (Duan *et al.*, 2015; Elong Ngono & Shresta, 2019; Mateus *et al.*, 2021; Moris *et al.*, 2011; Webster & Askonas, 1980). However, it is a topic of great importance in the context of vaccine development (Elong Ngono & Shresta, 2019; Mateus *et al.*, 2021; Tamura *et al.*, 2005; Webster & Askonas, 1980), as better understanding of the immunogenicity of cross-reactive responses can lead to "universal" vaccines that confer immunity to several influenza sub-types (Jang & Seong, 2019; Kumar *et al.*, 2018; Paules *et al.*, 2018; Zens *et al.*, 2016).

The original aim of this chapter was to use a deterministic model, together with experimental data, to study the effects of cross-reactivity. However, as demonstrated in Section 5.3.2, the parametrisation of this model was not possible with the available data, thus a different approach to the study of cross-reactivity using the model presented in Section 5.1 will be presented in Chapter 6. Instead, statistical analysis of novel experimental data provided by Jessica Gaevert at the

Figure 5.21: Results of ANOVA and Tukey's HSD on expansion rates for each epitope positive population of spleen circulating CD8$^+$ cells. Each node represents the slopes for a challenge infection, and a line between nodes represents a significant difference between the means of the slopes. When an edge connects two nodes the larger node is the one that is larger in the comparison. Note that in this context a larger node means that the population is expanding at a faster rate.

Paul Thomas laboratory in the Immunology department of St. Jude Children's Research Hospital was performed. A hypothesis that was derived from this analysis is the existence of *selfish*, and *selfless* epitopes. A selfish epitope is one that in general biases the immune response to be specific to itself; that is, the population of single positive cells specific to it has greater avidity that double and triple positive populations. Thus, when an infection occurs, the immune response will be biased towards single positive epitope-specific cells, and the cross-reactive capabilities will be diminished. A selfless epitope, on the other hand, is one that generates broadly cross-reactive responses which does not pull the response toward any specificity in particular; that is, the populations of cells that can recognise it (single, double, and triple positive), have similar avidity to the epitopes they can each recognise, meaning that when there is an infection, the immune response will not be biased towards single positive cells, and will have greater cross-reactive capabilities. This was originally hypothesised when preliminary analysis of the data showed that the N3A variant of the virus generated more broadly cross-reactive immune responses when compared to the T8A variant. The working hypothesis at the time was that a combination of the initial frequency of epitope-specific cells (initial conditions of the deterministic model), and the avidity of each clonotypes to the different variants (the $\kappa_v(i)$ parameter) could explain this difference in the behaviours. However, given the available data, structural identifiability analysis showed that only three parameters already widely reported in the literature (Ahmed *et al.*, 2015; Macallan *et al.*, 2003, 2019; Vrisekoop *et al.*, 2008) could be identified. For this reason, instead of parametrising the model with the data, statistical tools were used to gain a better understanding of the behaviours being observed.

The initial study of correlations between the different epitope-specific cells showed that the WT and T8A variants appeared to generate more cross-reactive responses than the N3A variant, based on the fact that very few significant correlations where found on the primary timepoint of these two infections. This was interpreted to mean that the sub-populations of clonotypes were not being coupled, and thus, the response was allowed to behave in a more stochastic fashion between each mouse. On the other hand, N3A-primed mice displayed several correlations, indicating that the responses were being guided towards specific populations. For example, the negative correlation between T8A-N3A

double positive cells and WT-T8A double positive cells on N3A-primed-T8A-challenged mice suggested that after the challenge infection the response was being pulled towards cross-reactivity with the WT or N3A variant, but not both. Moreover, positive correlations with the triple negative population suggested that the magnitude of the response was an indicator of its cross-reactive capabilities.

In order to corroborate these results ANOVA was used to find significant differences between the means of each of the possible epitope positive populations, both by epitope specificity (single, double, or triple positive) and by general epitope recognition (WT, T8A, or N3A-positive). It was these results that started to paint a clearer picture of the cross-reactive responses to the three virus variants. The first thing that was noted was that during the primary infection both WT and T8A-primed mice showed an increased population of their single positive specific cells, while the N3A-primed mice showed no significant differences between populations (see Figure 5.12). This contradicted what had been found with the correlations, so these results were compared with the ANOVA of general epitope positive cells. Here it was found that both WT and T8A-primed mice showed no significant differences between epitope positive cells at the primary timepoint, while N3A-primed mice had an significant increase in N3A-positive cells (see Figure 5.16). This explains the earlier contradiction between the ANOVA and the correlations, the N3A-primed show an increased number of single positive cells in general, but they are increased across the board, and thus, when looking at the individual populations the differences get lost in the random noise between samples. The second behaviour observed was that during challenge infections, the order had a noticeable effect on the composition of the response, as noted by the asymmetry of the challenge timepoints in Figure 5.12. This means that, at least in the particular case of influenza viruses, the order of infection is crucial in determining the shape of the immune response for future infections.

Further than this, the dynamics of the immune responses was compared between primary and challenge infections by considering an exponential decay model from primary infection to memory, and an exponential growth model from memory to challenge infection, and comparing the decay and growth of the data using ANOVA. In the case of contraction into memory (see Figure 5.18), it was found that while in general the decay of populations was not significantly different

between variants, some sub-populations did show differences between variants. At the site of infection these differences were only found in single positive cells, which can be explained by the fact that between single positive populations it is always the matching one to the infection which expands to a greater degree, and thus, must contract at a different rate in order to reach homeostatic levels with the other single positive populations.

When comparing the dynamics of expansion during the challenge infection it was found that the total immune response showed different behaviours between the primary infections. It was found that N3A-primed mice had no significant differences between the expansion rates, indicating that the N3A variant is, of the three studied, the one less likely to bias the repertoire towards a specific epitope specificity. On the other hand, the T8A variant is shown to have a clear preference towards T8A-positive responses, as evidenced by the increased expansion rate of the total immune response. Altogether, while the deterministic model could not be parametrised with the experimental data, statistical analysis showed that, in the case of influenza viruses, cross-reactivity is not symmetric between variants. Meaning that, the history of infection is a determining factor of the composition of the immune response. Not only that, but also that the selfishness of epitopes can be considered as a spectrum. Some epitopes will be selfish, when they generate a cross-reactive immune response it will still biased towards recognition of the primary infection's epitope, *e.g.*, the T8A variant, others can be selfless, generating an immune response that is broadly cross-reactive and is not biased to recall the original specificity, *e.g.*, the N3A variant, or they can somewhere in-bewteen selfish and selfless, where they generate a cross-reactive response, which is biased to recall the specificity of the primary infection to a lesser extent than a purely selfish epitope, *e.g.*, the WT variant.

5. PERTURBATION OF THE DISTRIBUTION OF T CELL CLONOTYPES BY VIRAL INFECTION

Chapter 6

Random recognition networks of viral peptides

In the context of heterologous infections, as defined in Section 5.3, T cell clonotypes have the potential to be cross-reactive and mount an effective immune response when presented with epitopes they have not encountered before (Duan *et al.*, 2015; Gaevert *et al.*, 2021; Souquette & Thomas, 2018). However, the biological mechanisms and cell population dynamics of these cross-reactive T cell immune responses are not yet fully understood (Adams *et al.*, 2016; Lang *et al.*, 2002; Mazza *et al.*, 2007; Nelson *et al.*, 2015; Yin & Mariuzza, 2009).

What is widely understood to be true, is that each TCR must have more than one cognate epitope (Lang *et al.*, 2002; Mason, 1998; Selin *et al.*, 1994; Sewell, 2012; van den Berg *et al.*, 2011; Yin & Mariuzza, 2009); that is, TCR-pMHC interactions do not follow a *"lock and key"* strategy, and each TCR has a set of multiple epitopes which it is able to recognise. Given the current incomplete understanding of TCR cross-reactivity, hypotheses have been made about why this phenomenon arises (whether by chance or with some underlying logic), and how it affects the dynamics of an immune response. One such hypothesis, proposed by Mason (1998), is that cross-reactivity can be either *focussed* or *unfocussed*. In his hypothesis, Mason defines focussed cross-reactivity as the case when T cell clonotypes that recognise a given peptide are more likely to be cross-reactive with each other for other peptides, and unfocussed when they are cross-reactive with other clonotypes for random peptides instead. In other words if a k-partite recognition network,

like the one defined in Chapter 5, is considered, unfocussed cross-reactivity would be represented by a network in which the edges between TCR and VDP nodes are completely random, and focussed cross-reactivity would be represented by a network in which there is an underlying structure that defines the edges that are present in the network. Network models have been considered in the past to model TCR-peptide interactions (De Boer, 1988; Jerne, 1974a). These network models are based on *idiotypic* interactions (De Boer, 1989; Jerne, 1974b, 1984); that is, random interactions between both receptors and their cognate particle, and receptors to other receptors, in order to consider self regulation of the population via cell-cell interactions. Given the random construction of these networks, they can be considered examples of the unfocussed hypothesis. An example of the focussed hypothesis is the concept of *shape space* for TCR-epitope interactions (Perelson & Oster, 1979), in which a multi-dimensional Euclidean vector space is considered with each dimension representing an antigenic determinant parameter. Using this space, neighbourhoods of epitope recognition for a given TCR are considered, and cross-reactivity between clonotyes is determined by the location and size of their recognition neighbourhoods.

The multi-variate competition process defined in Chapter 5, was proposed as a model for the dynamics of heterologous infection in the context of a known TCR-VDP recognition network. However, in order to study the focussed versus unfocussed hypothesis a more general approach to the construction of the TCR-VDP recognition network is required. That is, instead of considering a network representing the recognition profile of a specific pathogen as in Chapter 5, random networks will be considered. Since the recognition network is nothing more than a bipartite network in which the recognition profile of a T cell clonotype is encoded, random network generation algorithms can be used to construct recognition networks with different properties, such as having focussed or unfocussed cross-reactivity.

The main focus of this chapter is to explore the focussed and unfocussed hypothesis by using random network generation algorithms to construct the TCR-VDP recognition network. The networks generated will be small, consisting of a limited number of clonotypes and peptides, η and ε respectively, in order to study the effects that focussed and unfocussed cross-reactivity have on T cell mediated immune responses. The clustering coefficient of the networks is calculated in order

to compare the underlying community structure of the different types of networks being considered. In particular, the clustering coefficient considered here will measure the *cliquishness* of the networks; that is, the density of small connected sub-networks of the recognition network.

In Section 6.1 the random network generation algorithms used to generate small TCR-VDP recognition networks are introduced. Three network generation algorithms are considered: one which can be used for both focussed and unfocussed cross-reactivity, and two for focussed cross-reactivity only. Section 6.2 introduces the clustering coefficient that will be used to measure the degree of betweenness centrality of the random recognition networks (Newman, 2018, Chapter 7), and using the clustering coefficient proposed by Zhang *et al.* (2008), novel approximations of this coefficient for the different types of networks considered are found. Finally, in Section 6.3 the random networks are used together with the multi-variate competition model defined in Chapter 5 to study the dynamics of cross-reactive T cell immune responses. The analysis of the resulting dynamics suggests that all the random network generation algorithms considered yield biologically plausible immune responses, displaying immunodominance and distinct immune responses to heterologous infection. Given that the networks used exhibit both focussed and unfocussed cross-reactivity, this indicates that both hypotheses can lead to biologically relevant TCR-VDP recognition networks.

6.1 Random generation of TCR-VDP recognition networks

As in previous chapters, the sets considered for the TCR-VDP recognition network are: the set of T cell clonotypes being modelled ($\mathscr{C}$), and the sets of VDPs present during a first and second infection ($\mathscr{V}$ and $\mathscr{W}$, respectively), which need not be disjoint since a peptide can be presented during more than one infection (Boon *et al.*, 1994; Eickhoff *et al.*, 2019; Heiny *et al.*, 2007; Tan *et al.*, 2011; van der Bruggen *et al.*, 2002). Note that the names *first* and *second* are arbitrary here, given that the focus of this chapter is the effects of focussed and unfocussed cross-reactivity and not the specific order of the infections, their order is not relevant

and the names are used to make clear that the infections occur at different times. Since the main interest of this chapter is cross-reactivity in the context of infection, the set of self-pMHCs which can stimulate the clonotypes in $\mathscr{C}$ (namely $\mathscr{Q}$) will not be considered when constructing the random recognition networks. A general example of the types of networks that will be considered is shown in Figure 6.1.

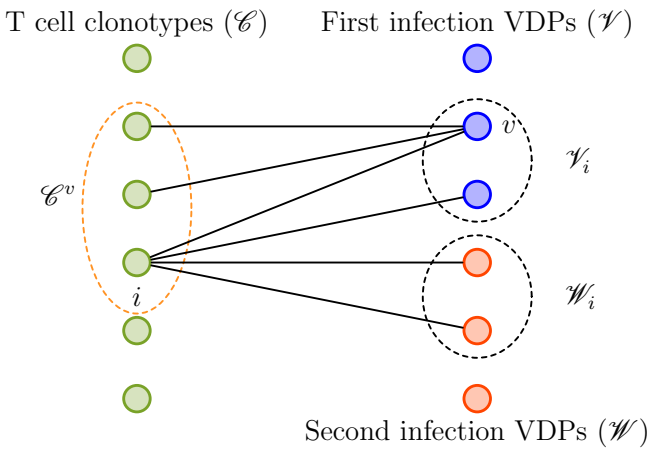


Figure 6.1: Bipartite recognition network of T cell clonotypes and VDPs for heterologous infection. The set of T cell clonotypes is $\mathscr{C}$, and the VDPs are separated into the set of peptides presented during the first infection, $\mathscr{V}$, and the set of those presented during the second infection, $\mathscr{W}$. Nodes belonging to $\mathscr{C}$, *i.e.*, clonotypes, are represented as green circles, and VDPs are represented as blue circles if they belong to $\mathscr{V}$, or red circles if they belong to $\mathscr{W}$. As in Chapters 4 and 5, each peptide $v \in \mathscr{V} \cup \mathscr{W}$ has a set $\mathscr{C}^v$ of clonotypes that can recognise it, and each clonotype $i \in \mathscr{C}$ has sets $\mathscr{V}_i$ and $\mathscr{W}_i$ of peptides, not necessarily disjoint, they are able to recognise.

In the case of unfocussed cross-reactivity, the desired property is that the bipartite recognition network is completely random; that is, every edge is added to the network independently of other edges. In this case, the Erdős-Rényi model for network generation will be introduced to define this type of recognition network (Erdős & Rényi, 1959; Gilbert, 1959). However, a different model, called the stochastic blockmodel, will also be discussed because of desirable properties that it has for the generation of TCR-VDP recognition networks (Airoldi *et al.*, 2008; Holland *et al.*, 1983; Wang & Wong, 1987). For focussed cross-reactivity two

different random networks are considered: a configuration model in which a degree sequence is used to define the degree of each clonotype node (Bollobás, 1980), and a modified preferential attachment network based on the Barabási-Albert model (Barabási & Albert, 1999).

6.1.1 Erdős-Rényi and stochastic blockmodel networks

In order to define the Erdős-Rényi network generation model, first for each VDP $v \in \mathscr{V} \cup \mathscr{W}$, p_v is defined to be the probability that a given TCR will be able to recognise v. If the value of p_v is the same for all v in $\mathscr{V} \cup \mathscr{W}$, then it is called an Erdős-Rényi network, denoted by $G(\eta, \varepsilon, p_v)$, where η is the number of clonotypes, and $\varepsilon = |\mathscr{V} \cup \mathscr{W}|$ the number of peptides. However, this is a strong assumption on the p_v probabilities, since structural differences between peptides, single amino acid changes, or structural differences between TCRs can have drastic effects on TCR-pMHC binding (Borg *et al.*, 2005; Gagnon *et al.*, 2005; Gras *et al.*, 2009; Ishizuka *et al.*, 2008). Therefore, a model closer to reality should consider some degree of variability for the p_v probabilities. This can be thought of as variable degenerate specificity of the TCRs themselves; that is, the variability of the number of epitopes each TCR can recognise (Joshi *et al.*, 2001; Stewart-Jones *et al.*, 2003), or conversely some peptides being more likely to be recognised due to their 3D structure (Carson *et al.*, 1997; Cole *et al.*, 2010; Day *et al.*, 2011; Turner *et al.*, 2005).

Since the Erdős-Rényi model considers every VDP to be exactly equal in terms of its recognition probability, it is unable to capture this variability of TCR-peptide interactions. One model that allows for this variability is the stochastic blockmodel (Holland *et al.*, 1983; Wang & Wong, 1987). In this model the set of VDPs is separated into M disjoint subsets B_ℓ, called blocks, such that $\mathscr{V} \cup \mathscr{W} = \bigcup_{\ell=1}^{M} B_\ell$. Each of these blocks is assigned a recognition, or edge, probability p_ℓ for $\ell = 1, \ldots, M$, so that every VDP $v \in B_\ell$ has recognition probability $p_v = p_\ell$. That is, the network is defined by the number of blocks, M, and the probability vector

$$\mathbf{p} = (p_1, p_2, \ldots, p_M). \tag{6.1}$$

If the p_ℓ are all the same, then the resulting network is equivalent to one generated by the Erdős-Rényi model. In any other case, the network will capture the

possibility that some peptides will be more or less likely to be recognised, since it is exactly the recognition probabilities that are being changed.

For the purposes of modelling TCR-VDP interactions, this model is an improvement over the Erdős-Rényi model. However, it still fails to capture variable recognition probabilities based on differences between TCRs. In order to account for this type of recognition variability a mixed membership stochastic blockmodel is considered. This is a special case of the stochastic blockmodel, in which each TCR has an associated mixed membership vector (Airoldi *et al.*, 2008). Denote by $\boldsymbol{\pi}_i$ the mixed membership vector of clonotype $i \in \mathscr{C}$,

$$\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iM}), \tag{6.2}$$

where $0 \leq \pi_{i\ell} \leq 1$ is the degree of membership of clonotype i to the block ℓ of peptides; that is, the probability of clonotype i recognising a VDP of block ℓ is given by $\pi_{i\ell}p_\ell$. Note that membership to each of the blocks is considered to be independent, and therefore the $\pi_{i\ell}$ do not need to add up to one. Then, it is easy to see that the recognition probabilities of clonotype i under this model are given by the Hadamard product of $\boldsymbol{\pi}_i$ and $\mathbf{p}$ (Marcus & Khan, 1959)

$$\boldsymbol{\pi}_i \circ \mathbf{p} = (\pi_{i1}p_1, \pi_{i2}p_2, \ldots, \pi_{iM}p_M), \tag{6.3}$$

where the $\pi_{i\ell}$ represent the effects of TCR structure on peptide recognition, and p_ℓ the effects of peptide structure. Figure 6.2 shows an illustrative example of a network consisting of three clonotypes, and two blocks of VDPs. In this example each clonotype has a different mixed membership vector. Clonotypes 1 and 3 can only recognise one of the blocks; that is, their mixed membership vectors are $\boldsymbol{\pi}_1 = (1, 0)$, and $\boldsymbol{\pi}_3 = (0, 1)$, while clonotype 2 has a membership vector $\boldsymbol{\pi}_2 = (1, 1)$, and is able to recognise both blocks. While this example considers only the extreme values 0 and 1 for $\pi_{i\ell}$, these probabilities can take any value between 0 and 1.

It is easy to see that this model is extremely versatile, allowing for the generation of recognition networks that consider degenerate TCR specificity in the form of mixed membership vectors, peptide recognition variability in the form of blocks of peptide nodes, both, or neither. This makes it a perfect candidate for an in depth study of cross-reactivity on large (repertoire wide) networks. However,

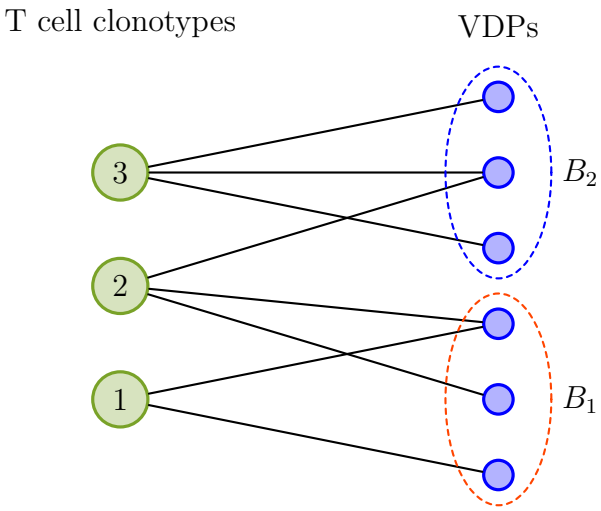


Figure 6.2: Example of a mixed membership stochastic blockmodel network. Two blocks of VDPs are considered, B_1 and B_2 (shown here in blue and red), and three clonotypes. Clonotypes 1 and 3 have mixed membership vectors $\boldsymbol{\pi}_1 = (1, 0)$, and $\boldsymbol{\pi}_3 = (0, 1)$, meaning they can only recognise one of the blocks. Clonotype 2 has a membership vector $\boldsymbol{\pi}_2 = (1, 1)$, and can recognise both blocks. In this example only the extreme values 0 and 1 are considered for $\pi_{i\ell}$, but these probabilities can take any value between 0 and 1.

since the focus here is on the effects of cross-reactivity in small scale and not large scale networks, the analysis of this model on repertoire wide networks is beyond the scope of this chapter. For this reason, in Section 6.3 only the extreme case with one block, $M = 1$, and all clonotypes having membership vector $\boldsymbol{\pi} = (1)$, which reduces the model to an Erdős-Rényi network, is considered. An outline of the network generation algorithm for this model is given in Algorithm 6.1.

6.1.2 Configuration model networks

The first model considered for the focussed cross-reactivity hypothesis is the configuration model (Bender & Canfield, 1978; Bollobás, 1980; Molloy & Reed, 1995). In this network generation model, instead of a recognition probability, two sequences of integers $K_C = (k_1, \ldots, k_\eta)$, and $K_P = (k'_1, \ldots, k'_\varepsilon)$, such that $\sum_{i=1}^{\eta} k_i = \sum_{i=1}^{\varepsilon} k'_i$, are considered. They are called the degree sequences, and

Algorithm 6.1: Mixed membership stochastic blockmodel network generation algorithm for a network with η clonotypes.

Input: $M \leftarrow$ number of peptide blocks,
$\mathbf{p} \leftarrow$ vector of peptide block recognition probabilities,
$\boldsymbol{\pi}_i \leftarrow$ mixed membership vectors for all clonotypes $i \in \mathscr{C}$.

```
for i = 1, ..., η do
    foreach ℓ = 1, ..., M do
        foreach v ∈ B_ℓ do
            Add the (i, v) edge to the network using Bernoulli trial with
             success probability p_ℓ π_iℓ;
        end
    end
end
```

they describe the degree of every node of TCR and VDP type in the network, respectively. Then, using these degree sequences, for each clonotype node i, a sample of size k_i is taken from the set of VDPs, $\mathscr{V} \cup \mathscr{W}$, to determine the VDPs that i will be able to recognise, and therefore the edges that must be added to the recognition network (Newman, 2018, Chapter 12). It is important to note that once a VDP node v has been sampled k'_v times, it must be removed from $\mathscr{V} \cup \mathscr{W}$ for future clonotypes, otherwise the degree sequence would be violated.

In general, when constructing a network using a degree sequence, it is necessary that the sum of the degrees of all nodes in the network is even. However, since the recognition network is bipartite, then it has in fact two degree sequences, one for clonotype nodes (K_C), and one for VDP nodes (K_P). More than this, their sums must be equal for the generation process to be successful (since a difference in the sums would imply there is an edge that starts on one of the sets but cannot end on the other), therefore the sum of the degree sequence for the entire network will always be even, regardless of the evenness of the sum of the degree sequences for clonotype and epitope nodes separately. Thus, this condition on the degree sequence can be forgone.

To simplify the analysis of this type of recognition network, the special case were the degree sequence $K_C = (k_1, \ldots, k_\eta)$ satisfies $k_1 = k_2 = \cdots = k_\eta = K_C^*$

will be considered, where K_C^* denotes the degree of all clonotype nodes; that is, the number of VDPs that every clonotype is able to recognise. An illustrative example of a configuration network is shown in Figure 6.3. For this example four clonotypes and six VDPs are considered. The degree of every clonotype node is $K_C^* = 3$, and the sequence of VDP nodes is $K_P = (1, 2, 3, 3, 2, 1)$. Note that the degree sequence of VDP nodes consists of different values, since the only condition is that its sum equals that of the sequence of clonotype node degrees.

