Mathematical Modelling of Cellular Receptor-Ligand Dynamics

Polly-Anne Jeffrey

Department of Mathematics

University of Leeds

A thesis submitted for the degree of

Doctor of Philosophy

January, 2022

The candidate confirms that the work submitted is her 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 where reference has been made to the work of others.

Joint publications

Almost all of the work in Chapter 3 has been refereed and published, as follows:

- **Jeffrey, P.A**, López-García, M, Castro, M, Lythe, G and Molina-París, C, (2020). On Exact and Approximate Approaches for Stochastic Receptor-Ligand Competition Dynamics - An Ecological Perspective. *Mathematics*, 8(6), p.1014.

Code availability: GitHub release of codes for Chapter 3

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

- Wilmes, S*, **Jeffrey, P.A***, Martinez-Fabregas, J, Hafer, M, Fyfe, P.K, Pohler, E, Gaggero, S, López-García, M, Lythe, G, Taylor, C, Guerrier, T *et al*, (2021). Competitive binding of STATs to receptor phospho-Tyr motifs accounts for altered cytokine responses. *Elife*, 10, p.e66014, (*: first co-authorship).

Code availability: GitHub release of codes for Chapter 4

Finally, some of the work in Chapter 5 has been accepted for publication, as follows:

- Lin, C.C, Suen, K.M*, **Jeffrey, P.A***, Wieteska, L, Stainthorp, A, Seiler, C, Koss, H, Molina-París, C, Miska, E, Ahmed, Z and Ladbury, J.E. Receptor tyrosine kinases regulate signal transduction through a liquid–liquid phase separated state. *Molecular Cell, accepted*, (*: second co-authorship).

Acknowledgements

This research was funded by the Engineering and Physical Science Research Council Doctoral Training Grant CASE Studentship, provided jointly by the University of Leeds and AstraZeneca UK Limited, and was coordinated by the Smith Institute.

First and foremost, I would like to thank my supervisors, Martín López-García, Carmen Molina-París and Grant Lythe for their continuous support and inspiring enthusiasm for mathematical biology. Their advice has been invaluable throughout the past four years and I am hugely appreciative of their patience and guidance.

From the University of Dundee I would like to thank Ignacio Moraga and Stephan Wilmes for the sharing of their experimental data and guidance in the development of the mathematical models in Chapter 4 of this thesis.

From the Faculty of Biological Sciences at the University of Leeds I would like to thank John Ladbury for his co-supervision, and Chi-Chuan Lin and Kin Man Suen for the sharing of experimental results and advice on Chapter 5 of this thesis.

From AstraZeneca I would like to thank Ian Barrret, for his co-supervision and support with Chapter 6 of this thesis, and Ana Narvaez for the sharing of her experimental data, used in Chapter 6. Also, James Yates for technical discussions and Domingo Salazar for his co-supervision over the first year of this project.

I would like to thank Charles Taylor, Mario Castro, John Paul Gosling, Priya Subramanian, Alastair Rucklidge, Jonathan Ward, Jonathan

Carruthers and Maria Nowicka for useful discussions and advice throughout my project.

I would also like to thank my PhD examiners, Rob Sturman and Martin Meier-Schellersheim for taking the time to read this thesis, and for the useful and interesting discussions in the viva.

I give thanks to the current and past PhD students in the group of mathematical biology and medicine, and others from the school, who I have worked alongside, and become friends with, over the past four years.

Finally, I give thanks to my family and friends, most notably to my parents and to Rochelle who have supported me and believed in me throughout my studies, and without whom this work would not have been possible.

Abstract

In this thesis, mathematical models are presented which are used to study cellular signalling pathways, initiated by the interaction between receptor molecules, residing on the surface of a cell, and ligand molecules, diffusing in the extra-cellular medium. Cell signalling pathways ultimately determine the *fate* of a cell, be it to divide, differentiate, migrate or even die, and the eventuality with occurs can be dependent on the receptor-ligand pairing. Different mathematical and statistical techniques are employed in this thesis, to study specific cell signalling pathways for which experimental data has been produced. Data collection has been carried out for such pathways due to their involvement in human disease when dysregulated, where disease progressions include autoimmune diseases and many types of cancer. As well as analysing mathematical models for specific pathways, new methodologies are developed in this thesis to analyse properties of a general model of the competition between multiple receptor types for a common ligand variety, which could be applicable to many cell types and signalling pathways. Both deterministic and stochastic models are used in this thesis, since the expression of different receptor and ligand types can be hugely variable, ranging from low copy numbers per cell, to very high copy numbers. This thesis aims to explore the role of receptors and ligands in different healthy and disease scenarios.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Biological introduction

The average human body is made up of approximately 10^{13} cells (Bianconi *et al.*, 2013). The vast majority of these cells do not survive for the lifetime of a human, but are constantly being turned over, with a rate dependent on the cell type. For example, some cells with the smallest turnover rate are gut epithelial cells, which have an average lifespan of only $3 - 5$ days, whereas cell types such as neurons, can survive for many years (Sender & Milo, 2021). The decision of a cell to die, known as apoptosis, is an example of a cellular *fate*, where other examples of fates include cell division, differentiation, and migration. How a cell determines its fate is dependent on signalling molecules in the extracellular environment and receptor molecules on the surface of a cell, with which these signalling molecules interact. Often, cells will present a diverse range of receptor types on their cell surface, where each receptor type can initiate a different functional response by interacting with the appropriate signalling molecules (Uings & Farrow, 2000). Receptors can be classified into different families, based on how they respond to signalling molecules, and two such families which will be discussed in this thesis are the cytokine receptors and the receptor tyrosine kinases (RTKs). Both of these receptor types are trans-membrane proteins, with an extracellular domain, a trans-membrane domain and an intracellular domain (Grötzinger, 2002; Schlessinger & Ullrich, 1992). The main difference between the receptor types is that cytokine receptors lack tyrosine kinase activity (Haan *et al.*, 2006), meaning that

they cannot initiate a cell signalling pathway, leading a cell to its fate, without the help of other intracellular proteins which *do* posses tyrosine kinase activity. As implied by their name, RTKs have a tyrosine kinase domain as part of their intracellular region, which allows them to phosphorylate (transfer a phosphate group to) other intracellular proteins. These phosphorylation events induce cell signalling pathways which lead to a fate for the cell. Cytokine receptors however, require association of other intracellular proteins, known as Janus kinases (JAKs) in order to initiate signalling. It is the JAK molecules which have the kinase activity, and a cytokine receptor which is not bound to a JAK molecule does not possess such activity (Haan *et al.*, 2006). Both receptor types bind to molecules, collectively known as ligands, in the extracellular medium, initiating the intracellular signalling pathways, however another difference between the receptor types is the ligands with which they bind. Cytokine receptors bind predominantly with cytokine molecules, whereas RTKs bind predominantly with other molecules, such as growth factors.

Cytokines are a class of ligand molecules, which are produced by, and utilised by, cells of the immune system (Altan-Bonnet & Mukherjee, 2019). Cytokines can be further sub-classified based on their functions, where some act as growth factors (initiating cell growth and division) and others act as pro- or anti-inflammatory molecules (Dinarello, 2007). There are hundreds of different cytokines (Cameron & Kelvin, 2013), and often different cytokines will have the ability to bind to the same receptor, but the specific receptor-ligand pairing appears to be crucial in determining the outcome of the initiated cell signalling (Grötzinger, 2002). Of particular interest in this thesis are a class of cytokines known as the interleukins, of which there are currently 40 known varieties, namely interleukin 1 (IL-1) to interleukin 40 (IL-40). Among the main functions of these cytokines are modulating growth and differentiation during inflammatory responses (Vaillant & Qurie, 2019). Two particular interleukins, IL-6 and IL-27, can act as both pro-inflammatory and anti-inflammatory cytokines, depending on the environment, however IL-6 is predominantly pro-inflammatory and IL-27 is predominantly anti-inflammatory (Hunter & Jones, 2015; Rose-John, 2018; Yoshida & Hunter, 2015a). Interestingly, these two interleukins share a common receptor subunit, namely glycoprotein 130 (GP130). In the case of IL-6 stimulation, a

hexameric receptor complex is formed, comprised of two copies of each of, IL-6 receptor α (IL-6Rα, an IL-6 specific receptor), GP130 and IL-6 (Boulanger *et al.*, 2003). Under IL-27 stimulation, IL-27 binds to its receptor IL-27 receptor α (IL-27Rα), which then forms a dimer with a single unbound GP130 unit (Yoshida & Hunter, 2015a). Both of these signalling dimers are then capable of initiating a signalling pathway known as the JAK/signal transducer and activator of transcription (STAT) pathway.

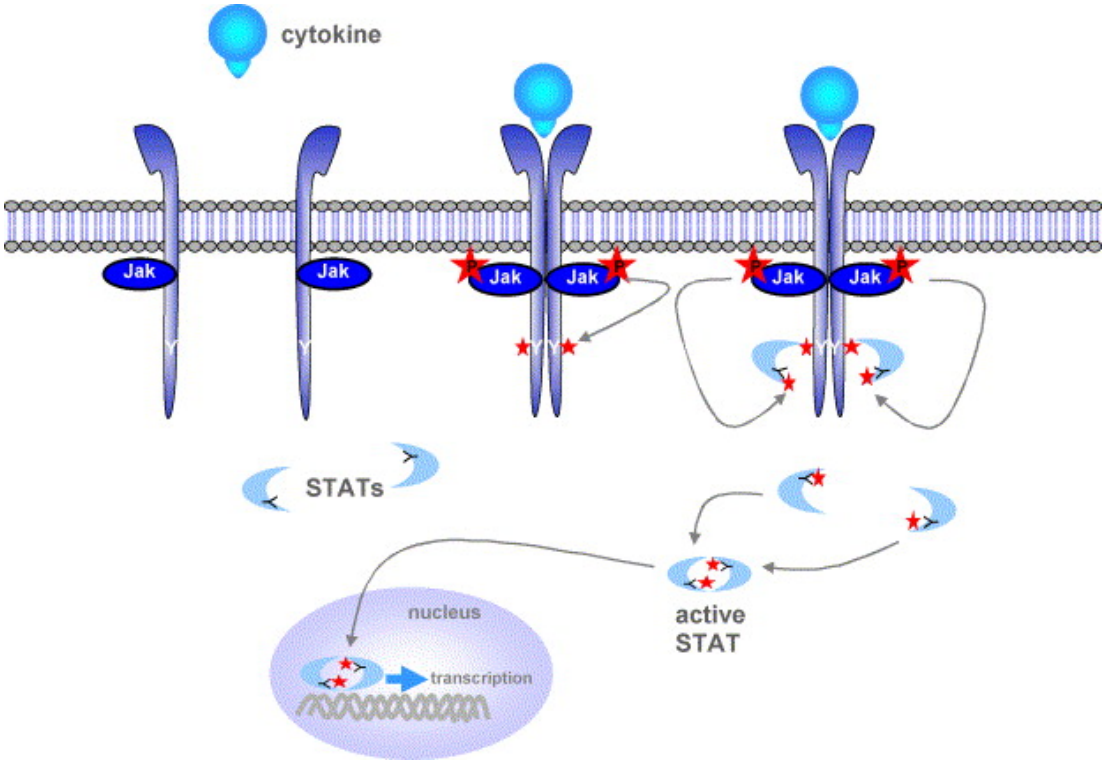

Figure 1.1: Diagram of the JAK/STAT signalling pathway, initiated by cytokine molecules, taken from Haan *et al.* (2006). The red stars attached to a species indicate that the species is phosphorylated.