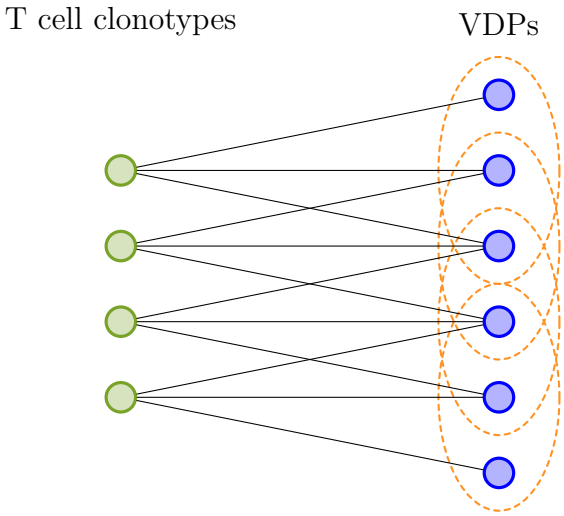


Figure 6.3: Example of a configuration model network. 4 clonotypes, and 6 VDPs are considered. The degree of every clonotype node is $K_C^* = 3$, and the degree sequence of the VDPs is $K_P = (1, 2, 3, 3, 2, 1)$. Note that in this example the degree sequence of the VDPs consists of multiple different values, since the only condition is that its sum must be the same as the sum of K_C.

For a network with η clonotype nodes, and ε VDPs, only values of K_C^* that satisfy

$$\frac{\varepsilon}{\eta} < K_C^* < \varepsilon \tag{6.4}$$

will be considered. Values of K_C^* such that $\eta K_C^* \leq \varepsilon$ are not considered since they can generate networks with no cross-reactivity, which are simply a collection of η independent clonotypes which do not compete with each other for stimulus provided by VDPs. The extreme case $K_C^* = \varepsilon$ is also not considered, since every clonotype having the same recognition profile (of all the VDPs in the network)

can be interpreted as a model with a single clonotype instead of a competition process of η clonotypes. An outline of the network generation algorithm for this model, with general degree sequences K_C and K_P, is given in Algorithm 6.2.

Algorithm 6.2: Configuration model network generation algorithm for a network with η clonotypes.

Input: $K_C \leftarrow$ degree sequence of clonotype nodes,
$K_P \leftarrow$ degree sequence of VDP nodes.

foreach $k_i \in K_C$ **do**
  $S_i \leftarrow$ sample of k_i epitopes from $\mathscr{V} \cup \mathscr{W}$;
  repeat
    $S_i' \leftarrow$ sample of $k_i - |S_i|$ epitopes from $\mathscr{V} \cup \mathscr{W}$;
    $S_i = S_i \cup S_i'$;
    foreach $v \in S_i$ **do**
      if $k_v = \min\{K_P\}$ **then**
        $S_i = S_i \setminus \{v\}$;
        $\mathscr{V} \cup \mathscr{W} = \mathscr{V} \cup \mathscr{W} \setminus \{v\}$;
        $K_P = K_P \setminus \{\min\{K_P\}\}$
      end
    end
  until $|S_i| = k_i$;
  foreach $v \in S_i$ **do**
    Add the (i, v) edge to the network;
  end
end

6.1.3 Preferential attachment networks

The final network generation model to be considered is the preferential attachment model (Barabási & Albert, 1999; Price, 1976, 1965). This is a model of network growth; that is, the network starts with a single clonotype node and all the VDPs nodes, and once the edges for this node are added to the network the remaining clonotype nodes are added one at a time.

Networks constructed using this type of model exhibit *preferential attachment* (Barabási & Albert, 1999; Metzig & Colijn, 2018), also called *cumulative advantage* (Price, 1976; Simon, 1955; Udny Yule, 1925). This behaviour is sometimes described as a "rich-get-richer" or "first mover advantage" effect, and it is caused by the resulting power law distribution of degrees of clonotype nodes. This causes some nodes to be better connected and therefore dominate the dynamics of the system associated to the network. In the case of the recognition network it means that there will always be a subset of clonotypes that will recognise, and therefore receive stimulus, from more VDPs than other clonotypes.

In the original preferential attachment model the probability of an edge being added is proportional to the degree of the node (Barabási & Albert, 1999; Price, 1976). However, in the model proposed here for the recognition network the probability is fixed for the initial sampling for VDPs, or equivalently edges, of every clonotype; that is, for every clonotype the initial sampling for edges behaves like the Erdős-Rényi model. After this, the set of VDPs each clonotype can recognise, $\mathscr{V}_i \cup \mathscr{W}_i$, is compared to that of other clonotypes in the network. If a clonotype j is found such that $(\mathscr{V}_i \cup \mathscr{W}_i) \cap (\mathscr{V}_j \cup \mathscr{W}_j) \neq \emptyset$, then the VDPs in $(\mathscr{V}_j \cup \mathscr{W}_j) \setminus (\mathscr{V}_i \cup \mathscr{W}_i)$ are sampled again with an adjusted probability based on the number of peptides shared after the initial sample.

In order to generate a network with η clonotypes Algorithm 6.3 is used. Initially, the network is considered to have only nodes of VDP type. Then, clonotype nodes are added to the network one at a time in the following way: first, for every clonotype i added to the network an initial sample of VDPs is taken using the Erdős-Rényi model, and the edges between those VDP nodes and clonotype i are added to the network. For clonotypes added to the network after the first one, their recognition profile is compared to that of the other clonotypes in the network. If there is a non-empty intersection with an existing clonotype j, then the cross-reactive recognition probability is calculated for VDPs in $(\mathscr{V}_j \cup \mathscr{W}_j) \setminus (\mathscr{V}_i \cup \mathscr{W}_i)$; that is, VDPs recognised by j but not by i, as follows

$$p_v^*(i,j) := \frac{|(\mathscr{V}_i \cup \mathscr{W}_i) \cap (\mathscr{V}_j \cup \mathscr{W}_j)|}{|\mathscr{V}_i \cup \mathscr{W}_i|} p^* p_v, \tag{6.5}$$

where $0 \leq p^* \leq 1$ is the base cross-reactivity constant of the network, which is used to modulate the strength of preferential attachment observed. That is, if $p^* = 0$,

then $p_v^*(i,j) = 0$ for all i, j, and v and the model becomes equivalent to the Erdős-Rényi model. On the other hand, as $p^* \to 1^-$ the effects of preferential attachment become more apparent, with the overlap for VDP recognition with other clonotypes driving the preferential attachment behaviour. Finally, new edges are added to the network using Bernoulli trials with success probabilities given by the cross-reactive recognition probabilities, $p_v^*(i,j)$, for every v in $(\mathscr{V}_j \cup \mathscr{W}_j) \setminus (\mathscr{V}_i \cup \mathscr{W}_i)$.

Algorithm 6.3: Preferential attachment network generation algorithm for a network with η clonotypes.

Input: $p_v \leftarrow$ probability of a VDP being recognised by a TCR,
$p^* \leftarrow$ base cross-reactivity constant.

for $i = 1, \ldots, \eta$ **do**
  Add edges for clonotype node i in the network using an Erdős-Rényi model with probability p_v;
  if $i = 1$ **then**
    Continue to $i = 2$;
  end
  foreach $j = 1, \ldots, i-1$ *such that* $(\mathscr{V}_i \cup \mathscr{W}_i) \cap (\mathscr{V}_j \cup \mathscr{W}_j) \neq \emptyset$ **do**
    $S_{ij} \leftarrow (\mathscr{V}_j \cup \mathscr{W}_j) \setminus (\mathscr{V}_i \cup \mathscr{W}_i)$;
    foreach $v \in S_{ij}$ **do**
      $p_v^*(i,j) \leftarrow \frac{|(\mathscr{V}_i \cup \mathscr{W}_i) \cap (\mathscr{V}_j \cup \mathscr{W}_j)|}{|\mathscr{V}_i \cup \mathscr{W}_i|} p^* p_v$;
      Add the (i,v) edge to the network using a Bernoulli trial with success probability $p_v^*(i,j)$;
    end
  end
end

An example of this cross-reactivity based preferential attachment of the network is shown in Figure 6.4. In this example clonotype j is already in the network and i is being added to the network. The initial sampling for VDPs that clonotype i can recognise is $\{v_1, v_2\}$, which has a non-empty intersection with the VDPs clonotype j can recognise. Then, using the probability of cross-reactive recognition $p_{v_3}^*(i,j)$, the edge (i, v_3) is considered again for inclusion in the network.

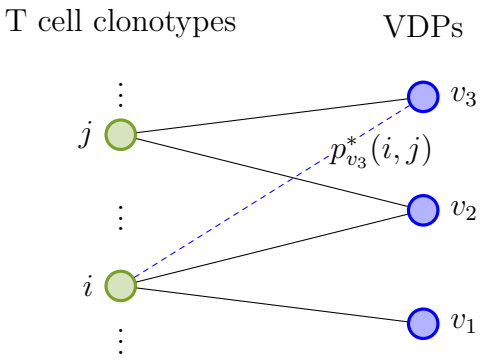


Figure 6.4: Example of the cross-reactivity principle in the preferential attachment model. Clonotype j is already part of the network, and clonotype i is being added. The initial sampling of VDPs for clonotype i is $\{v_1, v_2\}$. Since the intersection with VDPs recognised by j is not empty, the edge (i, v_3), shown in blue, is sampled for with success probability $p^*_{v_3}(i, j)$.

6.2 Clustering coefficient of recognition networks

In this section the clustering coefficient for the random networks presented in Section 6.1 will be calculated in order to measure, and compare, the community structure of the networks. That is, the division of clonotype and VDP nodes into groups within which network connections are dense, but between which there are fewer connections (Newman & Girvan, 2004). In general, for any network the local clustering coefficient of a given node i, $C_3(i)$, describes the average probability that two nodes that are connected to i are also connected to each other (Newman, 2018, Chapter 7). In its traditional definition, this coefficient depends on the density of closed paths of length two, or cycles of size three (see Figure 3.2), which is defined as the number of observed cycles of size three divided by the number of total possible cycles of size three (Watts & Strogatz, 1998). However, cycles of size three are a structure that cannot occur in bipartite networks (Robins & Alexander, 2004). Even more strongly, no cycles of odd size can occur in bipartite networks (Diestel, 2006, Chapter 1). Therefore, a clustering coefficient specific for bipartite networks needs to be considered in order to study the community structure of the TCR-VDP recognition network (Anthonisse, 1971; Freeman, 1977).

The definition of the clustering coefficient for bipartite networks that will be used, $C_{4,uv}(i)$, is the one proposed by Zhang *et al.* (2008), which is based on similar principles to the original $C_3(i)$ coefficient defined by Watts & Strogatz (1998). This definition of the clustering coefficient is suitable for bipartite networks because it depends on the density of butterflies, also called squares, of the network, where a butterfly is the shortest possible cycle that can be constructed in a bipartite network, an example of which is shown in Figure 6.5. Note that, in contrast to $C_3(i)$, this clustering coefficient depends not only on the clonotype node i, but also two peptide nodes u, and v. That is, this local coefficient measures the cliquishness of clonotype nodes of the network around clonotype node i and two given peptide nodes u, and v.

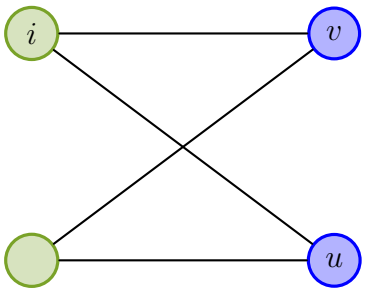


Figure 6.5: Example of a butterfly, or square, in a bipartite network containing i, u, and v. The clustering coefficient is defined in terms of this structure since it is the shortest possible cycle on a bipartite network. Note that a butterfly must always contain two nodes of each type.

In the case of the TCR-VDP recognition network, since the interest of this chapter is TCR cross-reactivity, the clustering coefficient of clonotype nodes will be considered over that of VDP nodes. This means that when calculating $C_{4,uv}(i)$, the nodes u and v will be VDPs, and i will be a clonotype node. Then, for nodes $i \in \mathscr{C}$, and $u, v \in \mathscr{V} \cup \mathscr{W}$ of the recognition network, the clustering coefficient is defined as (Zhang *et al.*, 2008)

$$C_{4,uv}(i) := \frac{q_{iuv}}{k_u + k_v - 2 - q_{iuv}}, \tag{6.6}$$

where q_{iuv} is the number of butterflies that i, u, v are part of, and k_u and k_v the degrees of u and v. Following the original idea of the C_3 clustering

coefficient (Watts & Strogatz, 1998), the global clustering coefficient of the network is defined as the expected value of the local coefficient for all the nodes. That is, the global clustering coefficient of clonotype nodes, denoted by C_4, is given by

$$C_4 \coloneqq \mathop{\mathbb{E}}_{i\in\mathscr{C},u,v\in\mathscr{V}\cup\mathscr{W}} \left[C_{4,uv}(i)\right]. \tag{6.7}$$

6.2.1 Mixed membership stochastic blockmodel

In order to approximate the global clustering coefficient of the stochastic blockmodel, the first step is to find an expression for the expected degree of a VDP node. From the definition of the network it is easy to see that the expected value of the degree of a VDP in block ℓ, $v \in B_\ell$, is

$$\langle k_v \rangle \coloneqq \mathbb{E}\left[k_v\right] = p_\ell \sum_{i\in\mathscr{C}} \pi_{i\ell}, \tag{6.8}$$

where p_ℓ is the recognition probability of the epitope block B_ℓ, and $\pi_{i\ell}$ is the degree of membership of clonotype i to this block.

Then, the local clustering coefficient for clonotype i, and two random VDPs u, and v can be approximated by

$$\hat{C}_{4,uv}(i) \coloneqq \frac{\langle q_{iuv}\rangle}{\langle k_u\rangle + \langle k_v\rangle - 2 - \langle q_{iuv}\rangle}, \tag{6.9}$$

where the number of butterflies that contain $i \in \mathscr{C}$, $u \in B_\ell$, and $v \in B_m$ is approximated, by considering the fact that a butterfly containing $i, j \in \mathscr{C}$, $u \in B_\ell$, and $v \in B_m$ is present in the network with probability $(\pi_{i\ell}p_u)(\pi_{im}p_v)(\pi_{j\ell}p_u)(\pi_{jm}p_v)$, as follows

$$\begin{aligned}
\langle q_{iuv}\rangle &\coloneqq \mathbb{E}\left[q_{iuv}\right] \\
&= \sum_{j\in\mathscr{C}\setminus\{i\}} (\pi_{i\ell}p_u)(\pi_{im}p_v)(\pi_{j\ell}p_u)(\pi_{jm}p_v) \\
&= (\pi_{i\ell}p_u)(\pi_{im}p_v) \sum_{j\in\mathscr{C}\setminus\{i\}} (\pi_{j\ell}p_u)(\pi_{jm}p_v) \\
&= (p_u p_v)^2 \pi_{i\ell}\pi_{im} \sum_{j\in\mathscr{C}\setminus\{i\}} \pi_{j\ell}\pi_{jm}.
\end{aligned} \tag{6.10}$$

Thus, replacing Eq. (6.10) in Eq. (6.9) the following expression for approximating the local clustering coefficient of a mixed membership stochastic blockmodel can be found

$$\hat{C}_{4,uv}(i) = \frac{(p_u p_v)^2 (\pi_{i\ell}\pi_{im}) \sum_{j\in\mathscr{C}\setminus\{i\}} \pi_{j\ell}\pi_{jm}}{\langle k_u\rangle + \langle k_v\rangle - 2 - (p_u p_v)^2 (\pi_{i\ell}\pi_{im}) \sum_{j\in\mathscr{C}\setminus\{i\}} \pi_{j\ell}\pi_{jm}}. \tag{6.11}$$

Then, the global clustering coefficient for a mixed membership stochastic blockmodel network with η clonotypes, and $\varepsilon = |\mathscr{V}\cup\mathscr{W}|$ VDPs is given by

$$C_4 = \frac{1}{\eta\binom{\varepsilon}{2}} \sum_{i\in\mathscr{C}} \sum_{u,v\in\mathscr{V}\cup\mathscr{W}} C_{4,uv}(i). \tag{6.12}$$

However, as mentioned in Section 6.1.1, only the extreme case that recovers the Erdős-Rényi model will be considered. In this case there is only one block of VDPs, $M = 1$, with recognition probability $p_{\overline{v}}$ for all $v \in B_1 = \mathscr{V}\cup\mathscr{W}$, and all clonotypes have the membership vector $\boldsymbol{\pi} = (1)$. Then, assuming again that the network consists of η clonotypes the degree of VDP v in Eq. (6.8) simplifies to

$$\langle k_v\rangle = p_v\eta. \tag{6.13}$$

Now, replacing this in Eq. (6.11) and simplifying results in the local clustering coefficient for the Erdős-Rényi model

$$\hat{C}_{4,uv}(i) = \frac{(p_v)^4(\eta-1)}{2p_v\eta - 2 - (p_v)^4(\eta-1)}. \tag{6.14}$$

Furthermore, since every clonotype and VDP behaves identically in this model, the global clustering coefficient matches the local one exactly; that is, for the Erdős-Rényi model the following equality is satisfied for any $i \in \mathscr{C}$ and $u, v \in \mathscr{V}\cup\mathscr{W}$

$$C_4 = \hat{C}_{4,uv}(i). \tag{6.15}$$

6.2.2 Configuration model

For the configuration model, the degree of the VDPs does not need to be approximated by its expected value, since the degree sequence is known, thus, the exact degree of every clonotype and VDP is already known. However, in order to

find the global clustering coefficient it is necessary to know the probability that a clonotype will be able to recognise a peptide. It is easy to see that for a VDP v of degree k_v this probability is given by

$$p_v = \frac{k_v}{\eta}, \tag{6.16}$$

where, as before, η is the number of clonotypes in the network.

Then, similarly to Eq. (6.10), for a clonotype i and two VDPs u, and v, the expected number of butterflies is given by

$$\langle q_{iuv} \rangle \coloneqq \mathbb{E}\left[q_{iuv}\right] = (p_u p_v)^2(\eta - 1) = \frac{k_u k_v}{\eta^2} \cdot \frac{k_u k_v}{\eta^2} \cdot (\eta - 1) = \frac{(k_u k_v)^2 (\eta - 1)}{\eta^4}, \tag{6.17}$$

since a butterfly containing $i, j \in \mathscr{C}$, $u, v \in \mathscr{V} \cup \mathscr{W}$ will be present in the network with probability $(p_u p_v)^2$. Now, replacing Eq. (6.17) in Eq. (6.6) the following expression for the local clustering coefficient of the configuration model is obtained

$$\begin{aligned}
\hat{C}_{4,uv}(i) &= \frac{\dfrac{(k_u k_v)^2 (\eta - 1)}{\eta^4}}{k_u + k_v - 2 - \dfrac{(k_u k_v)^2 (\eta - 1)}{\eta^4}} \\
&= \frac{\dfrac{(k_u k_v)^2 (\eta - 1)}{\eta^4}}{\dfrac{(k_u + k_v - 2)\eta^4 - (k_u k_v)^2 (\eta - 1)}{\eta^4}} \\
&= \frac{(k_u k_v)^2 (\eta - 1)}{(k_u + k_v - 2)\eta^4 - (k_u k_v)^2 (\eta - 1)}.
\end{aligned} \tag{6.18}$$

Finally, using the expression for the local clustering coefficient, Eq. (6.18), it is easy to see that the global clustering coefficient for the configuration model is given by

$$C_4 = \frac{1}{\binom{\varepsilon}{2}} \sum_{u,v \in \mathscr{V} \cup \mathscr{W}} \frac{(k_u k_v)^2 (\eta - 1)}{(k_u + k_v - 2)\eta^4 - (k_u k_v)^2 (\eta - 1)}, \tag{6.19}$$

where $\varepsilon = |\mathscr{V} \cup \mathscr{W}|$.

6.2.3 Preferential attachment

Similarly to the other models, the first step is to find the expected value of the degree of a VDP v. In this case, since the network is constructed by adding one clonotype node at a time, it is helpful to consider how the degree of a VDP node changes when a new clonotype node is added to the network. For this it is also necessary to first understand how the probability of VDP recognition changes when a new clonotype is added to the network. Another useful by-product of this network generation algorithm is that the clonotype nodes can be naturally enumerated using the order in which they are added to the network. This enumeration will prove useful in this section for the calculation of the clustering coefficient.

Define $\tilde{p}_v(i)$ to be the probability that the i-th clonotype added to the network will recognise VDP v. It is easy to see that in the case $i = 1$, the value of $\tilde{p}_v(i)$ is the base probability of recognition of v, namely p_v, since for a single clonotype the preferential attachment algorithm behaves like the Erdős-Rényi model. For $i > 1$, on the other hand, the value of $\tilde{p}_v(i)$ will be of the form $p_v + (1 - p_v)\rho$, for some probability ρ representing the fact that, after initially being unable to recognise v with probability $1 - p_v$, the preferential attachment algorithm will allow for the edge (i, v) to be considered for the recognition network again with probability ρ.

In order to find an expression for ρ, first consider some initial values of $\tilde{p}_v(i)$: In the case $i = 2$ there is only one other clonotype in the network. Thus, the probability that v is recognised after the preferential attachment algorithm, given that it was not recognised during the initial sampling, is $(1 - p_v)\tilde{p}_v(1)p_v^*(2, 1)$ (see Algorithm 6.3 lines: 8–11), and thus $\tilde{p}_v(2)$ is given by

$$\tilde{p}_v(2) = p_v + (1 - p_v)\tilde{p}_v(1)p_v^*(2, 1). \tag{6.20}$$

Now, when the third clonotype is added to the network, $i = 3$, there are two clonotypes in the network which provide two opportunities for the recognition probability of v. The first clonotype increases this probability by $(1 - p_v)\tilde{p}_v(1)p_v^*(3, 1)$, and the second one increases it by $(1 - p_v)\tilde{p}_v(2)p_v^*(3, 2)$. Thus, the value of $\tilde{p}_v(3)$ is given by

$$\tilde{p}_v(3) = p_v + (1 - p_v)\big(\tilde{p}_v(1)p_v^*(3, 1) + \tilde{p}_v(2)p_v^*(3, 2)\big). \tag{6.21}$$

From Eq. (6.20) and Eq. (6.21), it is easy to see that given an initial value for $\tilde{p}_v(1)$, a general recursive expression can be derived as follows

$$\begin{aligned}\tilde{p}_v(1) &= p_v, \\ \tilde{p}_v(i) &= p_v + (1-p_v)\sum_{j=1}^{i-1}\tilde{p}_v(j)p_v^*(i,j) \text{ for } i = 2,\ldots,\eta.\end{aligned} \tag{6.22}$$

Note, however, that this expression for the recognition probability depends on $p_v^*(i,j)$, which requires knowledge of the exact $\mathscr{V}_i \cup \mathscr{W}_i$ sets for the clonotypes considered. For a general preferential attachment network, like the one considered for the calculation of the clustering coefficient, these sets are not explicitly known. For this reason, the expected value of $p_v^*(i,j)$ will be calculated.

It is easy to see that the possible values of the overlap between the recognition profile of i and j, $\frac{|(\mathscr{V}_i\cup\mathscr{W}_i)\cap(\mathscr{V}_j\cup\mathscr{W}_j)|}{|\mathscr{V}_i\cup\mathscr{W}_i|}$, follow a binomial distribution. Let $\ell = |(\mathscr{V}_i \cup \mathscr{W}_i) \cap (\mathscr{V}_j \cup \mathscr{W}_j)|$, and $\varepsilon_i = |\mathscr{V}_i \cup \mathscr{W}_i|$. Then, the random variable describing this fraction of VDP overlap, $X = \left\{\frac{\ell}{\varepsilon_i}\right\}$, follows a binomial distribution with probability $p_v\tilde{p}_v(j)$, and ε_i trials; that is, $X \sim \text{Binomial}\left(p_v\tilde{p}_v(j), \varepsilon_i\right)$. Thus

$$\mathbb{E}\left[X\right] = \sum_{\ell=1}^{\varepsilon_i}\binom{\varepsilon_i}{\ell}\left(p_v\tilde{p}_v(j)\right)^{\ell}\left(1-p_v\tilde{p}_v(j)\right)^{\varepsilon_i-\ell}\frac{\ell}{\varepsilon_i}, \tag{6.23}$$

and from Eq. (6.5), given the fact that p^*, and p_v are constant, it is clear that

$$\langle p_v^*(i,j)\rangle \coloneqq \mathbb{E}\left[p_v^*(i,j)\right] = \mathbb{E}\left[X\right]p^*p_v, \tag{6.24}$$

which can be replaced in Eq. (6.22).