The JAK/STAT signalling pathway is depicted in Figure 1.1 for a generic receptor dimer induced by a cytokine molecule. This dimer could be, for example, the IL-6 or IL-27 induced receptor dimer. As described by Haan *et al.* (2006), when two receptors are brought together by a cytokine, the respective JAK molecules which they are bound to, trans-autophosphorylate themselves. The JAKs then phosphorylate tyrosine residues on the intracellular tails of the

receptors they are bound to. These phosphorylated tyrosine residues on the receptor tails then act as docking sites for other intracellular proteins such as STATs, allowing the STAT molecules to bind the receptors, become phosphorylated themselves, and then dissociate. Two phosphorylated STAT molecules can then form a dimer in the cell cytoplasm, and migrate to the nucleus of the cell. Within the nucleus, the STAT molecules act as transcription factors, where they bind to target genes, initiating the process of protein synthesis. It is this gene transcription which regulates cellular fates such as differentiation and division.

Unlike the cytokine receptors, RTKs have their own tyrosine kinase activity and thus do not depend on JAK molecules in order to initiate cell signalling. One important subclass of RTKs is the ErbB class, which consists of four members, namely the epidermal growth factor receptor (EGFR, also known as ErbB1 and Her1) and ErbB2-ErbB4. There are 12 known growth factors (ligands) which can each bind with at least one member of the ErbB family, inducing the formation of receptor dimers (Tebbutt *et al.*, 2013). These can be homodimers, if two receptors of the same type are brought together, or heterodimers, if the two receptors are of different types. Of the ErbB family, EGFR is the most studied of the receptors, which binds with its ligand EGF, whose primary function is to induce cell division, as well as differentiation, migration and growth. EGF induces a homodimer, comprised of two molecules of EGFR and two of EGF, where, upon dimerisation, the receptor molecules trans-autophosphorylate (Wee & Wang, 2017). Similarly to the cytokine receptor dimers, a dimer of EGFR is also capable of initiating a cell signalling pathway, due to the phosphorylated tyrosine residues on the intracellular tails of EGFR acting as docking sites for other proteins. In this case however, the primary pathway induced by EGF is not the JAK/STAT pathway, but a pathway known as the mitogen activated protein kinase (MAPK) pathway (Wee & Wang, 2017).

The MAPK signalling pathway is depicted in Figure 1.2 for a generic RTK forming a dimer induced by a ligand molecule. As described by Liu *et al.* (2018) and seen in Figure 1.2, there are many proteins involved in the MAPK pathway. Firstly, growth factor receptor bound protein 2 (Grb2), binds to the docking sites on the RTK dimer and is subsequently phosphorylated. Through protein-protein interactions, each of the molecules, SOS, RAS, RAF, MEK and ERK then become

phosphorylated. ERK can phosphorylate various cytoplasmic proteins, such as RSK, or can migrate to the nucleus, where it activates transcription factors such as ELK (Fang & Richardson, 2005). As in the case of the JAK/STAT pathway, the MAPK pathway promotes gene transcription and hence regulates the cell fate.

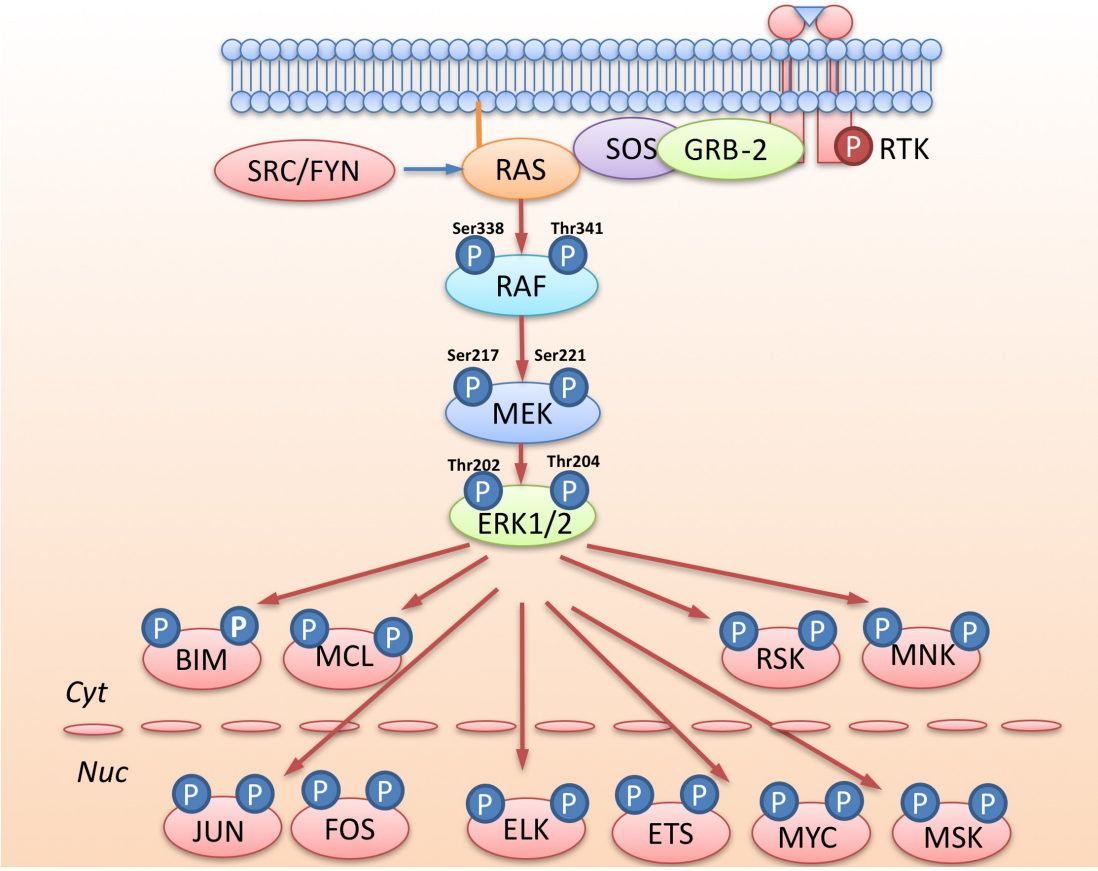

Figure 1.2: Diagram of the MAPK signalling pathway, initiated by a growth factor, taken from Liu *et al.* (2018). A letter "P" in a circle attached to a species indicates that this species is phosphorylated.

Another important subclass of RTKs which will be discussed in this thesis, is the fibroblast growth factor receptors (FGFRs), of which there are four members, namely FGFR1 - FGFR4, which possess tyrosine kinase activity and signal through interaction with 18 known fibroblast growth factor (FGF) ligands (Turner & Grose, 2010). The mechanism of FGFR activation and signal induction is very

similar to that for the ErbB class, whereby ligands induce receptor homo- or heterodimerisation causing the receptor tails to come into close proximity with one another, leading to trans-autophosphorylation (Sarabipour, 2017). Dependent on the receptor-ligand pairing, FGFR dimers are capable of inducing both STAT signalling pathways (similarly to the cytokine receptors) and the MAPK pathway (similarly to the ErbB family), among others (Ornitz & Itoh, 2015). Both of the cell signalling pathways discussed above are important processes in healthy cells, allowing for cells to grow, divide and even die, when necessary. However, dysregulation of parts of each pathway, through mutation or up-regulation of specific proteins, can have adverse effects on the cells and can cause disease in humans.

Systemic lupus erythematosus (SLE) is an autoimmune disease causing inflammation to tissues, and organ damage. Up-regulation of certain inflammatory cytokines is known to contribute to this disease, whereby the JAK/STAT pathway becomes overstimulated, leading to excessive inflammation and cell death (Ohl & Tenbrock, 2011). Another condition which is linked to excessive concentrations of pro-inflammatory cytokines is Crohn's disease, which causes inflammation of the bowels (Leon *et al.*, 2009). Given the significance of cytokines in both of these autoimmune conditions, it is important to better understand the JAK/STAT pathway and specifically, how this pathway is dysregulated in patients with these conditions.

Dysregulation of the MAPK pathway also has major implications in human disease. Over recent years, EGFR has become an important therapeutic target in many different types of cancer, particularly lung cancer, where increased levels of EGFR correlate with poor patient prognosis (Normanno *et al.*, 2006; Sharma *et al.*, 2007). As well as up-regulation of EGFR, there are now several well known mutant varieties of the protein, some of which do not require binding of the ligand extracellularly, to become activated and initiate the MAPK pathway (Bethune *et al.*, 2010). In both cases, the signalling pathway becomes highly activated, resulting in abnormal cell division or anti-apoptosis, which yields tumour growth. Although there are several marketed drugs to treat cancers caused by mutations in EGFR, by inactivating the protein, a common occurrence in patients being treated with such drugs, is the development of secondary mutations, meaning that the drug no longer inhibits cell signalling by EGFR (Huang & Fu, 2015).

There is therefore, still a huge importance in studying the EGFR induced MAPK signalling pathway and particularly the response of proteins in the pathway to novel drugs.

As explained by Turner & Grose (2010), the pathogenesis of multiple types of cancer can also be correlated with alterations and up-regulations of members of the FGFR family, where for example, approximately 50% of bladder cancers have mutations in FGFR3. The same authors describe how FGFR2 is mutated in approximately 12% of endometrial carcinomas, and how FGFR1 and FGFR2 have been found to be up-regulated in breast and gastric cancers, respectively. Similarly to the ErbB family, the FGFR family represents a class of molecules which are current targets for cancer therapies, where small molecule inhibitors are used to target the tyrosine kinase domains of the receptors, yielding them inactive. There are many such inhibitors already marketed and there is also scope for further development in this field (Casadei *et al.*, 2019; Ghedini *et al.*, 2018; Porta *et al.*, 2017), hence an ongoing need to understand fully the signalling initiated by FGFR family members.

1.2 Objectives of this thesis

It is clear that receptor-ligand mediated signalling pathways can be large, complex systems involving many proteins, and can be highly dependent on the cellular environment. Deterministic and stochastic mathematical modelling are approaches which have been used for many years in order to learn about specific parts of such systems, and introductions to each type of modelling are given by Allen (2010) and Allen (2007). Deterministic models are typically used to study biological processes in which there are large concentrations of the species (variables) in the model, and random fluctuations are ignored. On the other hand, stochastic modelling is useful if there is thought to be only a small quantity of at least one species in the model. This is because stochastic models account for the random fluctuations in copy numbers, seen in nature, and hence give more realistic outputs then deterministic models. The disadvantage of stochastic modelling is that it is usually more analytically complicated and computationally expensive than deterministic modelling (Simoni *et al.*, 2020).

Ligand concentrations, receptor densities and downstream signalling protein concentrations can vary greatly depending on the cell type and location, with some of these protein copy numbers becoming low enough that stochastic fluctuations should be taken into account (Chen *et al.*, 2009; Feinerman *et al.*, 2008; Fritsche-Guenther *et al.*, 2011). In Chapter 3 of this thesis, a general receptor-ligand binding model is considered and analysed stochastically, which could be relevant to any of the receptor types discussed in Section 1.1. In particular, and as discussed for both the cytokine receptors and the RTKs, often a single receptor molecule is capable of binding with multiple ligand varieties, or vice versa. This means that, on cells which express two or more receptor types which bind a common ligand, there is an element of natural competition between the receptors for the ligand, where two different receptor-ligand complexes can form. The activation, via ligand binding, of the different receptor types may ultimately, through signalling pathways, lead to different cellular fates, and hence it is interesting to study how different biological parameters may affect the fate of the cell. The competition process is modelled as a continuous-time Markov chain, as introduced in Section 2.2.1. Two stochastic descriptors are analysed for such a competition process, namely the steady state distribution of the two receptor-ligand complex types on the cell surface, and the time scales of formation of each receptor-ligand complex type. Properties of stochastic processes can be complex and computationally expensive to analyse exactly, and hence approximate methods are proposed in Chapter 3.

In some cases, where a biological system is governed by many reactions and involves many molecular species, the corresponding stochastic model can become too complex to analyse, and hence deterministic approximations can be employed. Many questions relating to signalling pathways are difficult to determine experimentally, but can be deduced by analysing properties of a deterministic mathematical model for the system. For example, one can use a mathematical model to infer summary statistics relating to a process, determine rate constants of reactions and even to decide between biological hypotheses about a process. In Chapter 4 of this thesis, a deterministic mathematical modelling approach is used, in combination with experimental data, to answer questions about the signalling induced by the cytokines IL-6 and IL-27, in different cell types. The experimental

datasets used in Chapter 4 have been provided by Dr. Ignacio Moraga and Dr. Stephan Wilmes from the School of Life Sciences at the University of Dundee. After formulation of the mathematical models of IL-6 and IL-27 induced signalling, Bayesian methods, namely those introduced in Section 2.5, are used to parametrise the models and to choose between two proposed hypotheses relating to internalisation of receptor molecules. The models are validated using additional experimental datasets and finally, predictions are made using the models, relating to signalling under different cellular conditions, in particular those found in patients with SLE and Crohn's disease. In Chapter 4, the concentrations of receptors, ligands, and other proteins in the signalling pathway, represented by variables in the mathematical models, are large enough that stochastic fluctuations need not be taken into account and hence the deterministic approach is reasonable.

In Chapters 5 and 6, the biological focus is turned to the RTKs, where in these chapters, FGFR2 and EGFR are the receptors of interest, respectively. It is well known that receptors, and other downstream signalling proteins, are capable of binding with other proteins to propagate a cellular signal, however a lesser reported event is the formation of ternary complexes, *i.e.* complexes comprised of three (potentially different) molecules, all bound together. This ternary complex formation was of interest to the group of molecular and cellular biology at the University of Leeds, and in particular, Dr. Chi-Chuan Lin has collected experimental data to evidence the formation of ternary complexes of FGFR2 with two other intracellular proteins known as Shp2 and Plcγ1. He has been able to show, through imagining experiments, that these ternary complexes come together to form high density droplets, and it is assumed that these droplets with high concentrations of signalling proteins are responsible for an increase in the signal produced by FGFR2. In Chapter 5 of this thesis, this ternary complex formation is modelled deterministically and through analysis of the steady states of the mathematical model, the experimental observation of the ternary complex is verified. The model is also simulated to determine the effects of varying protein concentrations on the formation of ternary complexes and the feasible parameter space is explored with relation to the model outputs and the stability of the steady state.

In Chapter 6, the EGFR initiated MAPK signalling pathway is studied. In particular, using data provided by AstraZeneca, the effect of eight different potential drug inhibitors of EGFR is analysed in terms of their effect on phosphorylated and total EGFR and three downstream proteins in the MAPK pathway, MEK, ERK and RSK. The inhibitors are third generation tyrosine kinase inhibitors (TKIs), which are able to pass through the cell membrane where they then form covalent bonds with residues on the tyrosine kinase domain of EGFR. The inhibitors compete with adenosine triphosphate (ATP), a molecule responsible for phosphorylating receptors, for binding to the receptor tails, thus inactivating the receptors by blocking further phosphorylation of downstream proteins in the signalling pathways. Statistical methods, introduced in Section 2.6, are used to determine whether there are any differences in protein expression, depending on the inhibitor type, the cell line (natural or mutant EGFR) and the concentration of the inhibitor. Finally, a review is given of the current mathematical modelling in the literature relating to the inhibition of EGFR by TKIs.

The broad aims of this thesis are to use methods from mathematical modelling, Bayesian statistics and frequentist statistics in combination with experimental data, to give useful insights into biological systems, which would be difficult to determine experimentally. In Chapter 3, new mathematical methods for analysing stochastic competition processes are developed, which can be useful in situations with low protein copy numbers. In Chapters 4, 5 and 6, mathematical modelling and statistical techniques are used to analyse properties of biological systems for which experimental data has been collected. Finally, Chapter 7 is a conclusion and discussion.

Chapter 2

Mathematical background

This chapter provides a background in probability theory and stochastic processes, as well as an overview of some of the mathematical methods and statistical techniques which will be used in this thesis.

2.1 Probability theory

In Chapter 3 of this thesis, analysis of a type of stochastic process called a Markov chain is carried out, and hence in this section the probability theory required to understand such a process is introduced. The definitions and explanations given here are based on the works by Allen (2010), Pinsky & Karlin (2010) and Casella & Berger (2002).

2.1.1 Random variables

The *sample space* of a random experiment is a set, Ω, of all possible outcomes of the experiment. For example, when a single coin is tossed, the sample space is $\Omega = \{H, T\}$ where H and T stand for heads and tails, respectively. A *random variable* X, is a real valued function that maps from the sample space to some subset of the real numbers, $\mathbb{R}$. This subset is known as the *support* (or range), A_X, of the random variable and is defined as

$$A_X = \{x \in \mathbb{R} : X(\omega) = x \text{ for some } \omega \in \Omega\}.$$

A *discrete random variable* is a variable whose support is finite or countably infinite, whereas a *continuous random variable* has an uncountably infinite support.

Definition 1. The *cumulative distribution function*, or cdf, of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = \mathbb{P}(X \leq x), \quad -\infty < x < +\infty.$$

Definition 2. If X is a discrete random variable, then the function

$$f_X(x) = \mathbb{P}(X = x), \text{ for } x \in A_X,$$

is the probability that X takes a particular value in its support, and is known as the *probability mass function* (pmf) of X.

Definition 3. If X is a continuous random variable with cdf $F_X(x)$ and there exists a nonnegative, integrable function $f : \mathbb{R} \to [0, \infty)$, such that

$$F_X(x) = \int_{-\infty}^{x} f_Y(y)dy,$$

then the function $f_X(x)$ is called the *probability density function*, or pdf, of X.

2.1.2 Exponential distribution

A well known distribution which is important in the study of stochastic processes is the *exponential distribution*.

Definition 4. A nonnegative continuous random variable X is said to follow an exponential distribution with parameter $\lambda > 0$, $X \sim Exp(\lambda)$, if its probability density function is

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

2.1.3 Uniform distribution

Definition 5. A continuous random variable X is said to follow a *continuous uniform distribution* with parameters $-\infty < a < b < +\infty$, $X \sim Unif(a, b)$, if its

probability density function is

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b], \\ 0 & \text{otherwise.} \end{cases}$$

2.1.4 Multivariate distributions

When several random variables $X_1, X_2, \ldots, X_n$ are associated with the same sample space, one can define a *multivariate probability density function* if the variables are continuous, or a *multivariate probability mass function* if the variables are discrete. For a random vector of two random variables, (X_1, X_2), the support of this vector is

$$A_{X_1,X_2} = \{(X_1(\omega), X_2(\omega)) | \omega \in \Omega\} \subseteq \mathbb{R}^2,$$

where Ω is the common sample space of X_1 and X_2.

Definition 6. A function $f_{X_1,X_2}(x_1, x_2)$ from A_{X_1,X_2} to $\mathbb{R}$ is called a joint probability mass function, or joint pmf, of the discrete random vector (X_1, X_2), if

$$\sum_{(x_1,x_2)\in A_{X_1,X_2}} f_{X_1,X_2}(x_1, x_2) = 1,$$

and for every $B \subset A_{X_1,X_2}$,

$$\mathbb{P}((X_1, X_2) \in B) = \sum_{(x_1,x_2)\in B} f_{X_1,X_2}(x_1, x_2).$$

The *marginal probability mass function* of X_1 is defined as

$$f_{X_1}(x_1) = \sum_{x_2} f_{X_1,X_2}(x_1, x_2),$$

and the marginal pmf of X_2 can be defined in a similar manner.

Definition 7. A function $f_{X_1,X_2}(x_1, x_2)$ from A_{X_1,X_2} to $\mathbb{R}$ is called a joint probability density function, or joint pdf, of the continuous random vector (X_1, X_2), if

$$\iint_{A_{X_1,X_2}} f_{X_1,X_2}(x_1, x_2) dx_1 dx_2 = 1,$$

and for every $B \subset A_{X_1,X_2}$,

$$\mathbb{P}((X_1, X_2) \in B) = \iint_B f_{X_1,X_2}(x_1, x_2)dx_1dx_2.$$

The *marginal probability density function* of X_1 is defined as

$$f_{X_1}(x_1) = \int_{\mathbb{R}} f_{X_1,X_2}(x_1, x_2)dx_2,$$

and the marginal pdf of X_2 can be defined in a similar manner.

Definition 8. Two random variables, discrete or continuous, are said to be *independent* if and only if

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2),$$

for all $(x_1, x_2) \in A_{X_1,X_2}$, otherwise they are said to be *dependent*.

2.1.5 Expectation, variance and covariance

In this section, the expectation, variance and covariance of random variables are defined.

Definition 9. The *expectation* of a discrete random variable X with support A_X, denoted by $\mathbb{E}[X]$, is defined as

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

where $f_X(x)$ is the pmf of X.
The expectation of a continuous random variable X with support A_X, is defined as

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

where $f_X(x)$ is the pdf of X.

Definition 10. The *variance* of a random variable X, denoted by $\mathrm{Var}(X)$, is

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

The standard deviation of X is $\sigma = \sqrt{\mathrm{Var}(X)}$.

Definition 11. The *covariance* of two jointly distributed random variables, X_1 and X_2, denoted by $\mathrm{Cov}(X_1, X_2)$, is defined as

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2].$$

If $\mathrm{Cov}(X_1, X_2) = 0$ then X_1 and X_2 are said to be *uncorrelated*.