There are two important things to note about this approximation. First, this approximation does not depend on clonotype i. This is due to the fact that when clonotype i is added to the network, its initial recognition probability for a VDP v before the preferential attachment algorithm is used is given by p_v. Then, for every clonotype i added after j, the value of $\langle p_v^*(i,j)\rangle$ will be the same; that is, $\langle p_v^*(i,j)\rangle = \langle p_v^*(k,j)\rangle$ for all $i,k > j$. Thus, $\langle p_v^*(\cdot,j)\rangle$ will be used to denote the cross-reactivity probability for any clonotype $i > j$. Second, given that $\tilde{p}_v(i)$ and $\langle p_v^*(i,j)\rangle$ are defined in terms of each other, and $\tilde{p}_v(i)$ is defined recursively, it may appear that these approximations are ill-defined. However, upon closer inspection

it can be seen that since the initial value $\tilde{p}_v(1)$ is known, the values of Eq. (6.22) and Eq. (6.24) are indeed well-defined and can be found sequentially as follows

$$\tilde{p}_v(1) \to \langle p_v^*(\cdot,1)\rangle \to \tilde{p}_v(2) \to \langle p_v^*(\cdot,2)\rangle \to \tilde{p}_v(3) \to \langle p_v^*(\cdot,3)\rangle \to \cdots . \tag{6.25}$$

Now, for a given VDP v, let $\langle k_v\rangle_\eta$ denote the expected value of the degree of v on a network with η clonotypes. When the network has a single clonotype node; that is, when $\eta = 1$, the expected value of k_v is clearly $\langle k_v\rangle_1 = p_v$. Then, once a second clonotype is added to the network the expected value increases by the probability that the new clonotype will recognise it, which is given by Eq. (6.22), and the expected value of the degree becomes

$$\langle k_v\rangle_2 = 2p_v + (1-p_v)\tilde{p}_v(1)p_v^*(2,1) \approx 2p_v + (1-p_v)\tilde{p}_v(1)\langle p_v^*(\cdot,1)\rangle, \tag{6.26}$$

where the first term counts the edges added to the network as a result of the initial sampling, and the second term the number of edges added after the preferential attachment algorithm is used. When a third clonotype is added to the network, it is easy to see that similarly like with the second clonotype, using Eq. (6.22) the expected degree of v becomes

$$\begin{aligned}\langle k_v\rangle_3 &= 3p_v + (1-p_v)\tilde{p}_v(1)\big(p_v^*(2,1) + p_v^*(3,1)\big) + (1-p_v)\tilde{p}_v(2)p_v^*(3,2)\\ &\approx 3p_v + (1-p_v)\big(2\tilde{p}_v(1)\langle p_v^*(\cdot,1)\rangle + \tilde{p}_v(2)\langle p_v^*(\cdot,2)\rangle\big)\end{aligned} \tag{6.27}$$

Comparing this to Eq. (6.26), it is easy to see a pattern emerge. The first term comes from the fact that all clonotypes can initially recognise VDP v with probability p_v, and the second term is the increase in the degree due to the preferential attachment algorithm comparing the recognition profiles of the second and third clonotypes with the first, as well as the third clonotype with the second.

In general, when clonotype $j = i+1$ is added to the network it is checked for cross-reactivity with all clonotypes up to i, and this increases expected value of k_v by $\sum_{\ell=1}^{i}(1-p_v)\tilde{p}_v(\ell)p_v^*(j,\ell)$. Then, for a recognition network with η clonotypes the expected value of the degree of VDP v is given by

$$\begin{aligned}\langle k_v\rangle_\eta &= p_v\eta + (1-p_v)\sum_{i=1}^{\eta-1}\tilde{p}_v(i)\sum_{j=i+1}^{\eta}p_v^*(j,i)\\ &\approx p_v\eta + (1-p_v)\sum_{i=1}^{\eta-1}(\eta-j)\tilde{p}_v(i)\langle p_v^*(\cdot,i)\rangle.\end{aligned} \tag{6.28}$$

Now, in order to calculate the clustering coefficient the number of butterflies is required first. Similarly to Eq. (6.10), and using Eq. (6.22) the number of expected butterflies can be written as

$$\langle q_{iuv} \rangle = \tilde{p}_u(i)\tilde{p}_v(i) \sum_{j \in \mathscr{C} \setminus \{i\}} \tilde{p}_u(j)\tilde{p}_v(j). \tag{6.29}$$

Finally, substituting Eq. (6.29) into Eq. (6.6) the following expression for the approximation of the local clustering coefficient is obtained

$$\hat{C}_{4,uv}(i) = \frac{\tilde{p}_u(i)\tilde{p}_v(i) \sum_{j \in \mathscr{C} \setminus \{i\}} \tilde{p}_u(j)\tilde{p}_v(j)}{\langle k_u \rangle_\eta + \langle k_v \rangle_\eta - 2 - \tilde{p}_u(i)\tilde{p}_v(i) \sum_{j \in \mathscr{C} \setminus \{i\}} \tilde{p}_u(j)\tilde{p}_v(j)}, \tag{6.30}$$

which can be used to calculate the global clustering coefficient as follows

$$C_4 = \frac{1}{\eta \binom{\varepsilon}{2}} \sum_{i \in \mathscr{C}} \sum_{u,v \in \mathscr{V} \cup \mathscr{W}} \frac{\tilde{p}_u(i)\tilde{p}_v(i) \sum_{j \in \mathscr{C} \setminus \{i\}} \tilde{p}_u(j)\tilde{p}_v(j)}{\langle k_u \rangle_\eta + \langle k_v \rangle_\eta - 2 - \tilde{p}_u(i)\tilde{p}_v(i) \sum_{j \in \mathscr{C} \setminus \{i\}} \tilde{p}_u(j)\tilde{p}_v(j)}. \tag{6.31}$$

In the special case where the recognition probability of all VDPs is assumed to be the same, the above equation simplifies to

$$C_4 = \frac{1}{\eta} \sum_{i \in \mathscr{C}} \frac{\tilde{p}_v(i)^2 \sum_{j \in \mathscr{C} \setminus \{i\}} \tilde{p}_v(j)^2}{2\langle k_v \rangle_\eta - 2 - \tilde{p}_v(i)^2 \sum_{j \in \mathscr{C} \setminus \{i\}} \tilde{p}_v(j)^2}, \tag{6.32}$$

for any $v \in \mathscr{V} \cup \mathscr{W}$.

6.2.4 Comparison of the recognition network models

Using the results of Sections 6.2.1, 6.2.2, and 6.2.3 it is possible to compare the network generation algorithms using their global clustering coefficient (see Figure 6.6). Note that, while the mixed membership stochastic blockmodel will not be considered in further sections, its clustering coefficient was still calculated and compared to the other models here.

Figure 6.6 shows the clustering coefficient as a function of p_v for the four types of random networks studied in this chapter with $\eta = 8$, and $|\mathscr{V} \cup \mathscr{W}| = 20$. In this Figure the preferential attachment model was assumed to have the same basic recognition probability for all epitopes, the configuration model was considered to have a constant degree sequence, and the stochastic blockmodel consists of two

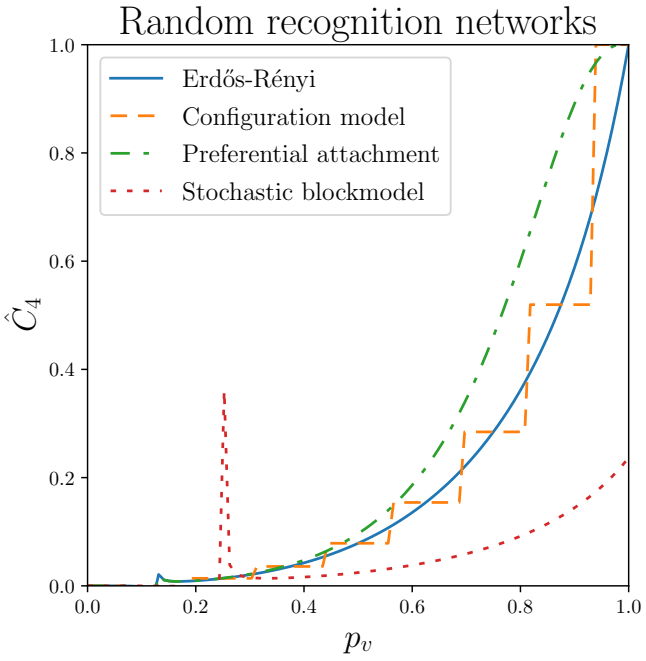


Figure 6.6: Global clustering coefficient, C_4, of the random networks described in Section 6.1 approximated using the method described in Section 6.2, with $\eta = 8$, and $|\mathscr{V} \cup \mathscr{W}| = 20$. In general the preferential attachment model generates networks with clustering coefficient greater than the other models, which is not unexpected given the fact that it is the only model that allows for resampling of edges. The Erdős-Rényi and configuration models show a similar behaviour, but since for the configuration model approximates the probability by changing the degree sequence it behaves less smoothly than the other models. The stochastic blockmodel generates networks with the lowest clustering coefficient overall.

blocks of epitopes. It is important to note that in the case of the configuration model the plot is not smooth because in order to calculate the clustering coefficient p_v has to be approximated using the closest integer solution of Eq. (6.16). This means that when two values of p_v are close enough to each other, the integer solution of Eq. (6.16) is the same for both of them. Another important aspect to note before continuing is the early peak of the stochastic blockmodel, due to an underestimation of the number of butterflies present in the network. This happens because, when considered as a function of $\langle q_{iuv} \rangle$, Eq. (6.12) is convex, and by Jensen's inequality (McShane, 1937), it can be shown that this causes the

expected value of C_4 to be underestimated.

Now, in order to compare the mixed membership and the Erdős-Rényi models, consider the mixed membership model with all VDPs in a single block; that is, all VDPs have the same recognition probability. In this case it is easy to see that the Erdős-Rényi model will have a greater clustering coefficient than the mixed membership model, since $(\pi_{iv}\pi_{iu})\sum_{j\in\mathscr{C}\setminus\{i\}}\pi_{jv}\pi_{ju}$ is less than or equal to $\eta - 1$. This difference between the models is to be expected, since the introduction of the membership vector decreases the number of edges added to the network in comparison to an Erdős-Rényi model by limiting the number of clonotypes that can recognise a VDP to only those with specific block membership values greater than zero, and decreasing the probability of recognition for those with a membership value strictly less than one, as evidenced in Figure 6.6. This lowers the local clustering coefficient of the network, which in turn lowers its global clustering coefficient. In general, when both the membership vector and multiple blocks are considered, the clustering coefficient will be less than that of an Erdős-Rényi network. This is because the effects of segregating the nodes using a membership vector greatly limits their capacity to form cliques and generate highly clustered networks.

The Erdős-Rényi and the preferential attachment models are comparable in terms of their clustering coefficients for low values of p_v, however the preferential attachment model has (unsurprisingly) a greater clustering coefficient as the value of p_v increases. This is easy to see by considering the case where $p^* = 0$, recovering the behaviour of an Erdős-Rényi network, and then gradually increasing the value of p^* (see Figure 6.7). To this end, note that if p^* is 0, and p_v is the same for all v in a preferential attachment network, then an Erdős-Rényi network is generated. By increasing the value of the cross-reactivity probability, the value of $\langle k_v \rangle$ increases, as well as the number of butterflies, which causes an increase in the clustering coefficient. This is expected since the second sampling for edges in the network facilitates the creation of cliques which are measured by the clustering coefficient.

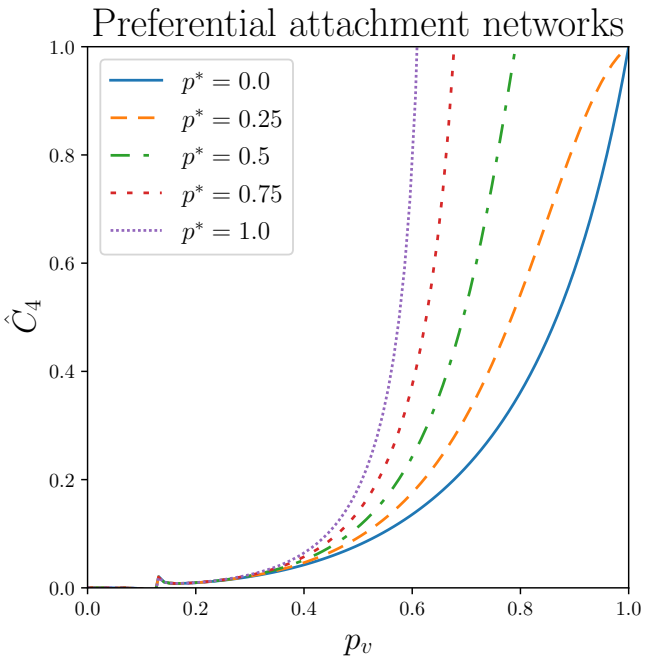


Figure 6.7: Global clustering coefficient, C_4, of preferential attachment random networks, Section 6.1.3, approximated using the method described in Section 6.2, with $\eta = 8$, and $|\mathscr{V} \cup \mathscr{W}| = 20$, for different values of p^*. It is easy to see that increasing the basic value of cross-reactivity causes an increase in the clustering coefficient of the network.

Finally, the clustering coefficient of the configuration model is qualitatively similar to the Erdős-Rényi and preferential attachment models if the degree sequence of clonotype nodes is considered to be $K_C = (K_C^*, \ldots, K_C^*)$. However, changing this degree sequence can have strong effects on the clustering coefficient. For example, consider the degree sequence $K_C = (K_C^*, \ldots, K_C^*, \mathcal{K}_C^*, K_C^*, \ldots, K_C^*)$, with $\mathcal{K}_C^* \gg K_C^*$. In this case most of the VDP nodes will be connected to the clonotype node with degree $\mathcal{K}_C^*$, and connections to other nodes will be less frequent. This causes the local clustering coefficients of clonotypes that are connected to this VDP node to drop greatly, since the number of total possible butterflies increases, but the number butterflies present in the network drops due to the fact that most edges are connected to a single node. A summary of the approximations of the clustering coefficients is presented in Table 6.1.

Comparison of global clustering coefficients for the random recognition networks	
Network Model	$\hat{C}_4$
Mixed membership stochastic blockmodel	$\frac{1}{\eta\binom{\varepsilon}{2}} \sum_{i\in\mathscr{C}} \sum_{u,v\in\mathscr{V}\cup\mathscr{W}} \frac{(p_u p_v)^2 (\pi_{i\ell}\pi_{im}) \sum_{j\in\mathscr{C}\setminus\{i\}} \pi_{j\ell}\pi_{jm}}{\langle k_u\rangle + \langle k_v\rangle - 2 - (p_u p_v)^2 \pi_{i\ell}\pi_{im} \sum_{j\in\mathscr{C}\setminus\{i\}} \pi_{j\ell}\pi_{jm}}$
Erdős-Rényi model	$\frac{(p_v)^4(\eta-1)}{2p_v\eta - 2 - (p_v)^4(\eta-1)}$
Configuration model	$\frac{1}{\binom{\varepsilon}{2}} \sum_{u,v\in\mathscr{V}\cup\mathscr{W}} \frac{(k_u k_v)^2 (\eta-1)}{(k_u + k_v - 2)\eta^4 - (k_u k_v)^2 (\eta-1)}$
Preferential attachment	$\frac{1}{\eta\binom{\varepsilon}{2}} \sum_{i\in\mathscr{C}} \sum_{u,v\in\mathscr{V}\cup\mathscr{W}} \frac{\tilde{p}_u(i)\tilde{p}_v(i) \sum_{j\in\mathscr{C}\setminus\{i\}} \tilde{p}_u(j)\tilde{p}_v(j)}{\langle k_u\rangle_\eta + \langle k_v\rangle_\eta - 2 - \tilde{p}_u(i)\tilde{p}_v(i) \sum_{j\in\mathscr{C}\setminus\{i\}} \tilde{p}_u(j)\tilde{p}_v(j)}$

Table 6.1: Global clustering coefficient for the network generation algorithms calculated in Sections 6.2.1, 6.2.2, and 6.2.3. The number of VDPs in the network is $\varepsilon = |\mathscr{V} \cup \mathscr{W}|$. The mixed membership stochastic blockmodel is in general less clustered than the Erdős-Rényi model due to the limitations introduced by the membership vector. The preferential attachment and the Erdős-Rényi models are comparable in terms of their clustering coefficient, but former has an increased coefficient due to its higher propensity for the formation of cliques.

6.3 Dynamics of cross-reactive T cell responses

Using the multi-variate competition process defined in Chapter 5, and the code presented in Appendix D, the Gillespie algorithm was used to simulate the population dynamics of a finite set of clonotypes during homeostasis, and during two perturbations due to heterologous viral infection (Gillespie, 1976, 1977). For the purposes of illustration and exposition, a set of three different clonotypes ($i = 1, 2, 3$) and a set of 18 pMHC complexes (nine for each of the viral challenges) are considered. Thus, the bipartite network has three clonotype nodes and 18 VDP nodes.

The three different network generation algorithms, defined in Section 6.1, are used to create the recognition networks using the code presented in Appendix D.1. The code in Appendix D.2 uses the Gillespie algorithm to simulate the T cell populations dynamics for a period of one year. Each infection is considered to last one week, and a six month period between infections is assumed. At time $t = 0$, each clonotype consists of five naive T cells. Thus, there are no memory or effector cells of any TCR specificity to start with.

The left panel of Figure 6.8 shows the bipartite recognition network used for the unfocussed cross-reactivity hypothesis generated using the Erdős-Rényi model defined in Section 6.1.1. For this example the recognition probability of the peptides was chosen to be $p_v = 8/18$, resulting in a clustering coefficient $C_4 \approx 0.132$. The middle and right panels show the mean of 10^4 stochastic simulations of a heterologous infection. From this panel, it can be noted that during the response to the initial viral challenge clonotype 2 expanded to a greater degree than the other clonotypes, and therefore was the dominant clonotype of the response. The behaviour of each separate phenotype of T cells for each clonotype is shown in the plots on the right panel; that is, the dynamics of naive, effector and memory populations of cells are plotted on the same timescale as the centre panel.

From the results of the stochastic realisations it can be seen that during the initial challenge, the population of naive cells was not completely depleted while

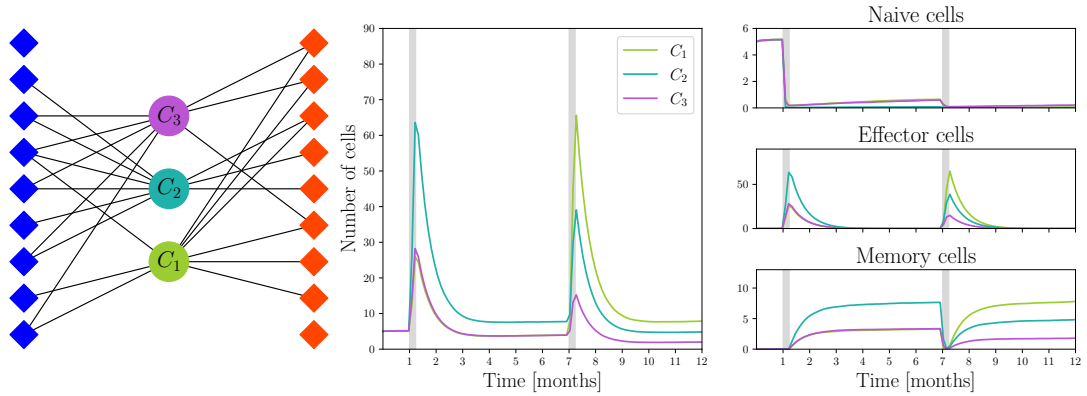


Figure 6.8: Average over 10^4 realisations of an immune response to two viral challenges with an unfocussed recognition network and dynamics defined in Chapter 5. The network was constructed using the Erdős-Rényi network generation algorithm described in Section 6.1.1 with probability of recognition $p_v = 8/18$ for all VDPs, and three clonotypes C_1, C_2, and C_3. The clustering coefficient of the network is $C_4 \approx 0.132$. Blue diamonds represent VDPs present during the first viral challenge and red diamonds represent those present during the second challenge. The grey bands represent the periods of infection. Note that not every VDP was recognised by the clonotypes, since the samples were drawn randomly, and there is always a probability $1 - p_v$ of no recognition between a clonotype and a VDP v. Figure generated using the code in Appendix D.

they were differentiating into effector cells, allowing the naive pool to recover to homeostatic levels after the first infection was cleared. Focussing on the effector cell compartment, it can be seen that during both infections it was indeed the population of effector phenotype cells that expanded and subsequently contracted once the infection is cleared, as expected from the definition of the multi-variate competition process. Finally, note that memory cells were generated after the first infection was cleared, and once the second infection ended, there was an increase in the population of memory cells, which was then homeostatically maintained.

In Figure 6.9 the same set of plots presented in the previous figure are shown, this time for the configuration model presented in Section 6.1.2, in order to show the effects of the focussed cross-reactivity hypothesis. The degree sequence considered was $(8, 8, 8)$ for the clonotype nodes, and the degree sequence of VDP nodes was

not specified. This degree sequence was chosen because the clustering coefficient of the resulting network is $C_4 \approx 0.131$, which is similar to that of the network in Figure 6.8. Finally, Figure 6.10 is another summary plot of the stochastic simulations for a focussed network generated using the preferential attachment model from Section 6.1.3. Similarly to the Erdős-Rényi model, the recognition probability was chosen to be $p_v = 8/18$. The cross-reactivity probability was chosen to be $p^* = 1$ in this case to be able to show the effects of preferential attachment on a network this small. The clustering coefficient of the resulting network is $C_4 \approx 0.135$. In both instances, a similar behaviour to that of the unfocussed hypothesis was observed on the simulations. That is, T cell responses that are dominated by a subset of the clonotypes, with the immunodominant clonotypes not necessarily being the same for both infections.

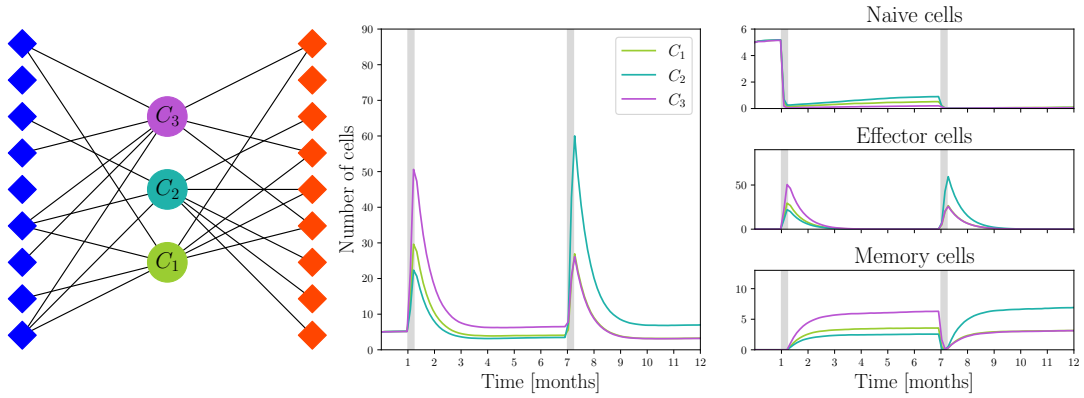


Figure 6.9: Average over 10^4 realisations of an immune response to two viral challenges with a configuration model recognition network and dynamics defined in Chapter 5. The network was constructed using the network generation algorithm described in Section 6.1.2 with degree $k = 8$ for all clonotypes, and three clonotypes C_1, C_2, and C_3. The clustering coefficient of the network is $C_4 \approx 0.131$. Blue diamonds represent VDPs present during the first viral challenge and red diamonds represent those present during the second challenge. Note that not every VDP is recognised by the clonotypes. Figure generated using the code in Appendix D.