Definition 12. The *correlation* of two jointly distributed random variables, X_1 and X_2, is defined as

$$\rho_{X_1 X_2} = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}},$$

where $\rho_{X_1 X_2}$ is known as the Pearson correlation coefficient. This coefficient can also be defined for paired sample data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ (n pairs) and is usually denoted r_{xy} where

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

with $\bar{x}$ and $\bar{y}$ representing the sample means, *i.e.*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

2.1.6 Laplace-Stieltjes transform

Definition 13. The *Laplace-Stieltjes transform* of a continuous random variable X with pdf $f_X(x)$, denoted by $\phi_X(z)$, is defined as

$$\phi_X(z) = \mathbb{E}[e^{-zX}] = \int_0^\infty e^{-zx} f_X(x)dx, \quad \mathrm{Re}(z) \geq 0,$$

where $z \in \mathbb{C}$. By taking derivatives of $\phi_X(z)$, the lth order moments of X, denoted by $\mathbb{E}[X^l]$, can be found, where

$$\mathbb{E}[X^l] = (-1)^l \frac{d^l}{dz^l}\phi_X(z)\bigg|_{z=0}, \quad l \geq 1.$$

2.2 Stochastic processes

In this section, stochastic processes are introduced, using definitions and explanations from Allen (2010), Pinsky & Karlin (2010), Latouche *et al.* (1999), Kulkarni (2016) and He (2014).

Definition 14. A *stochastic process* is a collection of random variables

$$\mathcal{X} = \{X_t(\omega) : t \in T, \omega \in \Omega\}$$

where T is some index set and Ω is the common sample space of the random variables. For each fixed t, $X_t(\omega)$ denotes a single random variable defined on Ω. For each fixed $\omega \in \Omega$, $X_t(\omega)$ corresponds to a function defined on T that is called a stochastic realisation of the process.

The values of the random variables $X_t(\omega)$ are known as the *states* of the process and the set of all possible values is known as the *state space* of the process and is denoted by $\mathcal{S}_{\mathcal{X}}$. For the stochastic processes considered in this thesis, the index set T, will be a set of times, and hence the stochastic processes will track how certain random variables evolve through time. If the index takes discrete values, for example $T = \{0, 1, 2, \dots\}$, then the process is discrete-time, whereas if the index takes continuous values, for example $T = [0, \infty)$, then the process is continuous-time. A stochastic process can also incorporate multiple random variables in a random vector. For example, a stochastic process with two random variables, $X_t^1(\omega)$ and $X_t^2(\omega)$, may be denoted $\{(X_t^1(\omega), X_t^2(\omega)) : t \in T, \omega \in \Omega\}$. Often, as will be the case for the remainder of this thesis, the variable ω is omitted and the random variables are denoted $X(t)$.

2.2.1 Continuous-time Markov chain

A certain type of stochastic process is one which obeys the *Markov property* and is referred to as "memoryless". In short, a Markovian process is a stochastic process in which the next state which the process will move to, depends only on the current state of the process and not on any previous states. In this thesis, *continuous-time Markov chains* are considered, whereby the index set relates to time and is continuous, $t \in [0, \infty)$. When the state space is discrete, as will be

the case in this thesis, the stochastic process is referred to as a Markov *chain*, whereas when the state space is continuous, it is known as a Markov *process*.

Definition 15. The stochastic process $\{X(t) : t \in [0, \infty)\}$, defined on the state space $\mathcal{S}_X$, is called a continuous-time Markov chain (CTMC) if it satisfies the following condition:
For any sequence of real numbers satisfying $0 \leq t_0 < t_1 < \cdots < t_n < t_{n+1}$,

$$\begin{aligned}\mathbb{P}(X(t_{n+1}) = x_{n+1} | X(t_0) = x_0, X(t_1) = x_1, \ldots, X(t_n) = x_n) \\ = \mathbb{P}(X(t_{n+1}) = x_{n+1} | X(t_n) = x_n),\end{aligned}$$

for any $x_0, x_1, \ldots, x_{n+1} \in \mathcal{S}_X$. This condition is known as the *Markov property*.

2.2.2 Transition probabilities

For a CTMC $\mathcal{X} = \{X(t) : t \in [0, \infty)\}$, there is a probability associated with the random variable $X(t)$ being in each state $i \in \mathcal{S}_X$, where $\mathcal{S}_X$ is the state space for the chain. These probabilities are

$$p_i(t) = \mathbb{P}(X(t) = i), \quad i \in \mathcal{S}_X.$$

Definition 16. A relation between random variables $X(t)$ and $X(s)$, $s < t$, is defined by the corresponding *transition probability*,

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

for $i, j \in \mathcal{S}_X$.

If the transition probabilities do not depend explicitly on the times s and t and instead depend only on the time interval $t - s$, then they are said to be *homogeneous* probabilities, and the resulting CTMC is known as a *time-homogeneous* CTMC. This will be the case throughout this thesis, and one can write that

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

for $s < t$. The transition probabilities can be arranged into a square matrix,

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

known as the *transition matrix*. The entries in each row of the transition matrix should sum to 1, since from state i the process must travel to another state $j \in \mathcal{S}_X$ or stay in the same state, in the time interval $[0, t]$.

2.2.3 Infinitesimal generator matrix

The transition probabilities $p_{ij}(\cdot)$ can be used to derive *transition rates*, q_{ij}.

Definition 17. The transition rates are defined as

$$q_{ij} = \begin{cases} \lim\limits_{\Delta t \to 0^+} \frac{p_{ij}(\Delta t) - p_{ij}(0)}{\Delta t} = \lim\limits_{\Delta t \to 0^+} \frac{p_{ij}(\Delta t)}{\Delta t}, & \text{for } i \neq j, \\ \lim\limits_{\Delta t \to 0^+} \frac{p_{ii}(\Delta t) - p_{ii}(0)}{\Delta t} = \lim\limits_{\Delta t \to 0^+} \frac{p_{ii}(\Delta t) - 1}{\Delta t}, & \text{for } i = j. \end{cases}$$

Using the fact that the transition probabilities sum to 1, it can be shown that

$$q_{ii} = -\sum_{j \in \mathcal{S}_X, j \neq i} q_{ij}, \quad \forall i \in \mathcal{S}_X.$$

The *infinitesimal generator matrix*, $\boldsymbol{Q} = (q_{ij})_{i,j \in \mathcal{S}_X}$, is a square matrix containing the transition rates where each row sum is 0 and the ith diagonal element is the negative of the sum of the off-diagonal elements in that row.

2.2.4 Interevent times

In order to simulate a CTMC, it is necessary to know the distribution of the time between successive events in the process, known as the *interevent time*. For a CTMC $\mathcal{X} = \{X(t) : t \in [0, \infty)\}$, the random variable for the interevent time is $T_i = W_{i+1} - W_i$, where W_i is the time at which the process makes the ith jump. It can be shown (see Allen (2010); Pinsky & Karlin (2010)) that T_i is an exponential random variable with parameter $\sum_{j \in \mathcal{S}_X, j \neq i} q_{ij}$, where q_{ij} are the transition rates as defined in Section 2.2.3. The expectation of a random variable $X \sim Exp(\lambda)$

is given by $\mathbb{E}[X] = \frac{1}{\lambda}$ and hence the expected time that a CTMC spends in state i is

$$\mathbb{E}[T_i] = \frac{1}{\sum_{k \in \mathcal{S}_\mathrm{X}, k \neq i} q_{ik}}.$$

Moreover, the probability that the process will move from state i to state j in one step is given by

$$p_{ij} = \frac{q_{ij}}{\sum_{k \in \mathcal{S}_\mathrm{X}, k \neq i} q_{ik}}.$$

The interevent times of a Markov chain being exponentially distributed can be intuited by considering that the exponential distribution is the only continuous probability distribution for which the *memoryless property* holds. This property says that

$$\mathbb{P}(T_i \geq t + \Delta t | T_i \geq t) = \mathbb{P}(T_i \geq \Delta t).$$

2.2.5 Kolmogorov differential equations

Definition 18. The *forward Kolmogorov differential equations* are a system of equations describing the rate of change of the transition probabilities. Formally,

$$\frac{dp_{ij}(t)}{dt} = \sum_{k \in \mathcal{S}_\mathrm{X}} q_{kj} p_{ik}(t), \quad \forall i, j \in \mathcal{S}_\mathrm{X},$$

and the equations can be written in matrix form as

$$\frac{d\boldsymbol{P}(t)}{dt} = \boldsymbol{Q}\boldsymbol{P}(t),$$

where $\boldsymbol{P}(\cdot)$ is the transition matrix and $\boldsymbol{Q}$ is the infinitesimal generator for the CTMC. This system of equations is also known as the *chemical master equation* (CME) or just the *master equation*.

2.2.6 Linear noise approximation

The solution of the chemical master equation,

$$\boldsymbol{P}(t) = \boldsymbol{P}(0) e^{\boldsymbol{Q}t},$$

gives the probability of being in each state of the state space at every time t. This solution, however, is usually computationally intractable to obtain, since the state space is typically large or even infinite (particularly in the case of multi-dimensional Markov chains), and hence one would need to compute large matrix exponentials. There are various methods of approximating the solution to the CME such as generating function techniques (Ammar *et al.*, 2016), and here the linear noise approximation is introduced as one such method. The *linear noise approximation* (LNA), also known as the system size expansion, was developed by Van Kampen (1976).

The LNA is a method often used when considering stochastic processes in chemistry, such as chemical reaction networks, whereby chemical reactions are assumed to be occurring within a fixed volume, Ψ (Elf & Ehrenberg, 2003a; Hayot & Jayaprakash, 2004). The method provides a second order approximation to the CME via a large volume expansion around the steady state, with the aim to obtain differential equations for different order moments of the random variables in the process. As explained by Elf & Ehrenberg (2003a), the method involves Taylor expanding the CME around the steady state of the system in powers of $\sqrt{\Psi}^{-1}$ where Ψ is the volume of the system in which the chemical reactions are occurring. Terms of first-order in $\sqrt{\Psi}^{-1}$ yield the deterministic equations describing the concentrations of the species in the system (which are participating in the chemical reactions) and terms of second order in $\sqrt{\Psi}^{-1}$ give a linear Fokker-Planck equation for the fluctuations of the numbers of molecules.

2.2.7 Birth-and-death process

In this section, a continuous-time *birth-and-death process* is introduced, a type of CTMC. The birth-and-death Markov chain $\mathcal{X} = \{X(t) : t \geq 0\}$ may have either a finite or infinite state space, $\mathcal{S}_\mathcal{X} = \{0, 1, 2, \ldots, N\}$ or $\mathcal{S}_\mathcal{X} = \{0, 1, 2, \ldots\}$. There are only two types of events in such a process, births with rate λ_n, moving the process from state n to state $n + 1$, and deaths with rate μ_n, moving the process from state n to state $n - 1$, as depicted in Figure 2.1.

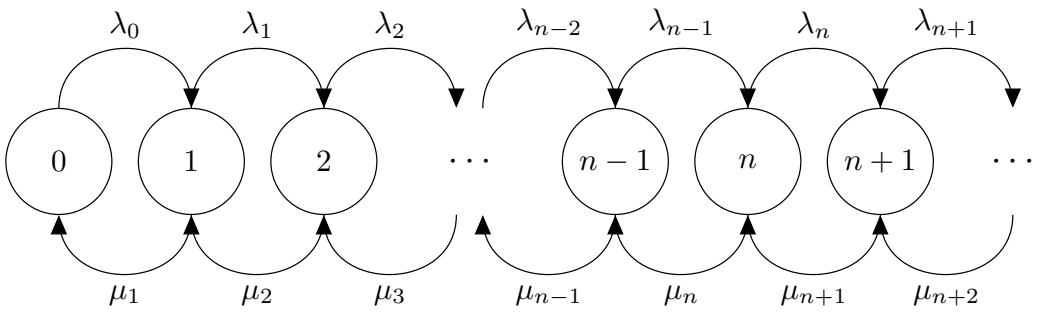

Figure 2.1: A depiction of a birth-and-death process.

Denoting by $\Delta X(t)$ the change in the state of the process from t to $t + \Delta t$, *i.e.*

$$\Delta X(t) = X(t + \Delta t) - X(t),$$

the transition probabilities for this process are

$$\begin{aligned} p_{i,i+j}(\Delta t) &= \mathbb{P}(\Delta X(t) = j | X(t) = i) \\ &= \begin{cases} \lambda_i \Delta t + o(\Delta t), & j = 1, \\ \mu_i \Delta t + o(\Delta t), & j = -1, \\ 1 - (\lambda_i + \mu_i)\Delta t + o(\Delta t), & j = 0, \\ o(\Delta t), & \text{otherwise.} \end{cases} \end{aligned}$$

The infinitesimal generator matrix for the process $\mathcal{X}$ is given by

$$\boldsymbol{Q} = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \dots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & \dots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & \dots \\ 0 & 0 & \mu_3 & -(\lambda_3 + \mu_3) & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

and it can be seen that $\boldsymbol{Q}$ is tridiagonal.

2.2.8 Quasi-birth-and-death process

A generalisation in higher dimensions of the birth-and-death process is the *quasi-birth-and-death process*. This process can be defined in n dimensions, but here a two-dimensional quasi-birth-and-death process (*i.e.* a bivariate CTMC) is discussed.

Definition 19. A quasi-birth-and-death (QBD) process is a bivariate CTMC $\mathcal{X} = \{(X_1(t), X_2(t)) : t \geq 0\}$ with state space $\mathcal{S}_\mathcal{X} = \{(i, j) : i = 0, 1, \ldots, J(j), j \geq 0\}$, where in one step, the process can move from state (i, j) to state (i', j') only if $j' = j$, $j + 1$, or $j - 1$. The coordinate j is called the level, while the coordinate i is called the phase of the state (i, j). The number of states in level j, $J(j) + 1$, can be finite or infinite.

In a QBD process the one-step transitions from a state can move freely in the phase, but can move only within the same level or to one of the two adjacent levels. The state space can be partitioned into levels,

$$\mathcal{S}_\mathcal{X} = \bigcup_{k \geq 0} L(k), \quad L(k) \;=\; \{(i, j) \in \mathcal{S}_\mathcal{X} : j = k\}, \quad k \geq 0.$$

The states inside a level can also be ordered as

$$L(k) = \{(0, k), (1, k), \ldots, (J(k), k)\}$$

and thus, with the states arranged into levels in this way, the infinitesimal generator matrix is block-tridiagonal and has the form

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{0,0} & \boldsymbol{Q}_{0,1} & \mathbf{0} & \mathbf{0} & \ldots \\ \boldsymbol{Q}_{1,0} & \boldsymbol{Q}_{1,1} & \boldsymbol{Q}_{1,2} & \mathbf{0} & \ldots \\ \mathbf{0} & \boldsymbol{Q}_{2,1} & \boldsymbol{Q}_{2,2} & \boldsymbol{Q}_{2,3} & \ldots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Q}_{3,2} & \boldsymbol{Q}_{3,3} & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Sub-matrices $\boldsymbol{Q}_{k,k'}$ have dimensions $(J(k) + 1) \times (J(k') + 1)$ and contain the transition rates from states in level $L(k)$ to states in level $L(k')$ where $k' \in$

$\{k, k-1, k+1\}$.

A specific example of a QBD process is seen in Figure 2.2 where there are infinitely many levels, represented by rows of the diagram (a example of a level is circled in blue), and the number of states within each level k is $J(k)+1=5$, with the phase of a state represented by the column. In this example, from a specific state (coloured in red), the process can move to any of four adjacent states (coloured in green), two of which are in the same level and two of which are in the adjacent levels.

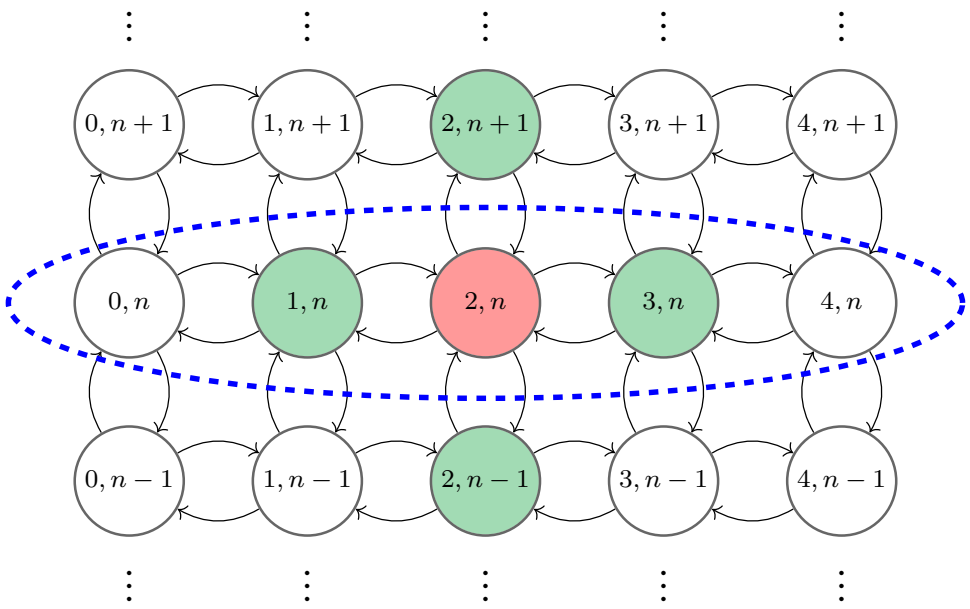

Figure 2.2: A depiction of a bivariate quasi-birth-and-death process. A level in the process is represented by a row of the diagram and a specific example of a level, level n, is circled in blue. From the state coloured in red, the process can move in one step to any of the four adjacent states, coloured in green.

2.2.9 Gillespie algorithm

The Gillespie algorithm developed by Gillespie (1976a), is a stochastic simulation algorithm which allows one to generate a single numerical realisation of a CTMC. The Gillespie algorithm is outlined in Algorithm 1 for a CTMC $\mathcal{X}$ defined on the

space of states $\mathcal{S}_X$ and with infinitesimal generator matrix $\boldsymbol{Q} = (q_{ij})_{i,j\in\mathcal{S}_X}$. The state of the process is denoted by x and the time is represented by t.

Algorithm 1 Gillespie algorithm.

1: Set T_{max}, the maximum time point until which to simulate the process $\mathcal{X}$. Set $t = 0$ and the state of the process equal to the initial state, $x = x_0$.
2: **while** $t < T_{max}$ **do**:
3: Compute the sum of all transition rates to states which can be reached from the current state in a single step, $q_x = \sum_{j\neq x} q_{xj}$.
4: If there are M states which can be reached from the current state, assign each state an index i so that a particular state can be denoted j_i, where $1 \leq i \leq M$.
5: Sample $u_1 \sim Unif(0, 1)$. The next state to which the process moves is state j_k, if $\sum_{i=1}^{k-1} \frac{q_{xj_i}}{q_x} < u_1 < \sum_{i=1}^{k} \frac{q_{xj_i}}{q_x}$.
6: Sample $u_2 \sim Unif(0, 1)$. Update the time as $t \to t - \frac{\log(u_2)}{q_x}$, since this implies (through inverse transform sampling), adding to the current time, a random variable sampled from an exponential distribution with parameter q_x.
7: Update the state of the process $x \to j_k$, as determined using u_1.
8: **end while**

2.3 Ordinary differential equations

As well as stochastic processes, in this thesis, ordinary differential equation (ODE) mathematical models will be considered, a method known as *deterministic modelling*. The word "deterministic" refers to the fact that, for given inputs to such a model, the resulting output is always the same. This is dissimilar to a stochastic model, where for given inputs, one can generate a collection of stochastic realisations of the process. As explained by Hahl & Kremling (2016), deterministic models, such as ODE models, are therefore an approximation to stochastic

models, since random fluctuations are not taken into account. ODE models are commonly used in biological sciences to describe many types of systems, for example, infectious disease modelling (Handel *et al.*, 2020; Rock *et al.*, 2014), predator-prey interactions (Zhang *et al.*, 2015a) and chemical reaction networks (Feinberg, 2019; Higham, 2008). ODE models assume spatial homogeneity and are often based upon the law of mass action kinetics, that the rate of a chemical reaction depends on the concentrations of reactants and the stoichiometry of the reaction (Ferner & Aronson, 2016). The information given in the remainder of this section is based on the book by Allen (2007) who describes how ODEs can be used to model systems in mathematical biology in general.

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

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

For the models considered in this thesis, $x(t)$ will represent a molecular species, for example the concentration of a population of receptors, where t will represent time. Only first-order differential equations will be used in this thesis, such as

$$a_1(t)\frac{dx}{dt} + a_0(t)x = g(t), \tag{2.1}$$

so that the ODE describes the rate of change of the molecular species with respect to time. If the coefficients $a_1(t)$ and $a_0(t)$ in Equation (2.1) are constants or functions of t, but not of x or $\frac{dx}{dt}$, then the equation is said to be *linear*, otherwise it is *nonlinear*. If the equation does not depend explicitly on t then it is said to be *autonomous*, otherwise it is *nonautonomous*. Equation (2.1) is *homogeneous* if $g(t) \equiv 0$ and *nonhomogeneous* otherwise.

Often, instead of just a single ODE, one is concerned with a system of ODEs, so that for example, the concentration of several molecular species can be tracked in parallel over time. This is useful since often molecular species can interact with one another to form new species, and such events can be formulated as a system

of ODEs. A first-order system of differential equations can be written as

$$\frac{d\mathbf{X}}{dt} = \mathbf{F}(\mathbf{X}(t), t), \tag{2.2}$$

where $\mathbf{X} = (x_1(t), x_2(t), \ldots, x_n(t))^T$, $\mathbf{F} = (f_1, f_2, \ldots, f_n)^T$ and $f_i \equiv f_i(x_1(t), x_2(t), \ldots, x_n(t), t)$. Similarly to the one variable case, the system of ODEs (2.2) is said to be *autonomous* if the right-hand side does not depend explicitly on t, as will be the case in this thesis. If the system can be written in the form

$$\frac{dx_i}{dt} = \sum_{j=1}^{n} a_{ij}(t)x_j + g_i(t),$$

for $i = 1, \ldots, n$, then it is said to be *linear*. When using ODE models to describe biological systems, the resulting equations are often *nonlinear* due to interactions between species. For example if two species x_1 and x_2 bind to form x_3, then the ODEs for each of the three species will contain a term proportional to $x_1(t)x_2(t)$ under the law of mass action kinetics. Finally, if the system is linear and $g_i(t) \equiv 0$ for $i = 1, \ldots, n$, then the system is *homogeneous*, otherwise it is *nonhomogeneous*.