From these preliminary and exploratory results, it can be seen that the three different hypotheses of TCR-VDP recognition considered in this chapter can lead to immune dynamics and clonal behaviour that are immunologically plausible

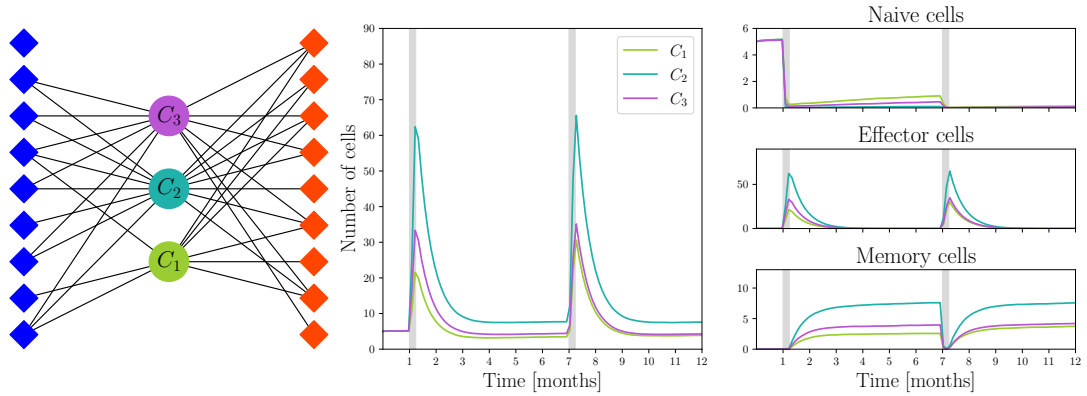


Figure 6.10: Average over 10^4 realisations of an immune response to two viral challenges with a preferential attachment recognition network and dynamics defined in Chapter 5. The network was constructed using the network generation algorithm described in Section 6.1.3 with probability of success $p = 8/18$, and cross-reactivity probability $p^* = 1$ for three clonotypes C_1, C_2, and C_3. The clustering coefficient of the network is $C_4 \approx 0.135$. Blue diamonds represent VDPs present during the first viral challenge and red diamonds represent those present during the second challenge. Note that not every VDP was recognised by the clonotypes. Figure generated using the code in Appendix D.

and realistic. For instance, in a given response to a viral infection, there is an immunodominant clonotype expanding to a greater degree than the rest of the clonotypes, and thus, driving the dynamics of the immune response (Akram & Inman, 2012; Yewdell & Bennink, 1999). This behaviour of immunodominance was observed on all three of the random networks generated, suggesting that either of the three network generation algorithms presented in Section 6.1 can plausibly describe a TCR-VDP recognition network found in nature. While a more in-depth study of the network models presented here was not possible due to time limitations, the methods presented can be used as a tool to link the network structure of TCR-VDP recognition networks with clonotype population dynamics. This tool can then be used to comprehensively explore the impact of different network types on immune response dynamics.

6.4 Discussion

Given the importance of cross-reactive immune responses in the context of memory generation and vaccination (Elong Ngono & Shresta, 2019; Mateus *et al.*, 2021; Moris *et al.*, 2011; Webster & Askonas, 1980), it is essential to gain a better understanding of the mechanisms, and underlying VDP recognition structure of cross-reactive immune responses. In this chapter, based on current immunological evidence (Bradley & Thomas, 2019; DeWitt III *et al.*, 2018; Duan *et al.*, 2015; Gaevert *et al.*, 2021; Kanduc, 2012; Souquette & Thomas, 2018), mathematically generated hypotheses of the mechanisms by which cross-reactivity emerges, and its underlying structure were proposed and studied from the point of view of a TCR-VDP bipartite recognition network. Three such hypotheses were proposed: an Erdős-Rényi model for unfocussed cross-reactivity, and a configuration model and a preferential attachment model for focussed cross-reactivity. These cross-reactive recognition networks were used to simulate an immune response to a heterologous infection, and all of them generated immune responses that show the qualitative traits of observed immune responses (Bevan, 2004; Selin *et al.*, 2006). Some of these traits are derived from the multi-variate competition process used to model the population dynamics, such as the expansion from the naive and memory compartments to the effector compartment during infection, and the contraction into memory that follows, and others arise from the random recognition network used to define the competition for stimulus, such as the immunodominant hierarchies observed during the expansion to the effector compartment.

Network representations, like those described in this chapter, can be used to define distances in the space of TCRs by defining a distance between TCR nodes in the network. For example, a distance can be defined based on the number of shared VDP nodes between clonotype nodes; that is, a distance measuring the similarity of recognition profiles of TCRs. Conversely, a distance between peptides based on which TCRs can recognise them can be also be defined (Hadfield *et al.*, 2018; Kitsak *et al.*, 2017; Lambiotte & Ausloos, 2005; Robins & Alexander, 2004; Schattgen *et al.*, 2020). However, these definitions of distance rely on having previous and precise knowledge of which peptides will be recognised by a given TCR, since the recognition network needs to be constructed before the distance

between nodes can be calculated. This is of course, not information that can be known a priori for most TCRs and pMHCs. For this reason, previously defined distances between TCRs are guided by structural information on their pMHC binding properties. For example, TCRdist is based on the structure of the CDR1, CDR2 and CDR3 loops (Dash *et al.*, 2017; Mayer-Blackwell *et al.*, 2020).

One method commonly used to define distances in bipartite networks is to do a one mode projection, and use the definition of the distance between two nodes in a network to define a distance between nodes of the same type on the pre-image of this projected network (Battiston & Catanzaro, 2004; Blond *et al.*, 2005; Morris *et al.*, 2005). That is, the network is projected onto one of the node types, and a distance is defined on this resulting network (Dankelmann *et al.*, 1996; Gutman & Yeh, 1995; Li & Song, 2014). However, this type of projection is prone to loss of information about the network structure (Tumminello *et al.*, 2011). More recently, advances have been made on projections with minimal loss of structural and community information (Dianati, 2016; Saracco *et al.*, 2017). Using this type of information preserving projections may allow for the study of distances, like TCRdist, from a network theoretical point of view as presented in this chapter. Another possible avenue of research would be to study the structure of the resulting family of bipartite networks generated by the pre-image of a one mode projection. This type of analysis could ostensibly be done using biological data to construct the mono-partite network used to calculate the pre-image.

The clustering coefficient was used to study the cliquishness of the networks proposed in this chapter. This coefficient can be interpreted as a measure of how connected to each other the nodes that are adjacent to a given node are (Newman, 2018, Chapter 7). Using this measure it was shown that both the focussed and unfocussed cross-reactivity hypotheses generate networks with comparable clustering coefficients. One other useful interpretation of this coefficient, is the betweenness centrality of the network (Borgatti & Everett, 2000; Burt, 2012). The betweenness centrality of a node measures how much "power" a node has over the others based on how many paths pass through it (Anthonisse, 1971; Freeman, 1977). While the clustering coefficient is not directly related to betweenness centrality, it was shown by Burt (2012) that they are strongly correlated, therefore the study of one can be extended to the study of the other. One particular

application of betweenness centrality is to find "structural holes" in a network, which are the missing edges that give a node more "power". An explicit example in the context of TCR-VDP recognition networks would be a TCR whose recognition profile includes VDPs which can only be recognised by said TCR. In this case, if this clonotype were to be removed from the network, this would mean that there are VDPs which cannot be recognised by any of the other clonotypes in the network, negatively impacting immune responses to pathogens that present those VDPs.

The definition of the clustering for bipartite networks used was the one proposed by Zhang *et al.* (2008). This particular definition was chosen because, in it the density of butterflies in the network considers only butterflies that are either present in the network, or can be added to the network by the inclusion of new edges. However, this is not the only possible definition of a clustering coefficient in bipartite networks. A similar formulation is proposed by Lind *et al.* (2005), with the key difference that it also considers butterflies created by the collapsing of two nodes. In the case of the clustering coefficient of clonotype nodes, as considered in Section 6.2, the two collapsed nodes are clonotype nodes. An example of this type of butterfly is show in Figure 6.11. Using this definition of butterflies the local clustering coefficient is given by

$$C_{4,uv}(i) \coloneqq \frac{q_{iuv}}{(k_u - 1 - q_{iuv})(k_v - 1 - q_{iuv}) + q_{iuv}}. \tag{6.33}$$

This definition of the clustering coefficient allows for a different type of study, since it considers collapsing two nodes of the same type. This can be interpreted as a measure of the similarity between families of clonotypes that have some overlap in their VDP recognition profile, since it considers both VDP recognition overlap with single clonotypes, and also pairs of clonotypes, thus extending the clustering from clonotype-to-clonotype to clonotype-to-clonotypes. This type of clustering measure could then be used to compare families of clonotypes within a TCR-VDP recognition network, and identify a community structure based on the interactions of these families of clonotypes. Analysis of this type of clustering would be useful in answering questions regarding the differences in immune responses between individuals, such as the prevalence of public TCR responses to some diseases and

infections (Elhanati *et al.*, 2018; Kedzierska *et al.*, 2006; Vujovic *et al.*, 2020; Zhao *et al.*, 2016).

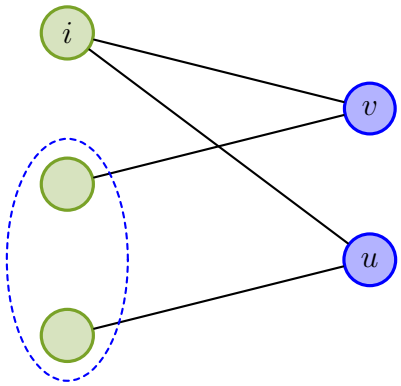


Figure 6.11: Example of a collapsed butterfly in a bipartite network containing i, u, and v. By considering this type of butterfly to be valid in the network, the local clustering coefficient can be used to measure the similarity between clonotypes that share VDPs.

The mixed membership stochastic blockmodel was introduced in Section 6.1.1, however, due to time limitations a full study of this type of networks was not performed. The type of network this generation algorithm is able to produce is in general very variable. As mentioned in Section 6.1.1, concepts such as variable degenerate specificity of TCRs (Joshi *et al.*, 2001; Stewart-Jones *et al.*, 2003), or increased likelihood of some peptides to be recognised due to their anchor residues or overall structure (Carson *et al.*, 1997; Cole *et al.*, 2010; Day *et al.*, 2011; Turner *et al.*, 2005) can be considered in this model, making it a very good candidate for modelling TCR-VDP recognition networks. However, this immense versatility makes it a difficult model to work with computationally and analytically, which is why it was considered beyond the scope of this chapter. Another type of network that was not considered in this chapter was convex bipartite networks (Yang, 2005). These are a special subset of bipartite networks that allow an enumeration such that all the nodes connected to a given node are adjacent to each other in the order given by the enumeration. It can be defined on the clonotype nodes, the VDP nodes, or both. It is a desired quality to have when considering the idea of defining a distance based on the recognition network, since, for example it can be

interpreted as structural similarity between TCRs that have similar recognition profiles, or conversely, structural or anchor residue similarity between peptides that are recognised by similar TCRs.

Chapter 7

Concluding remarks

The importance of T cells as a part of the immune system is two-fold: first, they are able to mount fast, in the scale of days (Murphy & Weaver, 2016, Chapter 9), and highly specific immune responses that allow for the swift clearance of invading pathogens, and second, they have the capability of generating immunological memory to infections the host has encountered in the past, conferring immunity to future encounters with the same pathogens. More than this, both the immune response itself, and the memory that it generates are highly diverse, being comprised of several different T cell clonotypes with different specificities. This combination of memory generation, and cross-reactivity of T cell clonotypes for different peptides makes it possible for T cells to generate immunity (however partial) to pathogens that have never been encountered before. If completely understood, this phenomenon could allow for the development of highly effective vaccines which could confer long lasting cross-reactive immunity to several strains of a given pathogen, for example influenza viruses, which are known to have highly conserved epitopes that are great candidates for this type of vaccine. The models presented in this Thesis were developed with the goal of expanding the current understanding of T cell cross-reactivity, and providing novel mathematical insights into the dynamics of T cell immune responses.

In this Thesis, a mathematical model was developed to study the population dynamics of T cell clonotypes during homeostasis and infection, with a focus on the phenomenon of cross-reactivity. In Chapter 4 a stochastic competition model for homeostasis was proposed, as a generalisation of the work presented by Stirk

et al. (2010) to more than two clonotypes. With this model, it was shown that naive T cell clonotypes engaging in competition for survival stimulus will become extinct with probability one. This highlights the importance of cross-reactivity, not only as a tool to generate broad immunity after infection, but as a necessary tool of the immune system to continue to provide protection against a wide variety of pathogens, even when some clonotypes become extinct before ever finding their cognate epitope. This model was developed so that even a thymic emigrant with a low probability of establishing itself in the periphery, has the capacity to perturb other clonotypes that it competes with for stimulus in order to study said perturbations. First step analysis of the model showed that even when a new clonotype is at a disadvantage compared to pre-established populations, it will still have a measurable negative impact on their average populations in a long timescale. One limitation of this model was the computational complexity of the stochastic descriptors defined, which prevented the study of competitions between a larger population of clonotypes.

In Chapter 5 the homeostatic model was extended to include the dynamics of differentiation into other T cell phenotypes, in particular the effector and memory phenotypes, to allow for the study of T cell dynamics during infection. However, the computational complexity of the stochastic model, together with the larger state space (due to the expansion of effector populations during infection), meant the model could not be used effectively to study infection. For this reason the linear noise approximation was used to find a deterministic approximation of this extended competition model. Structural identifiability was used to determine the viability of parametrising this deterministic model using novel data from a heterologous influenza infection experiment provided by Jessica Gaevert and Paul Thomas from St. Jude Children's Research Hospital. The aim of this parametrisation was to gain better understanding of the effects of affinity and avidity in cross-reactive immune responses. However, the parameters of interest could not be identified with the data available. Using statistical methods, the data set was analysed and it was found that cross-reactivity does not behave in a symmetrical manner. That is to say, developing cross-reactive immunity from pathogen A to pathogen B, is not the same as developing immunity from pathogen B to pathogen A. The analysis showed that some epitopes can generate

selfish cross-reactivity, where the immune response is indeed cross-reactive to other epitopes, but it displays a preference to recall cells specific to the original epitope. On the other hand, analysis showed that some epitopes can generate *selfless* cross-reactivity, where the memory generated by an initial infection will be more broadly cross-reactive with other epitopes, and the response will not experience a recall to epitope specific cells but instead to double or triple positive cells.

Finally, in Chapter 6 two different hypotheses on how cross-reactivity arises in the immune system were studied using the stochastic model defined in Chapter 5 and network theory to generalise the bipartite recognition network that was used in previous chapters to encode the cross-reactivity profiles of different TCRs. The hypotheses studied were the ideas of focussed and unfocussed cross-reactivity proposed by Mason (1998). Using random network generation algorithms, bipartite recognition networks that displayed focussed and unfocussed cross-reactivity were generated, and the clustering coefficient was used to compare the different community structures that the different random networks had. On top of this, using stochastic simulation the generated networks were shown to be able to generate immune dynamics and clonal behaviour that are biologically viable and could be observed in nature. The analysis of the community structure of these networks was limited to the clustering coefficient proposed by Zhang *et al.* (2008), but other definitions offer the possibility of gaining different insights into the structure of bipartite cross-reactive networks. More than this, other types of analysis like one-mode projections, or spectral graph theory may provide, again, different new insights into this area of research.

Appendix A

Code for the study of naive T cell homeostasis

Code used to calculate the stochastic descriptors defined in Chapter 4. The code presented here takes advantage of the quasi-birth-and-death structure of the matrices (Gómez-Corral & López-García, 2018; Kulkarni, 2017) and uses a linear level-reduction algorithm (Gaver *et al.*, 1984). The full code has been published and is available at https://doi.org/10.5281/zenodo.6342372 (Luque, 2022a).

A.1 QSD of the competition process

For the calculation of the QSD the Nåsell approximation (Nåsell, 1991, 2001), and the linear noise approximation (Elf & Ehrenberg, 2003; van Kampen, 2007) are considered, and compared using the Hellinger distance (Oosterhoff & van Zwet, 2012).

A.1.1 Stochastic approximation of the QSD

By taking advantage of the quasi-birth-and-death structure of the matrices, a linear level-reduction algorithm is constructed (see Algorithm 4.1) to find the QSD of the approximating processes $\mathcal{X}^{(1)}$, and $\mathcal{X}^{(2)}$. In order to solve the matrix equations in Eq. (4.42), first the $\mathbf{A}_{l,\ell}^{(j)}$ matrices must be calculated, which are then used to find the $\mathbf{H}_{k}^{(j)}$ matrices using Eq. (4.41). The $\mathbf{A}_{l,\ell}^{(j)}$ are calculated as follows

```
# Main diagonal matrices
for level_value in range(dimension, max_level_value + 1):
    matrix = main_diagonal_matrices_approximation(
        level_value,
        max_level_value,
        dimension,
        probability_values,
        mu_value,
        nu_value,
        stimulus_value,
        model_value,
    )
    matrices[0].append(matrix)

# Lower diagonal matrices
for level_value in range(
    dimension + 1, max_level_value + 1
):
    matrix = death_diagonal_matrices_approximation(
        level_value, dimension, mu_value, model_value
    )
    matrices[1].append(matrix)

# Upper diagonal matrices
for level_value in range(dimension, max_level_value):
    matrix = birth_diagonal_matrices_approximation(
        level_value, dimension, probability_values,
        nu_value, stimulus_value
    )
    matrices[2].append(matrix)
```

where the `_diagonal_matrices_approximation` functions calculate the block matrices using the birth and death rates of the approximating process, with the `model` parameter indicating which approximating process is used ($0 \to \mathcal{X}^{(1)}$, and $1 \to \mathcal{X}^{(2)}$). Then, using Eq. (4.41), the inverses of the $\mathbf{H}_k^{(j)}$ are calculated as follows

```
# Calculating the inverse of H_{N}^{(j)}
h_matrices = [inv(matrices[0][-1])]
```

```
# Calculating the remaining inverses of the H_{k}^{(j)} matrices
for level_value in range(len(matrices[0]) - 1):
    gc.collect()
    matrix = matrices[0][-(level_value + 2)]
    matrix_term = matrices[2][-(level_value + 1)].dot(
        h_matrices[-1].dot(matrices[1][-(level_value + 1)])
    )
    matrix -= matrix_term
    matrix = np.linalg.inv(matrix.todense())
    h_matrices.append(csc_matrix(matrix))
```

Finally, using Eq. (4.44), the quasi-stationary distribution is calculated as follows

```
# Setting the initial value of the distribution to 1
distribution = [np.array([1])]

for level_value in range(len(h_matrices) - 1):
    value = (
        distribution[level_value] * (-1) * matrices[2][
            level_value].dot(
             h_matrices[-(level_value + 2)]
        )
    )
    distribution.append(value.flatten())

# Normalising the values of the distribution
subTotals = [level.sum() for level in distribution]
total = sum(subTotals)
for level_value in range(len(distribution)):
    distribution[level_value] = distribution[level_value] / total
```

A.1.2 Deterministic approximation of the QSD

Using the results of the linear noise approximation, the QSD is approximated by calculating the positive steady states of Eq. (4.51) using Mathematica as follows

```
dn1 = n1(phi1*((probabilities[[1,1]]/n1) + (probabilities
   [[1,2]]/(n1 + n2)) + (probabilities[[1,3]]/(n1 + n3)) + (
   probabilities[[1,4]]/(n1 + n2 + n3))) - mu);
dn2 = n2(phi2*((probabilities[[2,1]]/n2) + (probabilities
   [[2,2]]/(n1 + n2)) + (probabilities[[2,3]]/(n2 + n3)) + (
```

```
    probabilities[[2,4]]/(n1 + n2 + n3))) - mu);
dn3 = n3(phi3*((probabilities[[3,1]]/n3) + (probabilities
    [[3,2]]/(n1 + n3)) + (probabilities[[3,3]]/(n2 + n3)) + (
    probabilities[[3,4]]/(n1 + n2 + n3))) - mu);

solution = Solve[{dn1==0, dn2==0, dn3==0}, {n1, n2, n3}];
If[Count[n1>0&&n2>0&&n3>0/. solution, True] == 1,
  populationResult = Extract[solution , Position[n1>0&&n2>0&&n3
     >0/. solution, True]]]
```

Then, the stability of the steady state is checked by calculating the Jacobian matrix and checking that all its eigenvalues are negative.

```
M = D[{dn1s, dn2s, dn3s}, {{n1, n2, n3}}];
eigenvectors = Eigenvalues[M/.populationResult[[1]]];
AllTrue[eigenvectors, #<0&]
```

A.1.3 Hellinger distance between distributions

To calculate the Hellinger distance between two 3-dimensional discrete probability distributions, as defined in Oosterhoff & van Zwet (2012), the function `hellinger_distance` is defined. This function takes two 3-dimensional distributions and calculates the distance between them as follows

```
shapes = (distributions[0].shape[0], distributions[1].shape[0])
max_index = np.argmax(shapes)
distance = 0
for i in range(shapes[max_index]):
    for j in range(shapes[max_index]):
        for k in range(shapes[max_index]):
            try:
                distance += (
                    math.sqrt(distributions[0][i][j][k])
                    - math.sqrt(distributions[1][i][j][k])
                ) ** 2
            except IndexError:
                distance += math.fabs(
                    distributions[max_index][i][j][k]
                )
distance = (1 / math.sqrt(2)) * math.sqrt(distance)
```

A.2 Mean time to extinction

The mean time to extinction is calculated by writing Eq. (4.68) in matrix form, and solving the equation as follows

```
M = coefficient_matrix(
    probability_values, max_level_value, mu_value, nu_value,
        stimulus_value
)
b = [-1] * int(comb(max_level_value, dimension))

Solution = spsolve(M, b)
```

where the `coefficient_matrix` function generates the coefficient matrix for Eq. (4.68).

A.3 Distribution of clonal sizes at the first extinction event

To calculate the distribution of clonal sizes at the time of extinction, first the $\mathbf{B}_{j,k}$, and $\mathbf{R}^{i}_{j,j+1}$ matrices must be calculated. The upper diagonal matrices $\mathbf{B}_{j,j+1}$ defined in Eq. (4.77), representing birth events, are calculated using the `birth_diagonal_matrices` function as follows

```
for level_value in range(dimension, max_level_value):
    b_matrices.append(
        birth_diagonal_matrices(
            level_value,
            dimension,
            probability_values,
            stimulus_value,
            mu_value,
            nu_value,
        )
    )
```

The lower diagonal matrices $\mathbf{B}_{j,j-1}$ defined in Eq. (4.76), representing death events, are calculated using the `death_diagonal_matrices` function as follows

```
for level_value in range(dimension + 1, max_level_value + 1):
    d_matrices.append(
        death_diagonal_matrices(
            level_value,
            max_level_value,
            dimension,
            probability_values,
            stimulus_value,
            mu_value,
            nu_value,
        )
    )
```

Finally, the absorption matrices $\mathbf{R}^{i}_{j,j+1}$ defined in Eq. (4.78), representing when the competition process reaches an absorbing state, are calculated as follows

```
for clone_number in range(dimension):
    for absorbing_level_value in range(dimension - 1,
       max_level_value):
        block_column = []
        for level_value in range(dimension, max_level_value + 1):
            if absorbing_level_value != level_value - 1:
                block_column.append(
                    dok_matrix(
                        (
                            int(comb(
                                level_value - 1,
                                dimension - 1)
                            ),
                            int(
                                comb(
                                    absorbing_level_value - 1,
                                    dimension - 2,
                                )
                            ),
                        )
                    ).tocsc()
                )
            else:
                block_column.append(
                    absorption_matrix(
```

```
                        level_value,
                        clone_number,
                        max_level_value,
                        dimension,
                        mu_value,
                        nu_value,
                        probability_values,
                        stimulus_value,
                    )
                )
        a_matrices[clone_number].append(block_column)
```

Now, the $\mathbf{H}_j^{-1}$ matrices used in Algorithm 4.2 are calculated recursively as follows

```
h_matrices = [identity(d_matrices[-1].shape[0], format="csc")]
for level_order in range(len(d_matrices)):
    gc.collect()
    matrix = identity(
        b_matrices[-(level_order + 1)].shape[0], format="csc"
    ) - b_matrices[-(level_order + 1)].dot(
        h_matrices[-1].dot(d_matrices[-(level_order + 1)])
    )
    matrix = np.linalg.inv(matrix.todense())
    h_matrices.append(csc_matrix(matrix))
```

which are then used to calculate the $\mathbf{K}_{j,k}^i$ as follows

```
k_matrices = [a_matrices[clone_number][column_number][-1]]
for level_order in range(
    len(a_matrices[clone_number][column_number]) - 1
):
    k_matrices.append(
        b_matrices[-(level_order + 1)].dot(
            h_matrices[level_order].dot(k_matrices[-1])
        )
        + a_matrices[clone_number][column_number][-(level_order +
            2)]
    )
```

Finally, the distribution of clonal sizes at the time of extinction defined in Eq. (4.74), $\mathbf{U}$, is calculated as

```
distribution_column = [h_matrices[-1].dot(k_matrices[-1])]
for level_order in range(len(k_matrices) - 1):
    matrix_term = (
        d_matrices[level_order].dot(distribution_column[-1])
        + k_matrices[-(level_order + 2)]
    )
    distribution_column.append(
        h_matrices[-(level_order + 2)].dot(matrix_term)
    )
distribution[clone_number].append(distribution_column)
```

A.4 Probability distribution of the number of divisions before extinction

In order to calculate the probability distribution of the number of divisions before extinction, first the $\mathbf{C}^{(i)}_{j,k}$ matrices must be calculated. The upper diagonal matrices defined in Eq. (4.83), $\mathbf{C}^{(i)}_{j,j+1}$, are calculated as follows

```
for level_value in range(max_level_value):
    b_matrices.append(
        birth_diagonal_matrices_division(
            level_value,
            dividing_clone,
            dimension_value,
            probability_values,
            stimulus_value,
            mu_value,
            nu_value,
        )
    )
```

and the lower diagonal matrices defined in Eq. (4.82), $\mathbf{C}^{(i)}_{j,j-1}$ are calculated as follows

```
for level_value in range(1, max_level_value + 1):
    d_matrices.append(
        death_diagonal_matrices_division(
            level_value,
```

```
            max_level_value,
            dividing_clone,
            dimension_value,
            probability_values,
            stimulus_value,
            mu_value,
            nu_value,
        )
    )
```

and these matrices are used to find the $\mathbf{H}_j^{-1}$ used in Algorithm 4.2 as follows

```
h_matrices = [identity(d_matrices[-1].shape[0], format="csc")]
for level_order in range(len(d_matrices)):
    gc.collect()
    matrix = identity(
        b_matrices[-(level_order + 1)].shape[0], format="csc"
    ) - b_matrices[-(level_order + 1)].dot(
        h_matrices[-1].dot(d_matrices[-(level_order + 1)])
    )
    matrix_inv = np.linalg.inv(matrix.todense())
    h_matrices.append(csc_matrix(matrix_inv))
```

Since the probability distribution of the number of divisions is found recursively, the vector of division probabilities defined in Eq. (4.86), $\mathbf{d}_j^{(i)}$, is calculated by

```
for current_level in range(max_level_value + 1):
    d_vectors.append(
        division_vector(
            current_level,
            dividing_clone,
            current_division,
            max_level_value,
            dimension_value,
            probability_values,
            stimulus_value,
            mu_value,
            nu_value,
            previous_division,
        )
    )
```

where the parameter `previous_division` contains the vector of probabilities calculated in the prior step.