For given initial conditions, *i.e.* the value of each variable in the system at time 0, ODE systems can be solved to find the functions $x_i(t)$ which tell us how the species x_i changes over time, for $i = 1, \ldots, n$. In some cases, most commonly for small linear systems of ODEs, the solution can be found analytically (Murphy, 2011), however for larger more complex nonlinear systems it is often impossible to find the analytic solution. In this case, as will be the case throughout this thesis, numerical methods (Kang & Cheek, 1972) are instead employed whereby the system is evolved over time using a method of numerical integration such as Euler's method (Griffiths & Higham, 2010).

2.3.1 Steady states

One important type of solution for an ODE system is the constant solution, $\mathbf{X}^*$, known as the steady state solution, which satisfies

$$\mathbf{F}(\mathbf{X}^*) = \mathbf{0}.$$

For a system of ODEs, this solution is found by setting the right hand of the ODEs to zero and solving the resulting expressions simultaneously. For a biological system, the ODE solution reaching steady state implies that the amount of each species in the model is no longer changing with time. Some ODE systems may have multiple solutions and therefore multiple steady states, however the only biologically relevant steady states, when the variables of the model represent a number of molecules (or people, animals etc.) or a concentration, are those where the steady state value for each variable is greater than or equal to 0.

2.3.2 Stability analysis

Steady states of a system of ODEs can be classified as either *locally stable*, *asymptotically stable* or *unstable*. In loose terms, a steady state is locally stable if a solution which starts close to the steady state solution remains close to this solution as $t \to \infty$. The steady state is asymptotically stable if it is locally stable, and solutions which start close to the steady state solution approach this solution as $t \to \infty$. A steady state for which neither of these conditions is met is classified as unstable. Formal definitions of local and asymptotic stability are as follows.

Definition 20. A steady state solution $\mathbf{X}^*$ of a system of ODEs is *locally stable* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every solution $\mathbf{X}(t)$ with initial condition $\mathbf{X}(t_0) = \mathbf{X_0}$,

$$\|\mathbf{X_0} - \mathbf{X}^*\| < \delta \implies \|\mathbf{X}(t) - \mathbf{X}^*\| < \varepsilon,$$

for all $t \geq t_0$, where $\|.\|$ denotes the Euclidean distance in $\mathbb{R}^n$.

Definition 21. A steady state solution $\mathbf{X}^*$ of a system of ODEs is *locally asymptotically stable* if it is locally stable and there exists $\delta > 0$ such that

$$\|\mathbf{X_0} - \mathbf{X}^*\| < \delta \implies \lim_{t\to\infty} \|\mathbf{X}(t) - \mathbf{X}^*\| = 0.$$

Whether or not a steady state is stable can be reasoned through linearisation of the system around the steady state. For example, in the one variable case, suppose that the ODE

$$\frac{dx}{dt} = f(x)$$

has a steady state at x^*. Then x^* can be perturbed slightly by adding to it a small positive number $u(t)$, so that the perturbation from the steady state can be written as $u(t) = x(t) - x^*$. It is then interesting to note how $u(t)$ changes with time, whereby if it grows with time, the solution is moving away from the steady state and hence the steady state is unstable, and if it decreases with time, the solution is moving towards the steady state and hence the steady state is asymptotically stable. Given that x^* is a constant, the time derivative of $u(t)$ is equal to the time derivative of $x(t)$ and hence,

$$\frac{du}{dt} = \frac{dx}{dt} = f(x) = f(u + x^*).$$

One can then Taylor expand around the steady state and truncate the expansion to the linear term since all higher order terms should be negligible. Hence,

$$\begin{aligned} \frac{du}{dt} &= f(x^*) + f'(x^*)(u + x^* - x^*) + O[(u + x^* - x^*)^2] \\ &\approx f(x^*) + f'(x^*)u \\ &= f'(x^*)u, \end{aligned} \tag{2.3}$$

since $f(x^*) = 0$. The solution to Equation (2.3) is

$$u(t) = u_0 e^{f'(x^*)t},$$

and hence if $f'(x^*) > 0$, $u(t)$ grows with time and the steady state is unstable and if $f'(x^*) < 0$ then $u(t)$ decays with time and the steady state is stable. The value $f'(x^*)$ is known as the *eigenvalue* of the linearised system.

For a system in two variables x and y,

$$\begin{aligned} \frac{dx}{dt} &= f(x, y) \\ \frac{dy}{dt} &= g(x, y) \end{aligned}$$

steady states of the system are solutions (x^*, y^*) which satisfy $f(x^*, y^*) = 0$ and $g(x^*, y^*) = 0$. A similar type of linear stability analysis to that in the one variable case can be carried out here, where f and g are Taylor expanded around

the steady state, using the change of variables $u = x - x^*$ and $v = y - y^*$ to define small perturbations from (x^*, y^*). In this case, it can be found that the system linearised about the steady state (x^*, y^*) is

$$\frac{d\mathbf{Z}}{dt} = \mathbf{J}\mathbf{Z},$$

where $\mathbf{Z} = (u, v)^T$ and $\mathbf{J}$ is the Jacobian matrix for the system, evaluated at the steady state, so that

$$\mathbf{J} = \begin{pmatrix} f_x(x^*, y^*) & f_y(x^*, y^*) \\ g_x(x^*, y^*) & g_y(x^*, y^*) \end{pmatrix}.$$

The elements of $\mathbf{J}$ are the partial derivatives of the functions $f(x, y)$ and $g(x, y)$ with respect to the variables x and y, evaluated at the steady state, for example

$$f_x(x^*, y^*) = \left.\frac{\partial f(x, y)}{\partial x}\right|_{(x,y)=(x^*,y^*)}.$$

This result can easily be extended to n variables, and the stability of the steady state then depends on the eigenvalues of the Jacobian matrix as follows. The steady state is asymptotically stable if and only if all eigenvalues of the Jacobian matrix evaluated at the steady state have negative real part. If one or more eigenvalues of the Jacobian matrix evaluated at the steady state have positive real part, then the steady state is unstable. If any eigenvalue has real part equal to zero, then linear stability analysis is inconclusive and nonlinear theory is required, which is not discussed in this thesis.

2.4 Global sensitivity analysis

In the following chapters of this thesis, mathematical models will be used to describe biological systems, with parameters such as rate constants of reactions, and concentrations of molecules often unknown. Inferring such parameters using Bayesian methods allows one to learn more about the underlying biological system. Before carrying out Bayesian inference, it is useful to employ a sensitivity analysis in order to determine which of the model parameters are most influential to the model output. Sobol sensitivity analysis is one such method, whereby a

"global" approach is taken, so that the parameters are varied simultaneously in order to determine both the individual contribution of a parameter to the output of the model, and the relative contribution of groups of parameters. In this section, a brief description of the Sobol sensitivity analysis method is given (Homma & Saltelli, 1996; Sobol, 1993; Zhang *et al.*, 2015b).

Firstly, let the output of interest of a mathematical model be denoted by $Y = f(\boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_s)$ is a vector of s model parameters, which can vary within known ranges. For example, the mathematical model could be a system of differential equations, and the output, some combination of the model variables at a particular instance in time, t. When a sample of parameters is taken from the known parameter ranges, and the model is simulated for this sample, the model output then has an associated variance. The idea of the Sobol method is to decompose the model variance, $\mathrm{Var}(Y)$, into contributions known as Sobol indices, from each individual parameter and combination of parameters, where the higher the value of the Sobol index, the more influential the parameter, or combination of parameters, is to the model output.

The contribution of a single parameter β_i to the variance of the model output is assessed via the computation of its corresponding *first-order Sobol index*. To compute this quantity, one can compute $\mathrm{Var}(\mathbb{E}[Y|\beta_i])$, where the variance is taken over the $s-1$ remaining parameters. This quantity tells us how much the model output varies when the parameter β_i is sampled from its known range. The first-order Sobol index for the parameter β_i, denoted S_i, is then defined by how much this variability contributes to the total variance of the model output, *i.e.*

$$S_i = \frac{\mathrm{Var}(\mathbb{E}[Y|\beta_i])}{\mathrm{Var}(Y)}.$$

To compute higher-order Sobol indices, quantifying the contribution of interactions between parameters to the model variance, it is noted that if the function $f(\cdot)$ is integrable over $[0,1]^s$ then it can be expanded as

$$Y = f(\boldsymbol{\beta}) = f_0 + \sum_{i=1}^{s} f_i(\beta_i) + \sum_{1 \le i < j \le s} f_{ij}(\beta_i, \beta_j) + \cdots + f_{1\ldots s}(\beta_1, \ldots, \beta_s),$$

i.e. Y can be decomposed into terms which depend on only individual model parameters or combinations of model parameters. Then, as proved by Sobol (1993), provided that each of the functions f in the expansion have zero mean, squaring both sides and integrating gives

$$\mathrm{Var}(Y) = \sum_{i=1}^{s} V_i + \sum_{1\leq i<j\leq s} V_{ij} + \cdots + V_{1\ldots s}, \tag{2.4}$$

where $V_i, V_{ij}, \ldots, V_{1\ldots s}$ are the variances of the functions $f_i, f_{ij}, \ldots, f_{1\ldots s}$ respectively. Hence,

$$\begin{aligned}
V_i &= \mathrm{Var}(\mathbb{E}[Y|\beta_i]),\\
V_{ij} &= \mathrm{Var}(\mathbb{E}[Y|\beta_i,\beta_j]) - V_i - V_j,\\
V_{ijk} &= \mathrm{Var}(\mathbb{E}[Y|\beta_i,\beta_j,\beta_k]) - V_{ij} - V_{ik} - V_{jk} - V_i - V_j - V_k,\\
&\vdots\\
V_{1\ldots s} &= \mathrm{Var}(Y) - \sum_{i=1}^{s} V_i - \sum_{1\leq i<j\leq s} V_{ij} - \cdots - \sum_{1\leq i_1<\cdots<i_{s-1}\leq s} V_{i_1\ldots i_{s-1}}.
\end{aligned}$$

The first s terms in Equation (2.4) can be used to compute the first-order Sobol indices. Other terms in the expansion can be used to compute higher-order Sobol indices, for example the *second-order indices*

$$S_{ij} = \frac{V_{ij}}{\mathrm{Var}(Y)},$$

which concern the variance of Y accounted for by the interaction between β_i and β_j. Finally, an additional index, which will be considered in this thesis, was introduced by Homma & Saltelli (1996), namely the *total-order Sobol index* for parameter i, denoted S_{Ti}. This index is a sum of all contributions to the model variance associated with β_i, *i.e.*

$$S_{Ti} = S_i + S_{ij} + S_{ik} + \cdots + S_{ijk} + \cdots + S_{i\ldots s}.$$

In this thesis, total-order Sobol indices will be computed in *Python*, making use of the package SALib.

2.5 Bayesian methods

In this section, two Bayesian inference methods are introduced, which allow the user to infer parameters of a mathematical model, given observed data. Both are variations of a method known as *Approximate Bayesian Computation* (ABC), which is based upon Bayes' theorem. Bayes' theorem is introduced by Blitzstein & Hwang (2019) and in many other probability and statistics texts, as a way of relating conditional probabilities of events A and B, and is given as

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

In statistical inference, Bayes' theorem is instead formulated as

$$\pi(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{D}|\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \pi(\boldsymbol{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{2.5}$$

where $\boldsymbol{\theta}$ is a model parameter, or vector of model parameters, and $\boldsymbol{D}$ is the observed data. In this formulation, $\pi(\boldsymbol{\theta})$ is known as the *prior distribution*, and encodes the users prior beliefs about the parameter(s). If the user has a strong prior knowledge of a parameter value then an *informative* prior can be used, such as a normal or beta distribution, giving more density to regions of the parameter space in which the true value is thought to lie. If however, the user has very little prior knowledge of a parameter value then a *non-informative* prior can be used, such as a uniform distribution, which would only require an upper and lower bound for the parameter value. $\pi(\boldsymbol{D}|\boldsymbol{\theta})$ is the *likelihood* of observing the data $\boldsymbol{D}$, given the parameter(s) $\boldsymbol{\theta}$. Finally, $\pi(\boldsymbol{\theta}|\boldsymbol{D})$ is the *posterior distribution* that the user aims to evaluate. The integral on the denominator of Equation (2.5) is just a normalisation constant, and hence a simpler form of Equation (2.5) is the proportionality equation,

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

Two methods of estimating the posterior distribution are given in Sections 2.5.1 and 2.5.2 which do not require computing the likelihood function, something that is often difficult to compute for mathematical models.

Often when modelling biological systems, a situation can arise in which two or more mathematical models could, in theory, describe the underlying system and one needs to decide which model is the most viable. Bayesian methods, namely Bayesian model selection, based upon approximate Bayesian computation, can be used to solve this problem and two such methods are explained in Section 2.5.3. The methods introduced in Sections 2.5.1, 2.5.2 and 2.5.3 are informed by Toni *et al.* (2009), however many others have explained and applied such methods in the literature (Beaumont *et al.*, 2002; Filippi *et al.*, 2013; Simola *et al.*, 2020; Sunnåker *et al.*, 2013).

2.5.1 Approximate Bayesian computation - rejection algorithm

Given observed data, $\boldsymbol{D}$, and a mathematical model, $\mathcal{M}$, parametrised by the vector, $\boldsymbol{\theta}$, the aim of approximate Bayesian computation (ABC) is to use model simulations to infer posterior distributions for the parameter values. The method allows the user to combine their prior beliefs about the parameters, $\pi(\boldsymbol{\theta})$, with comparisons between the data and model simulations to arrive at the posterior distributions. From a sampled parameter set $\boldsymbol{\theta^*} \sim \pi(\boldsymbol{\theta})$, one can simulate data from the model, $\boldsymbol{D}^* \sim \pi(\boldsymbol{D}|\boldsymbol{\theta}^*)$ and compare this simulated data with the observed data $\boldsymbol{D}$. If the simulated data is sufficiently close to the observed data, where sufficiently close is determined by some distance measure, $\delta(\cdot,\cdot)$, then the sample $(\boldsymbol{\theta}^*, \boldsymbol{D}^*)$ is accepted, otherwise it is rejected. The method continues by repeating this process until an accepted sample of size N is reached. The ABC rejection algorithm is summarised in Algorithm 2.

2.5.2 Approximate Bayesian computation - Sequential Monte Carlo

A major downfall of the ABC rejection algorithm is its computational inefficiency. In situations in which the parameter space is large, corresponding to a large number of model parameters and/or prior distributions spanning a large interval, the ABC rejection algorithm can be very slow to converge. This is because many

Algorithm 2 ABC rejection algorithm (Toni *et al.*, 2009).

1: Choose the posterior sample size N, the acceptance threshold ε, the distance measure $\delta(\cdot,\cdot)$ and set $n = 0$.
2: **while** $n < N$ **do**:
3: Sample $\boldsymbol{\theta}^*$ from $\pi(\boldsymbol{\theta})$.
4: Simulate a dataset $\boldsymbol{D}^*$ from $\pi(\boldsymbol{D}|\boldsymbol{\theta}^*)$.
5: If $\delta(\boldsymbol{D}, \boldsymbol{D}^*) \leq \varepsilon$, accept $\boldsymbol{\theta}^*$ and set $n = n + 1$.
6: **end while**

simulations are required in order to scan the whole parameter space effectively, and with a large number of parameters, the probability of sampling an acceptable parameter set is low, for each iteration. On this basis, Toni *et al.* (2009) developed the ABC - sequential Monte Carlo (ABC-SMC) method, an iterative method in which the user applies ABC multiple times in order to converge to the posterior distributions faster and more efficiently.

Using the same notation for the model, data and distance threshold as in Section 2.5.1, here a decreasing sequence of threshold values $\varepsilon_1 > \varepsilon_2 > \cdots > \varepsilon_Z$ is introduced, where Z is the number of iterations of the ABC to be run. Under this notation, in each iteration of the ABC one generates an accepted sample of size N, where each of the N elements, $\boldsymbol{\theta}^*$, is referred to as a *particle*. A whole sample of particles of size N will be referred to as a *population*, and the method iterates until there are Z populations, of accepted samples, where population Z comprises the final posterior distributions with all particles resulting in a distance measure $\delta(\boldsymbol{D}, \boldsymbol{D}^*) \leq \varepsilon_Z$.

In the first iteration, a rejection ABC is performed where the parameters are sampled from $\pi(\boldsymbol{\theta})$ and are accepted if they result in a distance measure $\delta(\boldsymbol{D}, \boldsymbol{D}^*) \leq \varepsilon_1$. Each particle in the first posterior distribution is assigned an equal weight, $w_1^{(n)} = 1/N$ for $n = 1, \ldots, N$. In each of the following iterations of the ABC, the parameters $\boldsymbol{\theta}^*$ are sampled from the posterior distributions of the previous iteration with weights $\mathbf{w}_{z-1}$, where z is an index for the iteration of the algorithm. The parameters are perturbed using a perturbation kernel $K_z(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$

and the model is simulated using the perturbed parameters and compared with the observed data. The perturbed parameter set is accepted into population z, if it results in a distance measure $\delta(\boldsymbol{D}, \boldsymbol{D}^*) \leq \varepsilon_z$. Each particle in the population z ($z \geq 2$) is assigned a weight based on the prior densities, the weights at the previous iteration and the perturbation kernel. This procedure continues until Z populations of size N are reached. The intuition for this method is that, in each round of the ABC one samples from a smaller parameter space than in the previous round, informed by the posterior distributions in the previous round. The perturbation kernel can be chosen by the user, and some common examples are the component wise uniform kernel and the multivariate normal kernel (Filippi *et al.*, 2013). The ABC-SMC algorithm is summarised in Algorithm 3.

2.5.3 Bayesian model selection

Bayesian model selection is a method of determining which of two or more mathematical models is most likely to describe the observed data, $\boldsymbol{D}$. Let us assume that there are two potential models, $\mathcal{M}_1$ and $\mathcal{M}_2$ which could describe the biological system from which the observed data is obtained. The aim of Bayesian model selection is to determine the *Bayes factor*,

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

where $\pi(\mathcal{M}_i|\boldsymbol{D})$ is the marginal posterior distribution of model $\mathcal{M}_i$, $i \in \{1, 2\}$ and $\pi(\mathcal{M}_i)$ is the prior for model $\mathcal{M}_i$, $i \in \{1, 2\}$. Assuming uniform priors, the Bayes factor becomes

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

and provides the odds in favour of $\mathcal{M}_1$ over $\mathcal{M}_2$. Algorithm 4 is based on the ABC rejection algorithm and provides a method of estimating the Bayes factor, B_{12}, where the estimated value is denoted $\hat{B}_{12}$.

Similarly to the rejection ABC method for parameter estimation, this method of model selection can be very inefficient for large and complex mathematical models. Hence, Toni *et al.* (2009) introduced a model selection algorithm based on the ABC-SMC method. The algorithm is very similar to the ABC-SMC method,

Algorithm 3 ABC-SMC algorithm (Toni *et al.*, 2009).

1: Choose the posterior sample size N, the sequence of threshold values $\varepsilon_1 > \cdots > \varepsilon_Z$ and the distance measure $\delta(\cdot,\cdot)$. Set the population indicator $z = 1$ and the particle indicator $n = 0$.

2: **while** $z < Z$ **do**:

3: **while** $n < N$ **do**:

4: If $z = 1$, sample $\boldsymbol{\theta}^{**}$ from $\pi(\boldsymbol{\theta})$. Else sample $\boldsymbol{\theta}^{*}$ from the previous populations posterior distributions, $\{\boldsymbol{\theta}_{z-1}^{(k)}\}$ for $k = 1, \ldots, N$, with weights $\mathbf{w}_{z-1}$ and perturb to obtain $\boldsymbol{\theta}^{**} \sim K_z(\boldsymbol{\theta}|\boldsymbol{\theta}^{*})$. If $\pi(\boldsymbol{\theta}^{**}) = 0$, re-sample $\boldsymbol{\theta}^{*}$ until $\pi(\boldsymbol{\theta}^{**}) \neq 0$.

5: Simulate a dataset $\boldsymbol{D}^{*}$ from $\pi(\boldsymbol{D}|\boldsymbol{\theta}^{**})$.

6: If $\delta(\boldsymbol{D}, \boldsymbol{D}^{*}) \leq \varepsilon_z$, set $\boldsymbol{\theta}_z^{(n)} = \boldsymbol{\theta}^{**}$ and set $n = n + 1$. Calculate the weight for particle $\boldsymbol{\theta}_z^{(n)}$ as,

$$w_z^{(n)} = \begin{cases} 1 & \text{if } z = 1, \\ \dfrac{\pi(\boldsymbol{\theta}_z^{(n)})}{\sum_{j=1}^{N} w_{z-1}^{(j)} K_z(\boldsymbol{\theta}_z^{(n)}|\boldsymbol{\theta}_{z-1}^{(j)})} & \text{if } z > 1. \end{cases}$$

7: **end while**

8: Normalise the weights and set $z = z + 1$.

9: **end while**

whereby the parameters are sampled iteratively from the posterior distributions of the previous iteration, perturbed, and assigned weights if they are accepted. As a step prior to sampling the parameters however, a model, $\mathcal{M}_i$, $i \in \{1, 2\}$, is firstly sampled from a prior distribution, $\pi(\mathcal{M})$. For example, if the user is selecting between two models, $\mathcal{M}_i$, $i \in \{1, 2\}$, with no prior information about which model is the most likely, then a discrete uniform distribution can be used as a prior distribution for the models with only two values $1 = \mathcal{M}_1$ and $2 = \mathcal{M}_2$, where each model then has prior probability, $\frac{1}{2}$. The parameters are sampled

Algorithm 4 Bayesian model selection algorithm (Toni *et al.*, 2009).