Then, similarly to the code presented in Appendix A.3, the $\mathbf{K}^i_{j,k}$ matrices are calculated

```
k_vectors = [d_vectors[-1]]
for level_order in range(len(b_matrices)):
    vector = (
        b_matrices[-(level_order + 1)].dot(
            h_matrices[level_order].dot(k_vectors[-1])
        )
        + d_vectors[-(level_order + 2)]
    )
    k_vectors.append(vector)
```

which are then used to calculate the probability of dividing ℓ times before becoming extinct defined in Eq. (4.87), $\mathcal{D}_{i,\ell}$, as follows

```
if num_divisions != 0:
    distribution.append(coo_matrix(([], ([], [])), [1, 1]).tocsc
        ())
else:
    distribution.append(coo_matrix(([1], ([0], [0])), [1, 1]).
        tocsc())
for level_order in range(max_level_value):
    matrix_term = (
        d_matrices[level_order].dot(distribution[-1])
        + k_vectors[-(level_order + 2)]
    )
    distribution_value = h_matrices[-(level_order + 2)].dot(
        matrix_term)
    distribution.append(distribution_value)
```

Appendix B

Tetramer specific population numbers for lung circulating and spleen resident cells

This Appendix contains the mean frequencies and mean absolute numbers of cells for lung resident and spleen circulating samples of WT-primed, T8A-primed, and N3A-primed mice from the experiments carried out by by Jessica Gaevert at the Paul Thomas laboratory in the Immunology department of St. Jude Children's Research Hospital, which are described in Section 5.3.1.

WT primary – Lung resident frequencies							
Timepoint	WT	T8A	N3A	WT+T8A	WT+N3A	T8A+N3A	Triple positive	Triple negative
Primary	0.0408	0.0158	0.0013	0.016	0.0012	0.0001	0.0004	0.9239
Memory	0.0287	0.0008	0.001	0.0007	0.0037	0	0	0.965
WT	0.1526	0.0036	0.0016	0.1442	0.0095	0	0.0095	0.679
T8A	0.0254	0.0106	0.0493	0.0448	0.0011	0.0023	0.0171	0.8495
N3A	0.0128	0.0006	0.0559	0.0031	0.036	0.0001	0.0272	0.8644

WT primary – Lung resident absolute numbers								
Timepoint	WT	T8A	N3A	WT+T8A	WT+N3A	T8A+N3A	Triple positive	Triple negative
Primary	43980	17124	1374.6	17428	1318.2	179.8	518.6	1002876.8
Memory	4405	166.25	260.75	123.5	740.5	0	0	229304
WT	953750	23210.4	10575.5	1211650	68327.5	101.5	74758	3065127.1
T8A	281493.333	97535	480666.667	444883.333	10295.5	23456.667	163121.667	8126881.167
N3A	85600	3943.333	353766.667	22093.333	239500	720.333	181603.333	5812773

Table B.1: Mean frequency and mean absolute number of tetramer positive lung resident cells in WT-primed mice.

T8A primary – Lung resident frequencies								
Timepoint	WT	T8A	N3A	WT+T8A	WT+N3A	T8A+N3A	Triple positive	Triple negative
Primary	0.0083	0.0436	0.0121	0.0252	0.0002	0.0002	0.0005	0.9099
Memory	0.0212	0.0152	0	0.0139	0	0	0	0.9497
WT	0.0042	0.0609	0.0004	0.0952	0	0	0.0006	0.8385
T8A	0.0009	0.1447	0.1686	0.0222	0.0005	0.0682	0.0053	0.5897
N3A	0.0007	0.0036	0.001	0.0019	0.0005	0	0.0003	0.992
T8A primary – Lung resident absolute numbers								
Timepoint	WT	T8A	N3A	WT+T8A	WT+N3A	T8A+N3A	Triple positive	Triple negative
Primary	6812	60500	7818	33053.75	115.875	237	955.125	1180883.25
Memory	2435	1828	0	1782.75	0	0	0	122104.25
WT	59050	799750	6444.5	1389500	578	109.25	11086.25	11973482
T8A	10884	2065800	2923940	292200	7744	1006552	103054	7199826
N3A	6084	33340	9616	16832	4318	163.6	2536	9383110.4

Table B.2: Mean frequency and mean absolute number of tetramer positive lung resident cells in T8A-primed mice.

N3A primary – Lung resident frequencies							
Timepoint	WT	T8A	N3A	WT+T8A	WT+N3A	T8A+N3A	Triple positive	Triple negative
Primary	0.0035	0.0007	0.1549	0.0009	0.0057	0.0001	0.0032	0.831
Memory	0.0021	0.0036	0.055	0.0014	0.0108	0.0014	0.0057	0.92
WT	0.0197	0.0005	0.2732	0.0112	0.1038	0.0007	0.0596	0.5313
T8A	0.0013	0.0021	0.2337	0.0076	0.0016	0.0012	0.0052	0.7473
N3A	0.0026	0.0016	0.2188	0.0053	0.0564	0.0011	0.005	0.7092

N3A primary – Lung resident absolute numbers							
Timepoint	WT	T8A	N3A	WT+T8A	WT+N3A	T8A+N3A	Triple positive	Triple negative
Primary	17103.333	3313.333	772666.667	4421.333	29266.667	446.667	16076.667	4276705.333
Memory	853.333	2203.333	37200	823.333	7920	1350	4856.667	583460
WT	217333.333	5753.333	3000000	130133.333	1138333.333	8216.667	646333.333	5843896.667
T8A	16637.5	25425	2875000	86850	20967.5	16367.5	61847.5	9121905
N3A	58300	31566.667	4286666.667	117533.333	1285666.667	18866.667	93966.667	13807433.333

Table B.3: Mean frequency and mean absolute number of tetramer positive lung resident cells in N3A-primed mice.

WT primary – Spleen circulating frequencies								
Timepoint	WT	T8A	N3A	WT+T8A	WT+N3A	T8A+N3A	Triple positive	Triple negative
Primary	0.0305	0.0135	0.0004	0.0101	0.0079	0.0006	0.0026	0.9344
Memory	0.015	0.0007	0.0002	0.0072	0.0006	0	0.0001	0.9763
WT	0.1186	0.0036	0.0002	0.1423	0.0019	0	0.0019	0.7314
T8A	0.0057	0.0489	0.0004	0.0448	0.0008	0.0001	0.0131	0.8861
N3A	0.0156	0	0.0008	0.0066	0.042	0.0003	0.054	0.8808
WT primary – Spleen circulating absolute numbers								
Timepoint	WT	T8A	N3A	WT+T8A	WT+N3A	T8A+N3A	Triple positive	Triple negative
Primary	43260	18044	333.8	13820	12228	632.8	4055.2	1199626.2
Memory	30906.667	1588.889	120	7526.556	619.889	0	393.333	1907923.556
WT	1057312.5	32473.25	1815	1130300	11830	0	14023.25	5072746
T8A	56733.333	364766.667	3535	352350	5855	1261.667	104765	7055733.333
N3A	108666.667	0	4400	38933.333	228436.667	2190	441666.667	3940873.333

Table B.4: Mean frequency and mean absolute number of tetramer positive spleen circulating cells in WT-primed mice.

T8A primary – Spleen circulating frequencies							
Timepoint	WT	T8A	N3A	WT+T8A	WT+N3A	T8A+N3A	Triple positive	Triple negative
Primary	0.0094	0.0962	0	0.0522	0	0	0	0.8421
Memory	0.0066	0.0061	0.0005	0.0016	0.0006	0	0.0001	0.9845
WT	0.0112	0.0458	0	0.1704	0	0	0.0002	0.7723
T8A	0.0008	0.2115	0.0008	0.0121	0	0.0008	0.0001	0.774
N3A	0.0021	0.0012	0.0055	0.0019	0.0015	0	0.0005	0.9873

T8A primary – Spleen circulating absolute numbers							
Timepoint	WT	T8A	N3A	WT+T8A	WT+N3A	T8A+N3A	Triple positive	Triple negative
Primary	37120	630000	0	359600	0	0	0	5109280
Memory	37821.429	28960	4265.714	12072.857	4557.143	0	447.143	6286590
WT	56175	224500	0	806250	335	0	742.5	3584497.5
T8A	11674	3523200	8098	159160	0	18100	2400	11359368
N3A	15536	9688	38720	13834	9610	0	4039.8	7176572.2

Table B.5: Mean frequency and mean absolute number of tetramer positive spleen circulating cells in T8A-primed mice.

N3A primary – Spleen circulating frequencies								
Timepoint	WT	T8A	N3A	WT+T8A	WT+N3A	T8A+N3A	Triple positive	Triple negative
Primary	0.0055	0.0044	0.0037	0.0026	0.0048	0.0003	0.0022	0.9765
Memory	0.0034	0.0003	0.0052	0	0.0007	0.0004	0.0017	0.9882
WT	0.0531	0.0016	0.0021	0.0999	0.1526	0.0001	0.0595	0.6311
T8A	0.0076	0.0118	0.0028	0.0215	0.0012	0.0003	0.0006	0.9542
N3A	0.0062	0.0016	0.067	0.0047	0.0895	0.0002	0.0003	0.8307
N3A primary – Spleen circulating absolute numbers								
Timepoint	WT	T8A	N3A	WT+T8A	WT+N3A	T8A+N3A	Triple positive	Triple negative
Primary	33133.333	25300	17440	18600	25800	2073.333	12033.333	5272286.66
Memory	16640	2010	21016.667	0	4233.333	430	1723.333	4688613.33
WT	1616333.333	24480	44333.333	1291666.667	5184000	776.667	1859666.66	17245410
T8A	40650	57687.5	12665	60595	5534.25	1497.5	2945	4330925.75
N3A	141743.333	42433.333	324666.667	117830	2385100	3433.333	7206.667	14290920

Table B.6: Mean frequency and mean absolute number of tetramer positive spleen circulating cells in N3A-primed mice.

B. TETRAMER SPECIFIC POPULATION NUMBERS FOR LUNG CIRCULATING AND SPLEEN RESIDENT CELLS

Appendix C

Code for the statistical analysis of IAV infection data

Code used to read flow-cytometry results from the experiment described in Section 5.3.1, and perform the statistical analysis described in Section 5.4. Python codes are for the visualisation of the data in Section 5.3.1, and part of the analysis in Section 5.4.1. R codes are used for the analysis in Sections 5.4.1, 5.4.2, 5.4.3, and 5.4.4. The full code has been published and is available at https://doi.org/10.5281/zenodo.7463077 (Luque, 2022b)

C.1 Reading flow-cytometry results

The experimental data presented in Section 5.3.1 is stored in `.csv` files (one file per primary infection and CD45 positivity combination), where each row represents a sample taken from a mouse, and columns represent: the tissue the sample was taken from, the timepoint at which it was taken, the total number of cells in the sample, the frequency of each tetramer-positive population, and the total number of cells of each tetramer-positive population. A function, `header_clipping`, is used to clear trailing whitespace from the column names, and standardise the names of all files.

Three classes are used to analyse the data read from the files: the `Mouse` class, used to represent a sample taken from a mouse, the `Timepoint` class, used to represent a collection of `Mouse` objects from a given timepoint, and the

`Experiment` class, used to represent a collection of `Timepoint` objects for a given primary infection and CD45 positivity.

The `Mouse` class requires the number of cells of each tetramer-specific population in order to create an object, as summarised in Table C.1. The `Timepoint` constructor creates an empty `Timepoint` object, and the class implements methods to assign a name to an instance (`Timepoint.change_name(name : str)`), and add `Mouse` objects to it (`Timepoint.add_mice(mice : list[Mouse])`). Finally the `Experiment` class constructor requires a name for the experiment (`name`), and a tag consisting of the name of the primary infection (`tag`). This also implements a method to add `Timepoint` objects to a dictionary (`Experiment.add_timepoint (timepoints : list[Timepoint], timepoint_names : list[str])`), in order to maintain a link between the timepoint names and their data.

Arguments of the `Mouse` class		
Argument	Type	Description
`wt`	`int`	WT single positive cells
`t8a`	`int`	T8A single positive cells
`n3a`	`int`	N3A single positive cells
`wt_t8a`	`int`	WT-T8A double positive cells
`wt_n3a`	`int`	WT-N3A double positive cells
`t8a_n3a`	`int`	T8A-N3A double positive cells
`triple_positive`	`int`	Triple positive cells
`triple_negative`	`int`	Triple negative cells

Table C.1: Arguments required to initialise an object of the `Mouse` class.

In order to extract the data from the `.csv` files into `Experiment` objects, the `data_extraction` function is used. This function takes the following parameters in order to create the `Experiment` object: `primary`, the virus variant used in the primary infection, `tissue`, the tissue the sample was taken from, `headers`, the list of the names of the columns in the file, `time_names`, the list of the names for the timepoints in the file, `standard_names`, a list of the desired names for the timepoints, `cd45`, the CD45 positivity of the sample, and `timepoints`, the number

of timepoints to be extracted from the file. These parameters are summarised in Table C.2.

Arguments of the `Mouse` class		
Argument	Type	Description
`primary`	`str`	Virus variant of the primary infection.
`tissue`	`str`	Tissue the sample was taken from.
`headers`	`list[str]`	List of columns names of the file.
`time_names`	`list[str]`	List of timepoint names in the file.
`standard_names`	`list[str]`	List of desired names for the timepoints.
`cd45`	`str`	CD45 positivity of the sample.
`timepoints`	`int`	Number of timepoints to be extracted.

Table C.2: Parameters used to extract the data from the `.csv` files using the `data_extraction` function.

C.2 Calculation of the Spearman rank correlation

To generate the results presented in Section 5.4.1, the `Experiment` class defines a method, `correlation_heatmap`, that calculates the Spearman rank correlation between the tetramer-specific populations, and generates a heatmap marking the correlations significant at the 0.05, and 0.01 levels. This function uses the `Timepoint.to_df()` method to transform the data of all the `Mouse` objects contained in the timepoint into a `pandas.DataFrame`. Then, the `corr(method="spearman")` method of the `DataFrame` is used to calculate the Spearman rank correlation between each of the tetramer-specific populations of cells. When determining the significance of the correlations, the `correlation_heatmap` method uses the Bonferroni adjustment to correct for multiple comparisons (Bland & Altman, 1995).

C.3 Tetramer specific ANOVA

To cluster the data into the appropriate groups to perform the ANOVA of tetramer-specific cells presented in Section 5.4.1, the `Experiment.to_df()` method is used to generate a `.csv` file containing all the population numbers indexed by tetramer-specificity. Then, the `aov` and `TukeyHSD` functions from the R programming language are used to perform the ANOVA and Tukey's HSD tests and save the results to the `ANOVA/Results/` directory, as follows

```
for (file_index in 1:4){
  Data <- read.csv(
        paste0("ANOVA/Data/Tetramers/",file_names[[file_index]])
  )

  for (time_index in 1:5){
    Timepoint <- stack(
      Data[Data$Challenge == time_names[time_index],],
      select=-Challenge
    )
    aovTimepoint <- aov(values ~ ind, Timepoint)
    tukeyTimepoint <- TukeyHSD(
      aovTimepoint, ordered=TRUE, conf.level=0.95
    )
    write.csv(
      as.data.frame(tukeyTimepoint[1]),
      paste0("ANOVA/Results/WT/Tukey-",
        time_file_names[[time_index]],"-",
        organ[[file_index]],"-",
        cd45[[file_index]],"-F.csv"
      )
    )
  }
}
```

C.4 Epitope specific ANOVA

To perform ANOVA on the epitope-specific cells presented in Section 5.4.2, the `positive_cells_df(experiments : list[Experiment], timepoint :`

`str, tetramer : str, file_name : str)` function is used. This function collects all the populations of cells positive for `tetramer` from the experiments in the list `experiments`, at `timepoint`, and generates a `.csv` file where these populations are clustered by their primary infection. Then, using R, the ANOVA and Tukey's HSD tests are done on the epitope-specific data, and saved to the `ANOVA/Results/<tetramer>/` directory, as follows

```
for (tetramer in tetramers){
  for (cd45 in cd45_list){
    for (tissue in tissue_list){
      for (timepoint in timepoint_list){
        Data <- read.csv(
          paste0("ANOVA/Data/Tetramers/", tetramer,
          timepoint, tissue, cd45, ".csv")
        )
        aovData <- aov(Cells ~ Experiment, Data)
        tukeyData <- TukeyHSD(
          aovData, ordered=TRUE, conf.level=0.95
        )
        write.csv(
          as.data.frame(tukeyData[1]),
          paste0("ANOVA/Results/Tetramers/Tukey-",
          tetramer, timepoint, tissue, cd45, ".csv")
        )
      }
    }
  }
}
```

C.5 Contraction and expansion ANOVA

To analyse the decay and expansion slopes as shown in Sections 5.4.3, and 5.4.4, they must be first calculated from the data, and clustered appropriately. To this end, the functions `decay_slopes_df`, and `expansion_slopes_df` are used. The former takes as parameters a list of `Experiment` objects for the different primary infections, a list of `int` representing the time at which the primary and memory samples are taken, a `str` for the tetramer-specific population being considered,

and a file name to save the clustered data as a `.csv` file. Once the files are generated for all tetramer-specific populations, and both tissues, the ANOVA and Tukey's HSD tests are done as follows

```
for (tissue in tissue_list){
  for (cd45 in cd45_list){
    for (tetramer in tetramer_positivity){
      Data <- read.csv(
        paste0("ANOVA/Data/Decay/",
        tissue, cd45, tetramer, ".csv")
      )
      if (length(unique(Data$Primary)) > 1){
        aovData <- aov(Slope ~ Primary, Data)
        tukeyData <- TukeyHSD(
          aovData, ordered=TRUE, conf.level=0.95
        )
        write.csv(
          as.data.frame(tukeyData[1]),
          paste0("ANOVA/Results/Decay/Tukey-",
          tissue, cd45, tetramer, ".csv")
        )
      }
    }
  }
}
```

On the other hand, the `expansion_slopes_df` takes as parameters a single `Experiment` object, instead of a list of object, as well as a list of `str` representing the challenge infections considered. All the other parameters are the same as those for `decay_slopes_df`. Now, once the data files have been generated, the ANOVA and Tukey's HSD test are performed as follows

```
for (tissue in tissue_list){
  for (cd45 in cd45_list){
    for (primary in infections){
      for (tetramer in tetramer_positivity){
        Data <- read.csv(
          paste0("ANOVA/Data/Expansion/", tissue,
          cd45, primary, tetramer, ".csv")
        )
        if (length(unique(Data$Challenge)) > 1){
```

```
            aovData <- aov(Slope ~ Challenge, Data)
            tukeyData <- TukeyHSD(
              aovData, ordered=TRUE, conf.level=0.95
            )
            write.csv(as.data.frame(tukeyData[1]),
              paste0(
                "ANOVA/Results/Expansion/Tukey-",
                tissue, cd45, primary, tetramer, ".csv"
              )
            )
          }
        }
      }
    }
}
```

C. CODE FOR THE STATISTICAL ANALYSIS OF IAV INFECTION DATA

Appendix D

Code for the analysis of random recognition networks

Python code used to generate random TCR-VDP recognition networks using the models described in Chapter 6, and simulate heterologous immune responses using the Gillespie algorithm (Gillespie, 1976, 1977) on the multi-variate competition model defined in Chapter 5. The full code has been published and is available at https://doi.org/10.5281/zenodo.5227343 (Luque, 2021).

D.1 Generation of random recognition networks

The network generation code is based on two main classes: a `Peptide` class used to represent VDPs, and a `Clonotype` class to represent T cell clonotypes. To create an object of `Peptide` type three arguments are required: `probability`, the probability of the peptide being recognised, `position`, the number of the peptide, and `stimulus`, the stimulus provided by the peptide. These arguments are summarised in Table D.1. When the class is instanced two other attributes are created: `clonotypes`, a list of `Clonotype` objects that can recognise the peptide, and `recognised`, the number of clonotypes that can recognise the peptide.

Objects of the `Clonotype` class require 12 arguments to be created: `position`, the number of the clonotype, `num_cells`, the initial number of cells of the clonotype, `naive_homeostatic_rate`, the homeostatic proliferation rate of a single naive cell, `naive_competition_matrix`, a matrix containing the self-pMHC

Arguments of the `Peptide` class		
Argument	Type	Description
`probability`	`float`	Probability of peptide being recognised.
`position`	`int`	Number of the peptide.
`stimulus`	`float`	Stimulus provided by the peptide.

Table D.1: Arguments required to initialise an object of the `Peptide` class.

competition probabilities, `memory_homeostatic_rate`, the birth rate of a single memory cell, `effector_division_constant`, the constant of effector division, `naive_death_rate`, the death rate of a single naive cell, `effector_death_rate`, the death rate of a single effector cell, `memory_death_rate`, the death rate of a single memory cell, `naive_differentiation_constant`, the differentiation constant from naive to effector, `memory_differentiation_constant`, the differentiation constant from memory to effector, and `effector_differentiation_rate`, the differentiation rate from effector to memory. These arguments are summarised in Table D.2. This class creates 6 other attributes when instanced: `effector`, the number of effector cells of the clonotype, `effector_dividing`, the number of effector cells currently dividing, `effector_division_times`, the list of times when dividing effector cells will finish dividing, `memory`, the number of memory cells, `peptides`, a list of `Peptide` objects that the clonotype can recognise, and `recognised`, the number of peptides the clonotype can recognise.

The main body of the code contains the parameters necessary for the network to be generated and simulated, they must be edited in place to generate the desired network. These parameters are: `network`, the type of network to be generated, 0 for an Erdős-Rényi network (see Section 6.1.1), 1 for a configuration model network (see Section 6.1.2), and 2 for a preferential attachment network (see Section 6.1.3), `num_clonotypes`, the number of clonotypes in the network, `starting_cells`, the number of initial cells for every clonotype, `num_peptides`, the number of peptides considered, `peptide_degree`, the degree of peptide nodes in the configuration model, and `peptide_probability`, the probability of a peptide being recognised. These parameters are summarised in Table D.3.

Arguments of the `Clonotype` class		
Argument	Type	Description
`position`	`int`	Number of the clonotype.
`num_cells`	`int`	Number of initial cells.
`naive_homeostatic_rate`	`float`	Homeostatic proliferation rate for a naive cell.
`naive_competition_matrix`	`numpy.ndarray`	Homeostatic competition sharing probability matrix.
`memory_homeostatic_rate`	`float`	Birth rate for a memory cell.
`effector_division_constant`	`float`	Constant of effector division.
`naive_death_rate`	`float`	Death rate for a naive cell.
`effector_death_rate`	`float`	Death rate for an effector cell.
`memory_death_rate`	`float`	Death rate for a memory cell.
`naive_differentiation_constant`	`float`	Differentiation constant from naive to effector.
`memory_differentiation_constant`	`float`	Differentiation constant from memory to effector.
`effector_differentiation_rate`	`float`	Differentiation rate from effector to memory.

Table D.2: Arguments required to initialise an object of the `Clonotype` class.

D. CODE FOR THE ANALYSIS OF RANDOM RECOGNITION NETWORKS

Parameters to generate a recognition network		
Parameter	Type	Description
`network`	`int`	0 → Erdős-Rényi network, 1 → Configuration model, 2 → Preferential attachment.
`num_clonotypes`	`int`	Number of clonotype.s
`starting_cells`	`int`	Starting number of cells for every clonotype.
`num_peptides`	`int`	Number of peptides.
`peptide_degree`	`int`	Number of recognised peptides on the configuration model network.
`peptide_probability`	`float`	Probability that a peptide will be recognised by a clonotype. If `None` the probability is calculated as `peptide_degree / num_peptides`.

Table D.3: Parameters used to generate a random recognition network using the models described in Chapter 6.

When the code is executed a list of `Peptide` objects is created according to the parameters set. Then, a `for` loop creates `Clonotype` objects one at a time, adds them to the list of clonotypes in the network, and generates their recognition profile according to the chosen network generation model as follows[1]:

```
peptides = [
    Peptide(peptide_prob_value, i, peptide_stimulus_value) for i
        in range(num_peptides)
]
initial_clones = []

for clone_index in range(num_clonotypes):
    initial_clones.append(Clonotype(*args))

    initial_clones[-1].add_peptide(
```

[1]The arguments of the `Clonotype` constructor have been replaced with `*args` in order to make the code easier to read.

```
        peptides, network, degree=peptide_degree, clone_list=
            initial_clones
    )
```

where the `network` parameter determines the network model the `add_peptide` method will use.

When `network` is 0 a network is generated using the Erdős-Rényi model, see Section 6.1.1. In this case a random number between 0 and 1 is generated for each `Peptide` to determine whether it is recognised by the `Clonotype` based on its `probability_value`:

```
for peptide in peptide_list:
    if peptide.position not in self.peptides:
        check_value = uniform(0.0, 1.0)
        if check_value < peptide.probability:
            self.peptides.append(peptide.position)
            self.recognised += 1
            peptide.add_clonotype(self)
```

If `network` is 1, the configuration model is used to generate the network, see Section 6.1.2. For this model a sample of the desired size is taken from the list of `Peptide` objects and added to the recognition profile as follows

```
peptide_sample = sample(list(range(len(peptide_list))), degree)
for peptide_index in peptide_sample:
    self.peptides.append(peptide_index)
    self.recognised += 1
    peptide_list[peptide_index].add_clonotype(self)
```

Finally, if `network` is 2 the preferential attachment model is used, see Section 6.1.3. In this model an initial sample of `Peptide` objects is taken using the Erdős-Rényi model. Then, this recognition profile is compared with that of the other `Clonotype` objects already in the network, and if there is cross-reactivity the `Peptide` objects not already shared are stored in an `extra_peptides` list:

```
self.add_peptide(peptide_list, 0)
if len(clone_list) > 1:
    extra_peptides = []
    for peptide in self.peptides:
        for clone in peptide_list[peptide].clonotypes:
```

```
            for peptide_index in clone_list[clone].peptides:
                if peptide_index not in self.peptides and
                    peptide_index not in extra_peptides:
                     extra_peptides.append(peptide_index)
    self.add_peptide(peptide_list, 0, subset=extra_peptides)
```

Then, `extra_peptides` is used to sample again for `Peptide` objects:

```
for peptide_index in subset:
    if peptide_index not in self.peptides:
        check_value = uniform(0.0, 1.0)
        if check_value < peptide_list[peptide_index].probability:
            self.peptides.append(peptide_index)
            self.recognised += 1
            peptide_list[peptide_index].add_clonotype(self)
```

D.2 Stochastic simulation of heterologous infection

In order to do stochastic simulations using the random recognition network generated in Appendix D.1, the first step is to create lists with copies of the the initial conditions of the simulation. These values will be used as the initial conditions of the stochastic simulation, and their changes are appended to the lists.

```
realisation_states = [deepcopy([clone.cells() for clone in
    current_clones])]
realisation_naive = [deepcopy([clone.naive for clone in
    current_clones])]
realisation_effector = [deepcopy([clone.effector + clone.
    effector_dividing for clone in current_clones])]
realisation_memory = [deepcopy([clone.memory for clone in
    current_clones])]
realisation_times = [0.0]
```

This allows for the desired number of realisations to be run with the same conditions without any risk of changing the original values.

Then, a function named `gillespie_step`, which performs one step of the Gillespie algorithm for the multi-variate competition process defined in Chapter 5,

is defined in the code. This function takes the arguments `clone_list`, a list of the clonotypes in the current state, `time`, the current time, `division_time`, the time for an effector cell to divide, `current_infection`, the current stimulus rate from the infection, `time_limit`, the maximum time for the simulation, and `peptide_list`, the list of peptides presented during the infection. After performing this Gillespie step the function returns the list of updated `Clonotype` objects, and the updated time, if `time_limit` is not `None` and the updated time would take the process over this time, then the function returns the unchanged state and `time_limit`. These parameters and results are summarised in Table D.4

Parameters of the `gillespie_step` function		
Parameter	Type	Description
`clone_list`	`list[Clonotype]`	List of clonotypes in the current state.
`time`	`float`	Current time.
`division_time`	`float`	Time for an effector cell to divide.
`current_infection`	`float`	Current stimulus rate available from infection.
`time_limit`	`float`	Maximum time simulated.
`peptide_list`	`list[Peptide]`	List of currently present peptides.

Results of the `gillespie_step` function	
Type	Description
`list[Clonotype]`	List of clonotypes in the updated state.
`float`	Updated time.

Table D.4: Parameters and results of the `gillespie_step` which perform one step of the Gillespie algorithm for the model defined in Chapter 5.

Using this function, the simulation is computed one step at a time as follows[1]:

```
while current_time != challenge_end:
```

[1]The arguments of the `gillespie_step` function have been replaced with `*args` in order to make the code easier to read

```
    current_clones, current_time = gillespie_step(*args)

    realisation_states.append(deepcopy([clone.cells() for clone
        in current_clones]))
    realisation_naive.append(deepcopy([clone.naive for clone in
        current_clones]))
    realisation_effector.append(deepcopy([clone.effector + clone.
        effector_dividing for clone in current_clones]))
    realisation_memory.append(deepcopy([clone.memory for clone in
         current_clones]))
    realisation_times.append(deepcopy(current_time))

    if [clone.cells() for clone in current_clones] == [0 for _ in
         current_clones]:
        break
```

First, the `gillespie_step` function is used to calculate the next state of the process and the time when this state was reached. Then, the lists containing the populations and times of the current realisation are updated with the results of the `gillespie_step` function. If all the populations have become extinct the simulation is stopped. Now, since the simulated scenario has periods of infection and homeostasis, the simulation is split into sections defined by their end-time and the peptides being presented during that section. Then, each section is simulated as shown above, and between simulations the current state is checked for total extinction:

```
if [clone.cells() for clone in current_clones] == [0 for _ in
    current_clones]:
    states.append(deepcopy(realisation_states))
    naive.append(deepcopy(realisation_naive))
    effector.append(deepcopy(realisation_effector))
    memory.append(deepcopy(realisation_memory))
    times.append(deepcopy(realisation_times))
    continue
```

If all populations have become extinct, all data from the current realisation is stored, and the next realisation is run.

D.3 Calculation of the clustering coefficient

Using the expressions for the clustering coefficient found in Section 6.2 (see Table 6.1), a function `clustering_coefficient` is defined to calculate the coefficients. This function takes the parameter `network`, which informs the function which type of network generation model is being considered as shown in Table D.5, and multiple keyword parameters depending on the type of network, which are summarised in Table D.6. The function returns the clustering coefficient for a network with the given parameters as a `float`, or -1 if an error occurs.

Description of the `network` parameter	
Type	Options
`str`	`mmsb` → Mixed membership stocahstic blockmodel, `er` → Erdős-Rényi model, `c` → Configuration model, `pa` → Preferential attachment model.

Table D.5: Description of the `network` parameter, and its options for the `clustering_coefficient` function.

Keyword parameters of the `clustering_coefficient` function			
Parameter	Type	Description	`network`
`clonotype_probabilities`	`list[list[float]]`	List of clonotype affiliation vectors.	`mmsb`
`epitope_probabilities`	`list[float]`	List of epitope clock recognition probabilities.	`mmsb`
`recognition_probability`	`float`	Epitope recognition probability.	`er, c, pa`
`clonotypes`	`int`	Number of clonotypes.	`re, c, pa`
`epitopes`	`int, list[int]`	Number of epitopes or list of epitope degrees.	`c, pa` (only `int`)
`crossreactivity`	`float`	Base cross-reactivity constant.	`pa`

Table D.6: Keyword parameters of the `clustering_coefficient` function and the values of `network` for which they are required.

References

Abdi, H. & Williams, L.J. (2010). Tukey's honestly significant difference (HSD) test. *Encyclopedia of research design*, **3**, 1–5. 25

Adachi, K. & Davis, M.M. (2011). T-cell receptor ligation induces distinct signaling pathways in naive vs. antigen-experienced T cells. *Proceedings of the National Academy of Sciences*, **108**, 1549–1554. 100

Adams, J.J., Narayanan, S., Birnbaum, M.E., Sidhu, S.S., Blevins, S.J., Gee, M.H., Sibener, L.V., Baker, B.M., Kranz, D.M. & Garcia, K.C. (2016). Structural interplay between germline interactions and adaptive recognition determines the bandwidth of TCR-peptide-MHC cross-reactivity. *Nature immunology*, **17**, 87–94. 141

Ahmed, R. & Gray, D. (1996). Immunological memory and protective immunity: understanding their relation. *Science*, **272**, 54–60. 4, 12, 94, 96, 98

Ahmed, R., Westera, L., Drylewicz, J., Elemans, M., Zhang, Y., Kelly, E., Reljic, R., Tesselaar, K., De Boer, R.J., Macallan, D.C. *et al.* (2015). Reconciling estimates of cell proliferation from stable isotope labeling experiments. *PLoS computational biology*, **11**, e1004355. 137

Airoldi, E.M., Blei, D.M., Fienberg, S.E. & Xing, E.P. (2008). Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, **9**, 1981–2014. 144, 146

Akram, A. & Inman, R.D. (2012). Immunodominance: a pivotal principle in host response to viral infections. *Clinical immunology*, **143**, 99–115. 169

Alam, S.M., Travers, P.J., Wung, J.L., Nasholds, W., Redpath, S., Jameson, S.C. & Gascoigne, N.R. (1996). T-cell-receptor affinity and thymocyte positive selection. *Nature*, **381**, 616–620. 11

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2002). Lymphocytes and the cellular basis of adaptive immunity. In *Molecular Biology of the Cell. 4th edition*, Garland Science. 31

Allen, L.J. (1994). Some discrete-time SI, SIR, and SIS epidemic models. *Mathematical biosciences*, **124**, 83–105. 2

Allen, L.J. (2007). *Introduction to mathematical biology*. Pearson/Prentice Hall. 22

Allen, L.J. (2010). *An introduction to stochastic processes with applications to biology*. CRC Press, Lubbock, TX, 2nd edn. 2, 17, 36, 58

Alt, F.W., Oltz, E.M., Young, F., Gorman, J., Taccioli, G. & Chen, J. (1992). VDJ recombination. *Immunology today*, **13**, 306–314. 9

Amador, J. & Gómez-Corral, A. (2020). A stochastic epidemic model with two quarantine states and limited carrying capacity for quarantine. *Physica A: Statistical Mechanics and its Applications*, **544**, 121899. 1

Andersen, M.H., Schrama, D., thor Straten, P. & Becker, J.C. (2006). Cytotoxic T cells. *Journal of Investigative Dermatology*, **126**, 32–41. 8

Anderson, G. & Takahama, Y. (2012). Thymic epithelial cells: working class heroes for T cell development and repertoire selection. *Trends in immunology*, **33**, 256–263. 11

Anthonisse, J.M. (1971). The rush in a directed graph. *Stichting Mathematisch Centrum. Mathematische Besliskunde*. 153, 171

Apetoh, L., Smyth, M.J., Drake, C.G., Abastado, J.P., Apte, R.N., Ayyoub, M., Blay, J.Y., Bonneville, M., Butterfield, L.H., Caignard, A. *et al.* (2015). Consensus nomenclature for $CD8^+$ T cell phenotypes in cancer. *Oncoimmunology*, **4**, e998538. 3

Appay, V. & Rowland-Jones, S.L. (2004). Lessons from the study of T-cell differentiation in persistent human virus infection. In *Seminars in immunology*, vol. 16, 205–212, Elsevier. 4

Arstila, T.P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J. & Kourilsky, P. (1999). A direct estimate of the human $\alpha\beta$ T cell receptor diversity. *Science*, **286**, 958–961. 8, 31

Aspinall, R. & Andrew, D. (2000). Thymic involution in aging. *Journal of clinical immunology*, **20**, 250–256. 85

Badovinac, V.P., Messingham, K.A., Jabbari, A., Haring, J.S. & Harty, J.T. (2005). Accelerated $CD8^+$ T-cell memory and prime-boost response after dendritic-cell vaccination. *Nature medicine*, **11**, 748–756. 95

Bain, L.J. & Engelhardt, M. (2000). *Introduction to probability and mathematical statistics*. Duxbury Press, Belmont, CA. 39

Bangs, S.C., McMichael, A.J. & Xu, X.N. (2006). Bystander T cell activation–implications for HIV infection and other diseases. *Trends in immunology*, **27**, 518–524. 112

Barabási, A.L. & Albert, R. (1999). Emergence of scaling in random networks. *science*, **286**, 509–512. 145, 150, 151

Baroyan, O.V., Rvachev, L.A., Basilevsky, U.V., Ermakov, V.V., Frank, K.D., Rvachev, M.A. & Shashkov, V.A. (1971). Computer Modelling of Influenza Epidemics for the Whole Country (USSR). *Advances in Applied Probability*, **3**, 224–226. 1

Battiston, S. & Catanzaro, M. (2004). Statistical properties of corporate board and director networks. *The European Physical Journal B*, **38**, 345–352. 171

Bender, E.A. & Canfield, E.R. (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, **24**, 296–307. 147

Bevan, M.J. (2004). Helping the $CD8^+$ T-cell response. *Nature Reviews Immunology*, **4**, 595–602. 170

Bevan, M.J. & Fink, P.J. (2001). The CD8 response on autopilot. *Nature immunology*, **2**, 381–382. 4, 94

Bland, J.M. & Altman, D.G. (1995). Multiple significance tests: the Bonferroni method. *Bmj*, **310**, 170. 115, 199

Blond, S.L., Guillaume, J.L. & Latapy, M. (2005). Clustering in p2p exchanges and consequences on performances. In *International Workshop on Peer-to-Peer Systems*, 193–204, Springer. 171

Bock, H.G. (1983). Recent advances in parameteridentification techniques for ode. *Numerical treatment of inverse problems in differential and integral equations*, 95–121. 2

Bollobás, B. (1980). A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, **1**, 311–316. 145, 147

Bollobás, B. (2002). *Modern Graph Theory*. Springer, 1st edn. 28

Boon, T., Cerottini, J.C., van den Eynde, B., van der Bruggen, P. & Van Pel, A. (1994). Tumor antigens recognized by T lymphocytes. *Annual review of immunology*, **12**, 337–365. 143

Borg, N.A., Ely, L.K., Beddoe, T., Macdonald, W.A., Reid, H.H., Clements, C.S., Purcell, A.W., Kjer-Nielsen, L., Miles, J.J., Burrows, S.R. *et al.* (2005). The CDR3 regions of an immunodominant T cell receptor dictate the 'energetic landscape' of peptide-MHC recognition. *Nature immunology*, **6**, 171–180. 145

Borgatti, S.P. & Everett, M.G. (2000). Models of core/periphery structures. *Social networks*, **21**, 375–395. 171

Borghans, J.A., Tesselaar, K. & de Boer, R.J. (2018). Current best estimates for the average lifespans of mouse and human leukocytes: Reviewing two decades of deuterium-labeling experiments. *Immunological reviews*, **285**, 233–248. 67, 71

Bortolussi, L., Cardelli, L., Kwiatkowska, M. & Laurenti, L. (2016). Approximation of probabilistic reachability for chemical reaction networks using the linear noise approximation. In *International Conference on Quantitative Evaluation of Systems*, 72–88, Springer. 20

Boyman, O., Krieg, C., Homann, D. & Sprent, J. (2012). Homeostatic maintenance of T cells and natural killer cells. *Cellular and molecular life Sciences*, **69**, 1597–1608. 11, 32

Bradley, P. & Thomas, P.G. (2019). Using T cell receptor repertoires to understand the principles of adaptive immune recognition. *Annual Review of Immunology*, **37**, 547–570. 170

Brauer, F., Castillo-Chavez, C. & Castillo-Chavez, C. (2012). *Mathematical models in population biology and epidemiology*, vol. 2. Springer. 1

Buis, R. (1991). On the generalization of the logistic law of growth. *Acta Biotheoretica*, **39**, 185–195. 1

Burt, R. (2012). Structural holes [1992]. *Contemporary Sociological Theory*, **204**, 9780429494468–63. 171

Busch, D.H. & Pamer, E.G. (1999). T cell affinity maturation by selective expansion during infection. *The Journal of experimental medicine*, **189**, 701–710. 91, 103

Butz, E.A. & Bevan, M.J. (1998). Massive expansion of antigen-specific $CD8^+$ T cells during an acute virus infection. *Immunity*, **8**, 167–175. 91

Cameron, P.J. (1994). *Combinatorics: Topics, Techniques, Algorithms*. Cambridge University Press, Cambridge. 46

Cardelli, L., Kwiatkowska, M. & Laurenti, L. (2016). Stochastic analysis of chemical reaction networks using linear noise approximation. *Biosystems*, **149**, 26–33. 20

Carson, R.T., Vignali, K.M., Woodland, D.L. & Vignali, D.A. (1997). T cell receptor recognition of MHC class II–bound peptide flanking residues enhances immunogenicity and results in altered TCR V region usage. *Immunity*, **7**, 387–399. 145, 173

Casadevall, A. & Pirofski, L.a. (2004). New concepts in antibody-mediated immunity. *Infection and immunity*, **72**, 6191–6196. 7

Casola, S., Otipoby, K.L., Alimzhanov, M., Humme, S., Uyttersprot, N., Kutok, J.L., Carroll, M.C. & Rajewsky, K. (2004). B cell receptor signal strength determines B cell fate. *Nature immunology*, **5**, 317–327. 7

Casrouge, A., Beaudoing, E., Dalle, S., Pannetier, C., Kanellopoulos, J. & Kourilsky, P. (2000). Size estimate of the $\alpha\beta$ TCR repertoire of naive mouse splenocytes. *The Journal of Immunology*, **164**, 5782–5787. 84

Chan, C.E.Z., Chan, A., Hanson, B., Ooi, E. *et al.* (2009). The use of antibodies in the treatment of infectious diseases. *Singapore Med J*, **50**, 663–672. 7

Chang, J.T., Palanivel, V.R., Kinjyo, I., Schambach, F., Intlekofer, A.M., Banerjee, A., Longworth, S.A., Vinup, K.E., Mrass, P., Oliaro, J. *et al.* (2007). Asymmetric T lymphocyte division in the initiation of adaptive immune responses. *science*, **315**, 1687–1691. 95

Chapman, T.J., Castrucci, M.R., Padrick, R.C., Bradley, L.M. & Topham, D.J. (2005). Antigen-specific and non-specific cd4$^+$ T cell recruitment and proliferation during influenza infection. *Virology*, **340**, 296–306. 112

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. 24

Chisari, F.V. *et al.* (1997). Cytotoxic T cells and viral hepatitis. *The Journal of clinical investigation*, **99**, 1472–1477. 8

Cibotti, R., Cabaniols, J., Pannetier, C., Delarbre, C., Vergnon, I., Kanellopoulos, J.M. & Kourilsky, P. (1994). Public and private V beta T cell receptor repertoires against hen egg white lysozyme (HEL) in nontransgenic versus HEL transgenic mice. *The Journal of experimental medicine*, **180**, 861–872. 14

Cole, D.K., Edwards, E.S., Wynn, K.K., Clement, M., Miles, J.J., Ladell, K., Ekeruche, J., Gostick, E., Adams, K.J., Skowera, A. *et al.* (2010). Modification of MHC anchor residues generates heteroclitic peptides that alter TCR binding and T cell recognition. *The Journal of Immunology*, **185**, 2600–2610. 145, 173

Correia-Neves, M., Waltzinger, C., Mathis, D. & Benoist, C. (2001). The shaping of the T cell repertoire. *Immunity*, **14**, 21–32. 32

Croft, N.P., Smith, S.A., Pickering, J., Sidney, J., Peters, B., Faridi, P., Witney, M.J., Sebastian, P., Flesch, I.E., Heading, S.L. *et al.* (2019). Most viral peptides displayed by class I MHC on infected cells are immunogenic. *Proceedings of the National Academy of Sciences*, **116**, 3112–3117. 95

Cumbers, S.J., Williams, G.T., Davies, S.L., Grenfell, R.L., Takeda, S., Batista, F.D., Sale, J.E. & Neuberger, M.S. (2002). Generation and iterative affinity maturation of antibodies in vitro using hypermutating B-cell lines. *Nature biotechnology*, **20**, 1129–1134. 7

Dankelmann, P., Oellermann, O.R. & Swart, H.C. (1996). The average steiner distance of a graph. *Journal of Graph Theory*, **22**, 15–22. 171

Darroch, J.N. & Seneta, E. (1967). On quasi-stationary distributions in absorbing continuous-time finite markov chains. *Journal of Applied Probability*, **4**, 192–196. 45

Dash, P., Fiore-Gartland, A.J., Hertz, T., Wang, G.C., Sharma, S., Souquette, A., Crawford, J.C., Clemens, E.B., Nguyen, T.H., Kedzierska, K. *et al.* (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, **547**, 89–93. 171

Day, E.B., Guillonneau, C., Gras, S., La Gruta, N.L., Vignali, D.A., Doherty, P.C., Purcell, A.W., Rossjohn, J. & Turner, S.J. (2011). Structural basis for enabling T-cell receptor diversity within biased virus-specific $CD8^+$ T-cell responses. *Proceedings of the National Academy of Sciences*, **108**, 9536–9541. 145, 173

De Boer, R.J. (1988). Symmetric idiotypic networks: connectance and switching, stability, and suppression. 142

De Boer, R.J. (1989). Information processing in immune systems: Clonal selection versus idiotypic network models. In *Cell to Cell Signalling*, 285–302, Elsevier. 142

De Boer, R.J. & Perelson, A.S. (1994). T cell repertoires and competitive exclusion. *Journal of theoretical biology*, **169**, 375–390. 31

De Sá, J.P.M. (2007). *Applied statistics using SPSS, statistica, Matlab and R*. Springer Science & Business Media. 25

Demkowicz Jr, W.E., Littaua, R.A., Wang, J. & Ennis, F.A. (1996). Human cytotoxic T-cell memory: long-lived responses to vaccinia virus. *Journal of virology*, **70**, 2627–2631. 12

Deuflhard, P. (1985). Recent progress in extrapolation methods for ordinary differential equations. *SIAM review*, **27**, 505–535. 2

DeWitt III, W.S., Smith, A., Schoch, G., Hansen, J.A., Matsen IV, F.A. & Bradley, P. (2018). Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *eLife*, **7**, e38358. 170

Dianati, N. (2016). A maximum entropy approach to separating noise from signal in bimodal affiliation networks. *arXiv preprint arXiv:1607.01735*. 171

Diestel, R. (2006). *Graph theory*. Graduate Texts in Mathematics, Springer, 3rd edn. 28, 30, 153

Doherty, P.C. & Christensen, J.P. (2000). Accessing complexity: the dynamics of virus-specific T cell responses. *Annual review of immunology*, **18**, 561–592. 3, 4, 94

Duan, S., Meliopoulos, V.A., McClaren, J.L., Guo, X.Z.J., Sanders, C.J., Smallwood, H.S., Webby, R.J., Schultz-Cherry, S.L., Doherty, P.C. & Thomas, P.G. (2015). Diverse heterologous primary infections radically alter immunodominance hierarchies and clinical outcomes following H7N9 influenza challenge in mice. *PLoS pathogens*, **11**. 4, 13, 83, 91, 135, 141, 170

Eickhoff, C.S., Terry, F.E., Peng, L., Meza, K.A., Sakala, I.G., Van Aartsen, D., Moise, L., Martin, W.D., Schriewer, J., Buller, R.M. *et al.* (2019). Highly conserved influenza T cell epitopes induce broadly protective immunity. *Vaccine*, **37**, 5371–5381. 94, 143

Ekiert, D.C., Bhabha, G., Elsliger, M.A., Friesen, R.H., Jongeneelen, M., Throsby, M., Goudsmit, J. & Wilson, I.A. (2009). Antibody recognition of a highly conserved influenza virus epitope. *Science*, **324**, 246–251. 94

Elf, J. & Ehrenberg, M. (2003). Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome research*, **13**, 2475–2484. 20, 33, 50, 52, 89, 179

Elhanati, Y., Sethna, Z., Callan Jr, C.G., Mora, T. & Walczak, A.M. (2018). Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunological reviews*, **284**, 167–179. 173

Elong Ngono, A. & Shresta, S. (2019). Cross-reactive T cell immunity to dengue and Zika viruses: New insights into vaccine development. *Frontiers in immunology*, **10**, 1316. 3, 92, 135, 170

Ely, K.H., Cookenham, T., Roberts, A.D. & Woodland, D.L. (2006). Memory T cell populations in the lung airways are maintained by continual recruitment. *The Journal of Immunology*, **176**, 537–543. 108

Erdős, P. & Rényi, A. (1959). On random graphs. *Publicationes Mathematicate*, **6**, 290–297. 144

Fadai, N.T., Baker, R.E. & Simpson, M.J. (2019). Accurate and efficient discretizations for stochastic models providing near agent-based spatial resolution at low computational cost. *Journal of the Royal Society Interface*, **16**, 20190421. 2

Fahse, L., Wissel, C. & Grimm, V. (1998). Reconciling classical and individual-based approaches in theoretical population ecology: a protocol for extracting population parameters from individual-based models. *The American Naturalist*, **152**, 838–852. 2

Farber, D.L. (2009). Biochemical signaling pathways for memory T cell recall. In *Seminars in immunology*, vol. 21, 84–91, Elsevier. 100

Farrant, J., Spickett, G., Matamoros, N., Copas, D., Hernandez, M., North, M., Chapel, H. & Webster, A. (1994). Study of B and T cell phenotypes in blood from patients with common variable immunodeficiency (CVID). *Immunodeficiency*, **5**, 159–169. 3

Feliciangeli, F., Dreiwi, H., Lopez Garcia, M., Castro Ponce, M., Molina-Paris, C. & Lythe, G. (2022). Why are cell populations maintained via multiple compartments? *Journal of the Royal Society. Interface*. 1

Feller, W. (1957). *An introduction to probability theory and its applications*, vol. 1. John Wiley & Sons, New York, 2nd edn. 46

Ferreira, C., Barthlott, T., Garcia, S., Zamoyska, R. & Stockinger, B. (2000). Differential survival of naive CD4 and CD8 T cells. *The Journal of Immunology*, **165**, 3689–3694. 31

Fraser, D. & Kaern, M. (2009). A chance at survival: gene expression noise and phenotypic diversification strategies. *Molecular microbiology*, **71**, 1333–1340. 2

Freeman, L.C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41. 153, 171

Gaevert, J.A., Luque Duque, D., Lythe, G., Molina-París, C. & Thomas, P.G. (2021). Quantifying T Cell Cross-Reactivity: Influenza and Coronaviruses. *Viruses*, **13**, 1786. 141, 170

Gagnon, S.J., Borbulevych, O.Y., Davis-Harrison, R.L., Baxter, T.K., Clemens, J.R., Armstrong, K.M., Turner, R.V., Damirjian, M., Biddison, W.E. & Baker, B.M. (2005). Unraveling a hotspot for TCR recognition on HLA-A2: evidence against the existence of peptide-independent TCR binding determinants. *Journal of molecular biology*, **353**, 556–573. 145

Gaimann, M.U., Nguyen, M., Desponds, J. & Mayer, A. (2020). Early life imprints the hierarchy of T cell clone sizes. *Elife*, **9**, e61639. 85

Gaud, G., Lesourne, R. & Love, P.E. (2018). Regulatory mechanisms in T cell receptor signalling. *Nature Reviews Immunology*, **18**, 485–497. 9

Gaver, D., Jacobs, P. & Latouche, G. (1984). Finite birth-and-death models in randomly changing environments. *Advances in applied probability*, **16**, 715–731. 50, 179

Gebhardt, T., Wakim, L.M., Eidsmo, L., Reading, P.C., Heath, W.R. & Carbone, F.R. (2009). Memory T cells in nonlymphoid tissue that provide enhanced local immunity during infection with herpes simplex virus. *Nature immunology*, **10**, 524–530. 108, 132

Gerriets, V.A. & Rathmell, J.C. (2012). Metabolic pathways in T cell fate and function. *Trends in immunology*, **33**, 168–173. 100

Gil, A., Yassai, M.B., Naumov, Y.N. & Selin, L.K. (2015). Narrowing of human influenza A virus-specific T cell receptor α and β repertoires with increasing age. *Journal of virology*, **89**, 4102–4116. 4, 91

Gilbert, E.N. (1959). Random graphs. *The Annals of Mathematical Statistics*, **30**, 1141–1144. 144

Gillespie, D.T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, **22**, 403–434. 166, 205

Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, **81**, 2340–2361. 2, 21, 22, 166, 205

Gómez-Corral, A. & López-García, M. (2012a). Extinction times and size of the surviving species in a two-species competition process. *Journal of mathematical biology*, **64**, 255–289. 59

Gómez-Corral, A. & López-García, M. (2012b). On the number of births and deaths during an extinction cycle, and the survival of a certain individual in a competition process. *Computers & Mathematics with Applications*, **64**, 236–259. 59

Gómez-Corral, A. & López-García, M. (2018). Perturbation analysis in finite LD-QBD processes and applications to epidemic models. *Numerical linear algebra with applications*, **25**, e2160. 47, 179

Goronzy, J.J., Lee, W.W. & Weyand, C.M. (2007). Aging and T-cell diversity. *Experimental gerontology*, **42**, 400–406. 85

Goronzy, J.J., Fang, F., Cavanagh, M.M., Qi, Q. & Weyand, C.M. (2015). Naive T cell maintenance and function in human aging. *The Journal of Immunology*, **194**, 4073–4080. 11

Gostic, K.M., Ambrose, M., Worobey, M. & Lloyd-Smith, J.O. (2016). Potent protection against H5N1 and H7N9 influenza via childhood hemagglutinin imprinting. *Science*, **354**, 722–726. 91

Gras, S., Saulquin, X., Reiser, J.B., Debeaupuis, E., Echasserieau, K., Kissenpfennig, A., Legoux, F., Chouquet, A., Le Gorrec, M., Machillot, P. *et al.* (2009). Structural bases for the affinity-driven selection of

a public TCR against a dominant human cytomegalovirus epitope. *The Journal of Immunology*, **183**, 430–437. 145

Gras, S., Kedzierski, L., Valkenburg, S.A., Laurie, K., Liu, Y.C., Denholm, J.T., Richards, M.J., Rimmelzwaan, G.F., Kelso, A., Doherty, P.C. *et al.* (2010). Cross-reactive CD8$^+$ T-cell immunity between the pandemic H1N1-2009 and H1N1-1918 influenza A viruses. *Proceedings of the National Academy of Sciences*, **107**, 12599–12604. 13

Gratie, D.E., Iancu, B. & Petre, I. (2013). ODE analysis of biological systems. In *International School on Formal Methods for the Design of Computer, Communication and Software Systems*, 29–62, Springer. 2

Griffiths, D.F. & Higham, D.J. (2010). Euler's method. In *Numerical Methods for Ordinary Differential Equations*, 19–31, Springer. 23

Guillemin, F. & Sericola, B. (2007). Stationary analysis of a fluid queue driven by some countable state space markov chain. *Methodology and Computing in Applied Probability*, **9**, 521–540. 42

Gutman, I. & Yeh, Y.N. (1995). The sum of all distances in bipartite graphs. *Mathematica Slovaca*, **45**, 327–334. 171

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T. & Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, **34**, 4121–4123. 170

Hahl, S.K. & Kremling, A. (2016). A comparison of deterministic and stochastic modeling approaches for biochemical reaction systems: on fixed points, means, and modes. *Frontiers in genetics*, **7**, 157. 2, 22

Hardin, G. (1960). The Competitive Exclusion Principle: An idea that took a century to be born has implications in ecology, economics, and genetics. *science*, **131**, 1292–1297. 79

Harty, J.T. & Badovinac, V.P. (2008). Shaping and reshaping $CD8^+$ T-cell memory. *Nature Reviews Immunology*, **8**, 107. 12, 96, 97

Heiny, A., Miotto, O., Srinivasan, K.N., Khan, A.M., Zhang, G., Brusic, V., Tan, T.W. & August, J.T. (2007). Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets. *PloS one*, **2**, e1190. 143

Henrickson, S.E. & von Andrian, U.H. (2007). Single-cell dynamics of T-cell priming. *Current opinion in immunology*, **19**, 249–258. 3

Hillaire, M.L., Vogelzang-van Trierum, S.E., Kreijtz, J.H., de Mutsert, G., Fouchier, R.A., Osterhaus, A.D.E. & Rimmelzwaan, G.F. (2013). Human T-cells directed to seasonal influenza A virus cross-react with 2009 pandemic influenza A (H1N1) and swine-origin triple-reassortant H3N2 influenza viruses. *Journal of General Virology*, **94**, 583–592. 4, 91

Holland, P.W., Laskey, K.B. & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, **5**, 109–137. 144, 145

Hong, H., Ovchinnikov, A., Pogudin, G. & Yap, C. (2019). SIAN: software for structural identifiability analysis of ODE models. *Bioinformatics*, **35**, 2873–2874. 25, 113, 114

Hong, H., Ovchinnikov, A., Pogudin, G. & Yap, C. (2020). Global Identifiability of Differential Models. *Communications on Pure and Applied Mathematics*, **73**, 1831–1879. 4, 25, 113

Howard M, T. & Karlin, S. (1998). *An introduction to stochastic modeling*. Academic Press, Oxford, 3rd edn. 36, 44, 57, 61

Huppa, J.B., Gleimer, M., Sumen, C. & Davis, M.M. (2003). Continuous T cell receptor signaling required for synapse maintenance and full effector potential. *Nature immunology*, **4**, 749–755. 12, 95

Iglehart, D.L. (1964). Multivariate competition processes. *The Annals of Mathematical Statistics*, **35**, 350–361. 2, 55, 57, 101, 102

Ishizuka, J., Stewart-Jones, G.B., van der Merwe, A., Bell, J.I., McMichael, A.J. & Jones, E.Y. (2008). The structural dynamics and energetics of an immunodominant T cell receptor are programmed by its Vβ domain. *Immunity*, **28**, 171–182. 145

Jameson, S.C. (2002). Maintaining the norm: T-cell homeostasis. *Nature Reviews Immunology*, **2**, 547–556. 128

Jang, Y.H. & Seong, B.L. (2019). The quest for a truly universal influenza vaccine. *Frontiers in cellular and infection microbiology*, **9**, 344. 135

Jenkins, M.K., Chu, H.H., McLachlan, J.B. & Moon, J.J. (2009). On the composition of the preimmune repertoire of T cells specific for peptide–major histocompatibility complex ligands. *Annual review of immunology*, **28**, 275–294. 31

Jerne, N.K. (1974a). Clonal selection in a lymphocyte network. *Society of General Physiologists series*, **29**, 39–48. 142

Jerne, N.K. (1974b). Towards a network theory of the immune system. *Ann. Immunol.*, **125**, 373–389. 142

Jerne, N.K. (1984). Idiotypic networks and other preconceived ideas. *Immunological reviews*, **79**, 5–24. 142

Jiang, X., Clark, R.A., Liu, L., Wagers, A.J., Fuhlbrigge, R.C. & Kupper, T.S. (2012). Skin infection generates non-migratory memory CD8$^+$ TRM cells providing global skin immunity. *Nature*, **483**, 227–231. 108, 132

Joshi, N.S., Cui, W., Chandele, A., Lee, H.K., Urso, D.R., Hagman, J., Gapin, L. & Kaech, S.M. (2007). Inflammation directs memory precursor and short-lived effector CD8$^+$ T cell fates via the graded expression of T-bet transcription factor. *Immunity*, **27**, 281–295. 95

Joshi, S.K., Suresh, P.R. & Chauhan, V.S. (2001). Flexibility in MHC and TCR recognition: degenerate specificity at the T cell level in the recognition of promiscuous Th epitopes exhibiting no primary sequence homology. *The Journal of Immunology*, **166**, 6693–6703. 145, 173

Kaech, S.M. & Cui, W. (2012). Transcriptional control of effector and memory CD8$^+$ T cell differentiation. *Nature Reviews Immunology*, **12**, 749. 12, 13, 94, 95, 98

Kaern, M., Elston, T.C., Blake, W.J. & Collins, J.J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, **6**, 451–464. 2

Kanduc, D. (2012). Peptide cross-reactivity: the original sin of vaccines. *Frontiers in Bioscience-Scholar*, **4**, 1393–1401. 170

Kang, S. & Cheek, J.B. (1972). *Numerical solution of differential equations*. Waterways Experiment Station. 23

Kawabe, T., Yi, J. & Sprent, J. (2021). Homeostasis of Naive and Memory T lymphocytes. *Cold Spring Harbor Perspectives in Biology*, a037879. 31

Kedzierska, K., Day, E.B., Pi, J., Heard, S.B., Doherty, P.C., Turner, S.J. & Perlman, S. (2006). Quantification of repertoire diversity of influenza-specific epitopes with predominant public or private TCR usage. *The Journal of Immunology*, **177**, 6705–6712. 83, 173

Kendall, M.G. (1948). Rank correlation methods. 115

Kieper, W.C. & Jameson, S.C. (1999). Homeostatic expansion and phenotypic conversion of naive T cells in response to self peptide/MHC ligands. *Proceedings of the National Academy of Sciences*, **96**, 13306–13311. 31

Kim, S.K., Cornberg, M., Wang, X.Z., Chen, H.D., Selin, L.K. & Welsh, R.M. (2005). Private specificities of CD8 T cell responses control patterns of heterologous immunity. *The Journal of experimental medicine*, **201**, 523–533. 14

Kitsak, M., Papadopoulos, F. & Krioukov, D. (2017). Latent geometry of bipartite networks. *Phys. Rev. E*, **95**, 032309. 170

Kjer-Nielsen, L., Clements, C.S., Purcell, A.W., Brooks, A.G., Whisstock, J.C., Burrows, S.R., McCluskey, J. & Rossjohn, J. (2003). A structural basis for the selection of dominant $\alpha\beta$ t cell receptors in antiviral immunity. *Immunity*, **18**, 53–64. 14

Kohlmeier, J.E., Miller, S.C. & Woodland, D.L. (2007). Cutting edge: Antigen is not required for the activation and maintenance of virus-specific memory CD8$^+$ T cells in the lung airways. *The Journal of Immunology*, **178**, 4721–4725. 108

Kolmogoroff, A. (1931). Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Mathematische Annalen*, **104**, 415–458. 43

Kostitzin, V. (1940). Sur la loi logistique et ses généralisations. *Acta Biotheoretica*, **5**, 155–159. 1

Kuhns, M.S., Davis, M.M. & Garcia, K.C. (2006). Deconstructing the form and function of the TCR/CD3 complex. *Immunity*, **24**, 133–139. 9

Kulkarni, V.G. (2017). *Modeling and analysis of stochastic systems*. CRC Press, New York, 3rd edn. 47, 179

Kumar, A., Meldgaard, T.S. & Bertholet, S. (2018). Novel platforms for the development of a universal influenza vaccine. *Frontiers in immunology*, **9**, 600. 135

Kussell, E. & Leibler, S. (2005). Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, **309**, 2075–2078. 2

La Gruta, N.L., Turner, S.J. & Doherty, P.C. (2004). Hierarchies in cytokine expression profiles for acute and resolving influenza virus-specific CD8$^+$ T cell responses: correlation of cytokine profile and TCR avidity. *The Journal of Immunology*, **172**, 5553–5560. 99

Lambiotte, R. & Ausloos, M. (2005). Uncovering collective listening habits and music genres in bipartite networks. *Phys. Rev. E*, **72**, 066107. 170

Lanfermeijer, J., Borghans, J.A. & van Baarle, D. (2020). How age and infection history shape the antigen-specific CD8$^+$ T-cell repertoire: Implications for vaccination strategies in older adults. *Aging Cell*, **19**, e13262. 4, 91

Lang, H.L., Jacobsen, H., Ikemizu, S., Andersson, C., Harlos, K., Madsen, L., Hjorth, P., Sondergaard, L., Svejgaard, A., Wucherpfennig, K. *et al.* (2002). A functional and structural basis for TCR cross-reactivity in multiple sclerosis. *Nature immunology*, **3**, 940–943. 8, 141

Lanzavecchia, A. & Sallusto, F. (2001). Regulation of T cell immunity by dendritic cells. *Cell*, **106**, 263–266. 4

Lauvau, G. & Soudja, S.M. (2015). Mechanisms of memory T cell activation and effective immunity. *Crossroads Between Innate and Adaptive Immunity V*, 73–80. 7, 96

Lawrence, C.W., Ream, R.M. & Braciale, T.J. (2005). Frequency, specificity, and sites of expansion of CD8$^+$ T cells during primary pulmonary influenza virus infection. *The Journal of Immunology*, **174**, 5332–5340. 91

Li, H., Ye, C., Ji, G. & Han, J. (2012). Determinants of public T cell responses. *Cell research*, **22**, 33–42. 14

Li, S. & Song, Y. (2014). On the sum of all distances in bipartite graphs. *Discrete Applied Mathematics*, **169**, 176–185. 171

Lieber, M.R. (2010). The mechanism of double-strand DNA break repair by the nonhomologous DNA end joining pathway. *Annual review of biochemistry*, **79**, 181. 9

Lieber, M.R. & Wilson, T.E. (2010). SnapShot: Nonhomologous DNA end joining (NHEJ). *Cell*, **142**, 496–496. 9

Lind, P.G., González, M.C. & Herrmann, H.J. (2005). Cycles and clustering in bipartite networks. *Physical review E*, **72**, 056127. 172

Lotka, A.J. (1925). *Elements of physical biology*. Williams & Wilkins. 2

Luciani, F., Sanders, M.T., Oveissi, S., Pang, K.C. & Chen, W. (2013). Increasing viral dose causes a reversal in $CD8^+$ T cell immunodominance during primary influenza infection due to differences in antigen presentation, T cell avidity, and precursor numbers. *The Journal of Immunology*, **190**, 36–47. 95

Luckey, M.A., Kimura, M.Y., Waickman, A.T., Feigenbaum, L., Singer, A. & Park, J.H. (2014). The transcription factor ThPOK suppresses Runx3 and imposes $CD4^+$ lineage fate by inducing the SOCS suppressors of cytokine signaling. *Nature immunology*, **15**, 638–645. 11

Luque, D. (2021). WolfDanny/X-reactivity-review: Quantifying cross-reactivity: influenza and coronaviruses. v, 205

Luque, D. (2022a). WolfDanny/Homeostatic_Competition: Homeostatic competition python codes. iv, 179

Luque, D. (2022b). WolfDanny/IAV_analysis: v1.0. iv, 197

Lynch, H.E., Goldberg, G.L., Chidgey, A., Van den Brink, M.R., Boyd, R. & Sempowski, G.D. (2009). Thymic involution and immune reconstitution. *Trends in immunology*, **30**, 366–373. 85

Lythe, G., Callard, R.E., Hoare, R.L. & Molina-París, C. (2016). How many TCR clonotypes does a body maintain? *Journal of Theoretical Biology*, **389**, 214–224. 32, 67, 71, 84

Macallan, D.C., Asquith, B., Irvine, A.J., Wallace, D.L., Worth, A., Ghattas, H., Zhang, Y., Griffin, G.E., Tough, D.F. & Beverley, P.C. (2003). Measurement and modeling of human T cell kinetics. *European journal of immunology*, **33**, 2316–2326. 137

Macallan, D.C., Busch, R. & Asquith, B. (2019). Current estimates of T cell kinetics in humans. *Current opinion in systems biology*, **18**, 77–86. 137

MacDonald, H.R., Casanova, J.L., Maryanski, J.L. & Cerottini, J.C. (1993). Oligoclonal expansion of major histocompatibility complex class I-restricted cytolytic T lymphocytes during a primary immune response in

vivo: direct monitoring by flow cytometry and polymerase chain reaction. *The Journal of experimental medicine*, **177**, 1487–1492. 32

Mackay, C.R., Marston, W.L. & Dudler, L. (1990). Naive and memory T cells show distinct pathways of lymphocyte recirculation. *The Journal of experimental medicine*, **171**, 801–817. 11

Malthus, T.R. (1986). An essay on the principle of population (1798). *The Works of Thomas Robert Malthus, London, Pickering & Chatto Publishers*, **1**, 1–139. 1

Manly, B.F. & Alberto, J.A.N. (2016). *Multivariate statistical methods: a primer*. Chapman and Hall/CRC, 4th edn. 119

Mantegazza, A.R., Magalhaes, J.G., Amigorena, S. & Marks, M.S. (2013). Presentation of phagocytosed antigens by MHC class I and II. *Traffic*, **14**, 135–152. 8

Marcus, M. & Khan, N.A. (1959). A note on the hadamard product. *Canadian Mathematical Bulletin*, **2**, 81–83. 146

Maryanski, J.L., Casanova, J.L., Falk, K., Gournier, H., Jaulin, C., Kourilsky, P., Lemonnier, F.A., Lüthy, R., Rammensee, H.G., Rötzschke, O. *et al.* (1997). The diversity of antigen-specific tcr repertoires reflects the relative complexity of epitopes recognized. *Human immunology*, **54**, 117–128. 8

Mason, D. (1998). A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunology today*, **19**, 395–404. 5, 8, 13, 141, 177

Masopust, D. & Soerens, A.G. (2019). Tissue-resident T cells and other resident leukocytes. *Annual review of immunology*, **37**, 521. 108, 132

Mateus, J., Dan, J.M., Zhang, Z., Rydyznski Moderbacher, C., Lammers, M., Goodwin, B., Sette, A., Crotty, S. & Weiskopf, D. (2021). Low-dose mRNA-1273 COVID-19 vaccine generates durable memory enhanced by cross-reactive T cells. *Science*, **374**, eabj9853. 3, 92, 135, 170

Mayer-Blackwell, K., Schattgen, S., Cohen-Lavi, L., Crawford, J.C., Souquette, A., Gaevert, J.A., Hertz, T., Thomas, P.G., Bradley, P. & Fiore-Gartland, A. (2020). TCR meta-clonotypes for biomarker discovery with tcrdist3: quantification of public, HLA-restricted TCR biomarkers of SARS-CoV-2 infection. *bioRxiv*. 171

Mazza, C., Auphan-Anezin, N., Gregoire, C., Guimezanes, A., Kellenberger, C., Roussel, A., Kearney, A., van der Merwe, P.A., Schmitt-Verhulst, A.M. & Malissen, B. (2007). How much can a T-cell antigen receptor adapt to structurally distinct antigenic peptides? *The EMBO journal*, **26**, 1972–1983. 141

McMichael, A.J., Gotch, F.M., Noble, G.R. & Beare, P.A. (1983). Cytotoxic T-cell immunity to influenza. *New England Journal of Medicine*, **309**, 13–17. 4, 91

McShane, E.J. (1937). Jensen's inequality. *Bulletin of the American Mathematical Society*, **43**, 521–527. 162

Metzig, C. & Colijn, C. (2018). Preferential attachment in systems and networks of constant size. *arXiv preprint arXiv:1811.04972*. 151

Miles, J.J., Elhassen, D., Borg, N.A., Silins, S.L., Tynan, F.E., Burrows, J.M., Purcell, A.W., Kjer-Nielsen, L., Rossjohn, J., Burrows, S.R. *et al.* (2005). CTL recognition of a bulged viral peptide involves biased TCR selection. *The Journal of Immunology*, **175**, 3826–3834. 14

Moler, C. & Van Loan, C. (1978). Nineteen dubious ways to compute the exponential of a matrix. *SIAM review*, **20**, 801–836. 20

Moler, C. & Van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, **45**, 3–49. 20

Molloy, M. & Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, **6**, 161–180. 147

Mondino, A., Khoruts, A. & Jenkins, M.K. (1996). The anatomy of T-cell activation and tolerance. *Proceedings of the National Academy of Sciences*, **93**, 2245–2252. 12

Montgomery, D.C. (2017). *Design and analysis of experiments*. John wiley & sons. 25

Moris, P., van der Most, R., Leroux-Roels, I., Clement, F., Dramé, M., Hanon, E., Leroux-Roels, G.G. & Van Mechelen, M. (2011). H5N1 influenza vaccine formulated with AS03A induces strong cross-reactive and polyfunctional CD4 T-cell responses. *Journal of clinical immunology*, **31**, 443–454. 3, 92, 135, 170

Morris, S., Yen, G.G. *et al.* (2005). Construction of bipartite and unipartite weighted networks from collections of journal papers. *arXiv preprint physics/0503061*. 171

Murphy, G.M. (2011). *Ordinary differential equations and their solutions*. Courier Corporation. 23

Murphy, K. & Weaver, C. (2016). *Janeway's immunobiology*. Garland science. 3, 7, 10, 12, 175

Murugan, A., Mora, T., Walczak, A.M. & Callan, C.G. (2012). Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, **109**, 16161–16166. 10, 32

Nåsell, I. (1991). On the quasi-stationary distribution of the Ross malaria model. *Mathematical Biosciences*, **107**, 187–207. 33, 42, 45, 46, 179

Nåsell, I. (2001). Extinction and quasi-stationarity in the Verhulst logistic model. *Journal of Theoretical Biology*, **211**, 11–27. 33, 45, 46, 179

Naylor, K., Li, G., Vallejo, A.N., Lee, W.W., Koetz, K., Bryl, E., Witkowski, J., Fulbright, J., Weyand, C.M. & Goronzy, J.J. (2005). The influence of age on T cell generation and TCR diversity. *The Journal of Immunology*, **174**, 7446–7452. 32

Nelson, R.W., Beisang, D., Tubo, N.J., Dileepan, T., Wiesner, D.L., Nielsen, K., Wüthrich, M., Klein, B.S., Kotov, D.I., Spanier, J.A. *et al.* (2015). T cell receptor cross-reactivity between similar foreign and self peptides influences naive cell population size and autoimmunity. *Immunity*, **42**, 95–107. 141

Newell, E.W., Sigal, N., Bendall, S.C., Nolan, G.P. & Davis, M.M. (2012). Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of $CD8^+$ T cell phenotypes. *Immunity*, **36**, 142–152. 3

Newman, M. (2018). *Networks*. Oxford university press. 28, 143, 148, 153, 171

Newman, M.E. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, **69**, 026113. 153

Nikolich-Žugich, J., Slifka, M.K. & Messaoudi, I. (2004). The many important facets of T-cell repertoire diversity. *Nature Reviews Immunology*, **4**, 123–132. 8, 31, 83

Okhrimenko, A., Grün, J.R., Westendorf, K., Fang, Z., Reinke, S., von Roth, P., Wassilew, G., Kühl, A.A., Kudernatsch, R., Demski, S. *et al.* (2014). Human memory T cells from the bone marrow are resting and maintain long-lasting systemic memory. *Proceedings of the National Academy of Sciences*, **111**, 9229–9234. 12

Oosterhoff, J. & van Zwet, W.R. (2012). A note on contiguity and Hellinger distance. In *Selected Works of Willem van Zwet*, 63–72, Springer. 87, 179, 182

Owens III, D.E. & Peppas, N.A. (2006). Opsonization, biodistribution, and pharmacokinetics of polymeric nanoparticles. *International journal of pharmaceutics*, **307**, 93–102. 8

Palmer, E. (2003). Negative selection—clearing out the bad apples from the T-cell repertoire. *Nature Reviews Immunology*, **3**, 383–391. 11

Park, J.E., Jardine, L., Gottgens, B., Teichmann, S.A. & Haniffa, M. (2020). Prenatal development of human immunity. *Science*, **368**, 600–603. 85

Park, J.H., Adoro, S., Guinter, T., Erman, B., Alag, A.S., Catalfamo, M., Kimura, M.Y., Cui, Y., Lucas, P.J., Gress, R.E. *et al.* (2010). Signaling by intrathymic cytokines, not T cell antigen receptors, specifies $CD8^+$ lineage choice and promotes the differentiation of cytotoxic-lineage T cells. *Nature immunology*, **11**, 257–264. 11

Patra, P. & Klumpp, S. (2013). Population dynamics of bacterial persistence. *PLoS One*, **8**, e62814. 2

Paules, C.I., Sullivan, S.G., Subbarao, K. & Fauci, A.S. (2018). Chasing seasonal influenza—the need for a universal influenza vaccine. *New England Journal of Medicine*, **378**, 7–9. 135

Pepper, M. & Jenkins, M.K. (2011). Origins of $CD4^+$ effector and central memory T cells. *Nature immunology*, **12**, 467–471. 13

Perelson, A.S. & Oster, G.F. (1979). Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *Journal of theoretical biology*, **81**, 645–670. 142

Pinsky, M. & Karlin, S. (2010). *An introduction to stochastic modeling*. Academic press, 4th edn. 17

Price, D.d.S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science*, **27**, 292–306. 150, 151

Price, D.J.D.S. (1965). Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science*, **149**, 510–515. 150

Qi, H., Chen, X., Chu, C., Lu, P., Xu, H. & Yan, J. (2014a). Follicular T-helper cells: controlled localization and cellular interactions. *Immunology and Cell Biology*, **92**, 28–33. 8

Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J.Y., Olshen, R.A., Weyand, C.M., Boyd, S.D. & Goronzy, J.J. (2014b). Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences*, **111**, 13139–13144. 8, 31, 83, 85

Reiser, J.B., Darnault, C., Grégoire, C., Mosser, T., Mazza, G., Kearney, A., van der Merwe, P.A., Fontecilla-Camps, J.C., Housset, D. & Malissen, B. (2003). CDR3 loop flexibility contributes to the degeneracy of TCR recognition. *Nature immunology*, **4**, 241–247. 9

Robins, G. & Alexander, M. (2004). Small worlds among interlocking directors: Network structure and distance in bipartite graphs. *Computational & Mathematical Organization Theory*, **10**, 69–94. 153, 170

Robins, H.S., Campregher, P.V., Srivastava, S.K., Wacher, A., Turtle, C.J., Kahsai, O., Riddell, S.R., Warren, E.H. & Carlson, C.S. (2009). Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood, The Journal of the American Society of Hematology*, **114**, 4099–4107. 8, 31

Rock, K.L., Reits, E. & Neefjes, J. (2016). Present yourself! By MHC class I and MHC class II molecules. *Trends in immunology*, **37**, 724–737. 8

Rosenblum, M.D., Way, S.S. & Abbas, A.K. (2016). Regulatory T cell memory. *Nature Reviews Immunology*, **16**, 90–101. 12, 14

Rudd, B.D., Venturi, V., Li, G., Samadder, P., Ertelt, J.M., Way, S.S., Davenport, M.P. & Nikolich-Žugich, J. (2011). Nonrandom attrition of the naive $CD8^+$ T-cell pool with aging governed by T-cell receptor:pMHC interactions. *Proceedings of the National Academy of Sciences*, **108**, 13694–13699. 11, 32

Rvachev, L.A. (1968). Modelling experiment of a large-scale epidemic by means of a computer. *Doklady Akademii Nauk*, **180**, 294–296. 1

Saracco, F., Straka, M.J., Di Clemente, R., Gabrielli, A., Caldarelli, G. & Squartini, T. (2017). Inferring monopartite projections of bipartite networks: an entropy-based approach. *New Journal of Physics*, **19**, 053022. 171

Sarkar, S., Kalia, V., Haining, W.N., Konieczny, B.T., Subramaniam, S. & Ahmed, R. (2008). Functional and genomic profiling of effector CD8 T cell subsets with distinct memory fates. *The Journal of experimental medicine*, **205**, 625–640. 95

Schattgen, S.A., Guion, K., Crawford, J.C., Souquette, A., Barrio, A.M., Stubbington, M.J., Thomas, P.G. & Bradley, P. (2020). Linking T cell receptor sequence to transcriptional profiles with clonotype neighbor graph analysis (CoNGA). *bioRxiv*. 170

Schatz, D.G. & Ji, Y. (2011). Recombination centres and the orchestration of V(D)J recombination. *Nature Reviews Immunology*, **11**, 251–263. 9

Schatz, D.G. & Swanson, P.C. (2011). V(D)J recombination: mechanisms of initiation. *Annual review of genetics*, **45**, 167–202. 9, 10

Seddon, B. & Zamoyska, R. (2002). TCR signals mediated by Src family kinases are essential for the survival of naive T cells. *The Journal of Immunology*, **169**, 2997–3005. 31

Selin, L.K., Nahill, S.R. & Welsh, R.M. (1994). Cross-reactivities in memory cytotoxic T lymphocyte recognition of heterologous viruses. *The Journal of experimental medicine*, **179**, 1933–1943. 8, 141

Selin, L.K., Brehm, M.A., Naumov, Y.N., Cornberg, M., Kim, S.K., Clute, S.C. & Welsh, R.M. (2006). Memory of mice and men: $CD8^+$ T-cell cross-reactivity and heterologous immunity. *Immunological reviews*, **211**, 164–181. 170

Sewell, A.K. (2012). Why must T cells be cross-reactive? *Nature Reviews Immunology*, **12**, 669–677. 8, 13, 141

Shinar, G., Milo, R., Martínez, M.R. & Alon, U. (2007). Input–output robustness in simple bacterial signaling systems. *Proceedings of the National Academy of Sciences*, **104**, 19931–19935. 1

Simon, H.A. (1955). On a class of skew distribution functions. *Biometrika*, **42**, 425–440. 151

Simoni, G., Vo, H.T., Priami, C. & Marchetti, L. (2020). A comparison of deterministic and stochastic approaches for sensitivity analysis in computational systems biology. *Briefings in bioinformatics*, **21**, 527–540. 2

Simpson, M.L., Cox, C.D., Allen, M.S., McCollum, J.M., Dar, R.D., Karig, D.K. & Cooke, J.F. (2009). Noise in biological circuits. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology*, **1**, 214–225. 2

Singer, A., Adoro, S. & Park, J.H. (2008). Lineage fate and intense debate: myths, models and mechanisms of CD4-versus CD8-lineage choice. *Nature Reviews Immunology*, **8**, 788–801. 11

Souquette, A. & Thomas, P.G. (2018). Past life and future effects—How heterologous infections alter immunity to influenza viruses. *Frontiers in immunology*, **9**, 1071. 141, 170

Sprent, J. & Surh, C.D. (2011). Normal T cell homeostasis: the conversion of naive cells into memory-phenotype cells. *Nature immunology*, **12**, 478–484. 11, 32

Sridhar, S., Begom, S., Bermingham, A., Hoschler, K., Adamson, W., Carman, W., Bean, T., Barclay, W., Deeks, J.J. & Lalvani, A. (2013). Cellular immune correlates of protection against symptomatic pandemic influenza. *Nature medicine*, **19**, 1305–1312. 4, 91

Srinivasan, J., Lancaster, J.N., Singarapu, N., Hale, L.P., Ehrlich, L.I. & Richie, E.R. (2021). Age-related changes in thymic central tolerance. *Frontiers in immunology*, **12**, 676236. 85

Staneková, Z. & Varečková, E. (2010). Conserved epitopes of influenza A virus inducing protective immunity and their prospects for universal vaccine development. *Virology journal*, **7**, 1–13. 112

Starr, T.K., Jameson, S.C. & Hogquist, K.A. (2003). Positive and negative selection of T cells. *Annual review of immunology*, **21**, 139–176. 11

Stewart-Jones, G.B., McMichael, A.J., Bell, J.I., Stuart, D.I. & Jones, E.Y. (2003). A structural basis for immunodominant human T cell receptor recognition. *Nature immunology*, **4**, 657–663. 14, 145, 173

Stirk, E.R., Molina-París, C. & van den Berg, H.A. (2008). Stochastic niche structure and diversity maintenance in the T cell repertoire. *Journal of theoretical biology*, **255**, 237–249. 2, 3, 32, 33, 34, 36, 39, 40, 41, 70, 71

Stirk, E.R., Lythe, G., van den Berg, H.A. & Molina-París, C. (2010). Stochastic competitive exclusion in the maintenance of the naïve T cell repertoire. *Journal of theoretical biology*, **265**, 396–410. 2, 3, 32, 33, 34, 83, 175

Surh, C.D. & Sprent, J. (2005). Regulation of mature T cell homeostasis. *Seminars in immunology*, **17**, 183–191. 11, 32

Surh, C.D. & Sprent, J. (2008). Homeostasis of naive and memory T cells. *Immunity*, **29**, 848–862. 11, 98

Takada, K. & Jameson, S.C. (2009). Naive T cell homeostasis: from awareness of space to a sense of place. *Nature Reviews Immunology*, **9**, 823–832. 11

Tamura, S.i., Tanimoto, T. & Kurata, T. (2005). Mechanisms of broad cross-protection provided by influenza virus infection and their application to vaccines. *Japanese journal of infectious diseases*, **58**, 195. 135

Tan, J.T., Ernst, B., Kieper, W.C., LeRoy, E., Sprent, J. & Surh, C.D. (2002). Interleukin (IL)-15 and IL-7 jointly regulate homeostatic proliferation of memory phenotype $CD8^+$ cells but are not required for memory phenotype $CD4^+$ cells. *The Journal of experimental medicine*, **195**, 1523–1532. 98

Tan, P.T., Khan, A.M. & August, J.T. (2011). Highly conserved influenza A sequences as T cell epitopes-based vaccine targets to address the viral variability. *Human vaccines*, **7**, 402–409. 94, 143

Thompson, C.B. (1995). New insights into V(D)J recombination and its role in the evolution of the immune system. *Immunity*, **3**, 531–539. 9

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature*, **302**, 575–581. 7

Topham, D.J., Tripp, R.A. & Doherty, P.C. (1997). $CD8^+$ T cells clear influenza virus by perforin or Fas-dependent processes. *The Journal of Immunology*, **159**, 5197–5200. 108

Trotter, J., DeJong, L.J. & Smith, M.E. (1986). Opsonization with antimyelin antibody increases the uptake and intracellular metabolism of myelin in inflammatory macrophages. *Journal of neurochemistry*, **47**, 779–789. 8

Tsimring, L.S. (2014). Noise in biology. *Reports on Progress in Physics*, **77**, 026601. 2

Tukey, J.W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, **5**, 99–114. 119

Tumminello, M., Micciche, S., Lillo, F., Piilo, J. & Mantegna, R.N. (2011). Statistically validated networks in bipartite complex systems. *PloS one*, **6**, e17994. 171

Turner, S.J., Kedzierska, K., Komodromou, H., La Gruta, N.L., Dunstone, M.A., Webb, A.I., Webby, R., Walden, H., Xie, W., McCluskey, J. *et al.* (2005). Lack of prominent peptide–major histocompatibility complex features limits repertoire diversity in virus-specific $CD8^+$ T cell populations. *Nature immunology*, **6**, 382–389. 145, 173

Turner Jr, M.E., Blumenstein, B.A. & Sebaugh, J.L. (1969). 265 note: A generalization of the logistic law of growth. *Biometrics*, 577–580. 1

Turner Jr, M.E., Bradley Jr, E.L., Kirk, K.A. & Pruitt, K.M. (1976). A theory of growth. *Mathematical Biosciences*, **29**, 367–373. 1

Tynan, F.E., Borg, N.A., Miles, J.J., Beddoe, T., El-Hassen, D., Silins, S.L., van Zuylen, W.J., Purcell, A.W., Kjer-Nielsen, L., McCluskey, J. *et al.* (2005). High resolution structures of highly bulged viral epitopes bound to major histocompatibility complex class I: implications for T-cell receptor engagement and T-cell immunodominance. *Journal of Biological Chemistry*, **280**, 23900–23909. 14

Tynan, F.E., Reid, H.H., Kjer-Nielsen, L., Miles, J.J., Wilce, M.C., Kostenko, L., Borg, N.A., Williamson, N.A., Beddoe, T., Purcell, A.W. *et al.* (2007). AT cell receptor flattens a bulged antigenic peptide presented by a major histocompatibility complex class I molecule. *Nature immunology*, **8**, 268–276. 14

Udny Yule, G. (1925). A mathematical theory of evolution, based on the conclusions of Dr. JC Willis. *Philosophical Transactions of the Royal Society of London Series B*, **213**, 21–87. 151

van den Berg, H., Rand, D. & Burroughs, N. (2001). A reliable and safe T cell repertoire based on low-affinity T cell receptors. *Journal of Theoretical Biology*, **209**, 465–486. 34

van den Berg, H.A., Molina-París, C. & Sewell, A.K. (2011). Specific T-cell activation in an unspecific T-cell repertoire. *Science progress*, **94**, 245–264. 8, 141

van den Boorn, J.G., Caroline Le Poole, I. & Luiten, R.M. (2006). T-cell avidity and tuning: the flexible connection between tolerance and autoimmunity. *International reviews of immunology*, **25**, 235–258. 99

van der Bruggen, P., Zhang, Y., Chaux, P., Stroobant, V., Panichelli, C., Schultz, E.S., Chapiro, J., van den Eynde, B.J., Brasseur, F. & Boon, T. (2002). Tumor-specific shared antigenic peptides recognized by human T cells. *Immunological reviews*, **188**, 51–64. 143

van Doorn, E.A. & Scheinhardt, W.R. (1997). A fluid queue driven by an infinite-state birth-death process. In *Teletraffic Contributions for the Information Age*, vol. 2 of *Teletraffic Science and Engineering*, 465–475, Elsevier. 42

van Kampen, N.G. (2007). *Stochastic processes in physics and chemistry*, vol. 1. Elsevier, 3rd edn. 20, 21, 33, 50, 51, 53, 54, 87, 89, 92, 103, 179

van Stipdonk, M.J., Lemmens, E.E. & Schoenberger, S.P. (2001). Naive CTLs require a single brief period of antigenic stimulation for clonal expansion and differentiation. *Nature immunology*, **2**, 423–429. 93, 95

Varma, R. (2008). TCR triggering by the pMHC complex: valency, affinity, and dynamics. *Sci. Signal.*, **1**, pe21–pe21. 7, 8, 31

Venturi, V., Kedzierska, K., Price, D.A., Doherty, P.C., Douek, D.C., Turner, S.J. & Davenport, M.P. (2006). Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proceedings of the National Academy of Sciences*, **103**, 18691–18696. 15

Venturi, V., Price, D.A., Douek, D.C. & Davenport, M.P. (2008). The molecular basis for public T-cell responses? *Nature Reviews Immunology*, **8**, 231–238. 14

Verhulst, P.F. (1838). Notice sur la loi que la population suit dans son accroissement. *Corresp. Math. Phys.*, **10**, 113–126. 1

Verhulst, P.F. (1845). Resherches mathematiques sur la loi d'accroissement de la population. *Nouveaux memoires de l'academie royale des sciences*, **18**, 1–41. 1

Verkaik, N.S., Esveldt-van Lange, R.E., Heemst, D.v., Brüggenwirth, H.T., Hoeijmakers, J.H., Zdzienicka, M.Z. & Gent, D.C.v. (2002). Different types of V(D)J recombination and end-joining defects in DNA double-strand break repair mutant mammalian cells. *European journal of immunology*, **32**, 701–709. 9

Vignali, D.A., Collison, L.W. & Workman, C.J. (2008). How regulatory T cells work. *Nature reviews immunology*, **8**, 523–532. 13

Volterra, V. (1926). Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Memor. Accad. Lincei.*, **6**, 31–113. 2

Vrisekoop, N., den Braber, I., de Boer, A.B., Ruiter, A.F., Ackermans, M.T., van der Crabben, S.N., Schrijver, E.H., Spierenburg, G., Sauerwein, H.P., Hazenberg, M.D. *et al.* (2008). Sparse production but preferential incorporation of recently produced naive T cells in the human peripheral pool. *Proceedings of the National Academy of Sciences*, **105**, 6115–6120. 137

Vujovic, M., Degn, K.F., Marin, F.I., Schaap-Johansen, A.L., Chain, B., Andresen, T.L., Kaplinsky, J. & Marcatili, P. (2020). T cell receptor sequence clustering and antigen specificity. *Computational and Structural Biotechnology Journal*, **18**, 2166–2173. 173

Wang, Y.J. & Wong, G.Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, **82**, 8–19. 144, 145

Watts, D.J. & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *nature*, **393**, 440–442. 153, 154, 155

Weaver, C.T., Hatton, R.D., Mangan, P.R. & Harrington, L.E. (2007). IL-17 family cytokines and the expanding diversity of effector T cell lineages. *Annual review of immunology*, **25**, 821–852. 8

Webster, R.G. & Askonas, B.A. (1980). Cross-protection and cross-reactive cytotoxic T cells induced by influenza virus vaccines in mice. *European journal of immunology*, **10**, 396–401. 3, 92, 135, 170

Weinreich, M.A. & Hogquist, K.A. (2008). Thymic emigration: when and how T cells leave home. *The journal of immunology*, **181**, 2265–2270. 11

Wherry, E.J. & Ahmed, R. (2004). Memory CD8 T-cell differentiation during viral infection. *Journal of virology*, **78**, 5535–5545. 4

Wilmes, S., Jeffrey, P.A., Martinez-Fabregas, J., Hafer, M., Fyfe, P.K., Pohler, E., Gaggero, S., López-García, M., Lythe, G., Taylor, C. *et al.* (2021). Competitive binding of STATs to receptor phospho-Tyr motifs accounts for altered cytokine responses. *Elife*, **10**, e66014. 1

Wynn, K.K., Fulton, Z., Cooper, L., Silins, S.L., Gras, S., Archbold, J.K., Tynan, F.E., Miles, J.J., McCluskey, J., Burrows, S.R. *et al.* (2008). Impact of clonal competition for peptide-MHC complexes on the $CD8^+$ t-cell repertoire selection in a persistent viral infection. *Blood, The Journal of the American Society of Hematology*, **111**, 4283–4292. 32

Yager, E.J., Ahmed, M., Lanzer, K., Randall, T.D., Woodland, D.L. & Blackman, M.A. (2008). Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus. *The Journal of experimental medicine*, **205**, 711–723. 83

Yang, J., Zou, M., Chu, X., Floess, S., Li, Y., Delacher, M. & Huehn, J. (2022). Inflammatory perturbations in early life long-lastingly shape the transcriptome and TCR repertoire of the first wave of regulatory T cells. *Frontiers in immunology*, **13**, 991671. 4, 91

Yang, S.J. (2005). Efficient algorithms to solve the link-orientation problem for multi-square, convex-bipartite, and convex-split networks. *Information Sciences*, **171**, 475–493. 173

Yewdell, J.W. & Bennink, J.R. (1999). Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annual review of immunology*, **17**, 51. 169

Yewdell, J.W. & Hill, A.B. (2002). Viral interference with antigen presentation. *Nature immunology*, **3**, 1019–1025. 12

Yin, Y. & Mariuzza, R.A. (2009). The multiple mechanisms of T cell receptor cross-reactivity. *Immunity*, **31**, 849–851. 8, 141

Zarnitsyna, V., Evavold, B., Schoettle, L., Blattman, J. & Antia, R. (2013). Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Frontiers in immunology*, **4**, 485. 32

Zarnitsyna, V.I., Handel, A., McMaster, S.R., Hayward, S.L., Kohlmeier, J.E. & Antia, R. (2016). Mathematical model reveals the role of memory CD8 T cell populations in recall responses to influenza. *Frontiers in immunology*, **7**, 165. 1

Zens, K.D., Chen, J.K. & Farber, D.L. (2016). Vaccine-generated lung tissue–resident memory T cells provide heterosubtypic protection to influenza infection. *JCI insight*, **1**. 135

Zhang, H., Weyand, C.M. & Goronzy, J.J. (2021). Hallmarks of the aging T-cell system. *The FEBS journal*, **288**, 7123–7142. 85

Zhang, P., Wang, J., Li, X., Li, M., Di, Z. & Fan, Y. (2008). Clustering coefficient and community structure of bipartite networks. *Physica A: Statistical Mechanics and its Applications*, **387**, 6869–6875. 143, 154, 172, 177

Zhao, Y., Nguyen, P., Ma, J., Wu, T., Jones, L.L., Pei, D., Cheng, C. & Geiger, T.L. (2016). Preferential use of public TCR during autoimmune encephalomyelitis. *The Journal of Immunology*, **196**, 4905–4914. 173

Zhu, J. & Paul, W.E. (2008). CD4 T cells: fates, functions, and faults. *Blood, The Journal of the American Society of Hematology*, **112**, 1557–1569. 8

Ziemssen, T. & Ziemssen, F. (2005). The role of the humoral immune system in multiple sclerosis (MS) and its animal model experimental autoimmune encephalomyelitis (EAE). *Autoimmunity reviews*, **4**, 460–467. 8