1: Choose the posterior sample size N, the acceptance threshold ε and the distance measure $\delta(\cdot,\cdot)$. Set $i = 1$ and $r_1 = r_2 = 0$.
2: **while** $i < 2$ **do**:
3: Set n = 0.
4: **while** $n < N$ **do**:
5: Sample $\boldsymbol{\theta}_i^*$ from $\pi(\boldsymbol{\theta}_i)$.
6: Simulate a dataset $\boldsymbol{D}_i^*$ from $\pi(\boldsymbol{D}|\boldsymbol{\theta}_i^*)$.
7: If $\delta(\boldsymbol{D}, \boldsymbol{D}_i^*) \leq \varepsilon$, set $r_i = r_i + 1$.
8: Set $n = n + 1$.
9: **end while**
10: Set $i = i + 1$.
11: **end while**
12: Compute $p_i = \frac{r_i}{N}$ for $i \in \{1, 2\}$.
13: **return** $\hat{B}_{12} = \frac{p_1}{p_2}$.

for this particular model from either the prior distributions (first iteration) or the previous iterations posteriors (iteration $z = 2, \ldots, Z$, in which case they are also perturbed), and again the particle is accepted if the model simulation has distance $\delta(\boldsymbol{D}, \boldsymbol{D}_i^*) \leq \varepsilon_z$ for iteration z. This procedure continues until a sample of size N is reached consisting of parameters sets from any model. If the models are initially given equal weight, then for a large enough sample size N, they will each be sampled roughly the same number of times. However, as the iterations of the algorithm proceed, one would expect that if one model results in simulations that better resemble the data, this model will have more parameter sets accepted per iteration as z approaches Z. The ABC-SMC model selection algorithm is given in Algorithm 5. In this algorithm, for a specific model m^*, $\boldsymbol{\theta}(m^*)$ denotes a particle relating to the model m^*. Likewise, $\boldsymbol{D}(m^*)^*$ denotes a simulated dataset from model m^*, and $\mathbf{w}(m^*)$ denotes weights corresponding to particles accepted into a sample for model m^*.

Algorithm 5 ABC-SMC model selection algorithm (Toni *et al.*, 2009).

1: Choose the posterior sample size N, the sequence of threshold values $\varepsilon_1 > \cdots > \varepsilon_Z$ and the distance measure $\delta(\cdot,\cdot)$. Set the population indicator $z = 1$ and the particle indicator $n = 0$.

2: **while** $z < Z$ **do**:

3: **while** $n < N$ **do**:

4: Sample a model m^* from $\pi(\mathcal{M})$.

5: If $z = 1$, sample $\boldsymbol{\theta}^{**}$ from $\pi(\boldsymbol{\theta}(m^*))$. Else sample $\boldsymbol{\theta}^*$ from the previous populations posterior distributions, $\{\boldsymbol{\theta}(m^*)_{z-1}^{(k)}\}$ for $k = 1, \ldots, N$, with weights $\mathbf{w}(m^*)_{z-1}$ and perturb to obtain $\boldsymbol{\theta}^{**} \sim K_z(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$. If $\pi(\boldsymbol{\theta}^{**}) = 0$, re-sample $\boldsymbol{\theta}^*$ until $\pi(\boldsymbol{\theta}^{**}) \neq 0$.

6: Simulate a dataset $\boldsymbol{D}(m^*)^*$ from $\pi(\boldsymbol{D}|\boldsymbol{\theta}^{**}, m^*)$.

7: If $\delta(\boldsymbol{D}, \boldsymbol{D}(m^*)^*) \leq \varepsilon_z$, set $\boldsymbol{\theta}(m^*)_z^{(n)} = \boldsymbol{\theta}^{**}$ and set $n = n + 1$. Set $\boldsymbol{m}_z^{(n)} = m^*$ and calculate the weight for particle $\boldsymbol{\theta}(m^*)_z^{(n)}$ as,

$$w_z^{(n)} = \begin{cases} 1 & \text{if } z = 1, \\ \frac{\pi(\boldsymbol{\theta}_z^{(n)})}{\sum_{j=1}^{N} w_{z-1}^{(j)} K_z(\boldsymbol{\theta}_z^{(n)}|\boldsymbol{\theta}_{z-1}^{(j)})} & \text{if } z > 1. \end{cases}$$

8: **end while**

9: Normalise the weights and set $z = z + 1$.

10: **end while**

The output of Algorithm 5 is a series of N dimensional vectors $\mathbf{m}_z$, for $z = 1, \ldots, Z$. When using the algorithm to select between two models $\mathcal{M}_i$, $i \in \{1, 2\}$, each vector $\mathbf{m}_z$ will contain the numbers 1 and 2, and one can compute,

$$f_i^z = \text{frequency of the value } i \text{ in vector } \mathbf{m}_z, \text{ for } i \in \{1, 2\} \text{ and } z = 1, \ldots, Z.$$

Using these frequencies, the relative probability of each model at each iteration

can be computed as

$$p_i^z = \frac{f_i^z}{N}, \text{ for } i \in \{1, 2\} \text{ and } z = 1, \ldots, Z.$$

Thus an approximation of the Bayes factor B_{12} can be found iteratively using Algorithm 5 as

$$\hat{B}_{12}^z = \frac{p_1^z}{p_2^z}, \text{ for } z = 1, \ldots, Z.$$

2.6 Statistical analysis

In this section, statistical tests are introduced, which will be used in Chapter 6. In particular, analysis of variance (ANOVA) is a method used to find differences between the means of two or more groups of data. When there are more than two groups, the result of an ANOVA will conclude whether or not there is a statistically significant difference between the means of at least one pair of the groups, but it will not tell us which pairing was significantly different. To infer this information, one can use a post-hoc analysis, such as Tukey's honest significant difference (HSD) test, as is introduced in Section 2.6.2.

2.6.1 Analysis of variance (ANOVA)

Here, ANOVA is described, based on the introduction given by Marques de Sá (2003).

One-way ANOVA

Assuming that there are c independent samples, an ANOVA tests whether the null hypothesis, that the means of the c groups are equal, is true, *i.e.*

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_c,$$

against the alternative hypothesis, that the means of at least one pair are unequal,

$$H_1 : \mu_i \neq \mu_j,$$

for some pair i, j. The simplest form of this method is one-way ANOVA, which is applied when there is only one categorical grouping level. The variable being tested is known as the *dependent variable*, and the variable with the groupings is known as the *independent variable*. Given that ANOVA tests only for differences in the means of the groups of data, an assumption of the test is that the variance in each group is equal and the data in each group is normally distributed. The general principle of the method is to decompose the total variance of the data, into contributions to the variance, within-groups, and between-groups.

Suppose one has a sample of size n, split into c groups with sizes $n_1, n_2, \ldots, n_c$, and with sample means $\bar{x_1}, \bar{x_2}, \ldots, \bar{x_c}$. Any value in the total sample can be denoted x_{ij}, where $i \in \{1, \ldots, c\}$ is the group index and $j \in \{1, \ldots, n_i\}$ is an index for the value within the group, with n_i being the number of observations in group i. The total variance is then related to the total sum of squares (SST) of deviations from the global sample mean $\bar{x}$, where

$$\text{SST} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2.$$

Adding and subtracting $\bar{x_i}$ to the deviations $x_{ij} - \bar{x}$, one can arrive at the following expression,

$$\text{SST} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ij} - \bar{x_i})^2 + \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\bar{x_i} - \bar{x})^2 + 2 \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ij} - \bar{x_i})(\bar{x_i} - \bar{x}).$$

The last term in this expression can be shown to be equal to 0, and hence,

$$\text{SST} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ij} - \bar{x_i})^2 + \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\bar{x_i} - \bar{x})^2. \tag{2.6}$$

The first term in Equation (2.6) is known as the *within-group sum of squares* (SSW) or the *error sum of squares* (SSE), since it represents the deviations of the values in a particular group from the sample mean of that group. The second term in the equation is known as the *between-group sum of squares* (SSB), since it is accounting for the deviations of the group means from the global mean.

Therefore one can write Equation (2.6) as

$$\text{SST} = \text{SSW} + \text{SSB},$$

i.e. the total sum of squares is contributed to by the within-group sum of squares and the between-group sum of squares. Each sum of squares can then be expressed in terms of variances as,

$$(n-1)v = (n-c)v_W + (c-1)v_B,$$

where v_W is the within-group variance and v_B is the between-group variance. The total variance, v, has $(n-1)$ degrees of freedom (dof), whilst the within-group and between-group variances have $(n-c)$ and $(c-1)$ dof respectively. If the null hypothesis, that the means of the c groups are equal, is *not* true, one would expect the variation between means to be large relative to the variation within samples. The within-group variance, or mean square error (MSE), can be written as

$$v_W = \text{MSE} = \frac{\text{SSW}}{n-c},$$

and the between-group variance, or mean square between (MSB), can be written as

$$v_B = \text{MSB} = \frac{\text{SSB}}{c-1}.$$

Then if the null hypothesis is true, one would expect the ratio,

$$\frac{v_B}{v_W} = \frac{\text{MSB}}{\text{MSE}},$$

to be close to 1, and if there is evidence against the null hypothesis, then the ratio should be significantly larger than 1. This ratio defines the test statistic, and will be compared with the F distribution to test the validity of H_0. Firstly, the sum of squares of k independent random variables, with standard normal distribution, follows the chi-square, $\chi^2(k)$, distribution. The F distribution can then be defined as the ratio of two independent χ^2 random variables. If two

independent random variables Q_1 and Q_2 are χ^2 distributed with d_1 and d_2 degrees of freedom, respectively, then

$$\frac{Q_1/d_1}{Q_2/d_2},$$

follows an $F(d_1, d_2)$ distribution. The test statistic for a one-way ANOVA is

$$F^* = \frac{v_B}{v_W} = \frac{\text{MSB}}{\text{MSE}} \sim F(c-1, n-c),$$

under H_0.

Two-way ANOVA

Whereas a one-way ANOVA only examines the effect of one categorical variable (or grouping) on the dependent variable, a two-way ANOVA examines the effect of two categorical variables. Two-way ANOVA follows the same assumptions as the one-way case, where each sample is independent, there is equality in the variances between samples, and each sample is normally distributed. Let us assume that the first independent variable, R_1, has r_1 groupings and the second independent variable, R_2, has r_2 groupings. There are now three pairs of null and alternative hypotheses:

H_{01} : the means of all r_1 samples defined by R_1 groupings are equal, vs,
H_{11} : there is at least one of the r_1 samples with unequal mean,

H_{02} : the means of all r_2 samples defined by R_2 groupings are equal, vs,
H_{12} : there is at least one of the r_2 samples with unequal mean,

H_{03} : there is no interaction between R_1 and R_2, vs,
H_{13} : there is interaction between R_1 and R_2.

Similarly to the one-way case, here the total variance in the two-way case is split into contributions from the factor R_1, the factor R_2, the interaction between R_1 and R_2 and the within-group variance. Assuming that each grouping has the

same size, n, the total sum of squares (SST) (with $n-1$ dof) can be expressed as

$$\text{SST} = \text{SS}R_1 + \text{SS}R_2 + \text{SS}R_{12} + \text{SSW},$$

where,

$$\begin{aligned}
\text{SS}R_1 &= R_1 \text{ main effect sum of squares with } (r_1 - 1) \text{ dof,} \\
\text{SS}R_2 &= R_2 \text{ main effect sum of squares with } (r_2 - 1) \text{ dof,} \\
\text{SS}R_{12} &= \text{interaction effect sum of squares with } (r_1 - 1)(r_2 - 1) \text{ dof, and} \\
\text{SSW} &= \text{error sum of squares with } n - r_1 r_2 \text{ dof.}
\end{aligned}$$

Two-way ANOVA then has three test statistics, F_1^*, corresponding to the R_1 main effect, F_2^*, corresponding to the R_2 main effect and F_{12}^*, corresponding to the interaction effect. These test statistics can be defined by taking ratios of the following mean squares corresponding to each effect in the model, where,

$$\begin{aligned}
v_{R1} &= \text{MS}R_1 = \frac{\text{SS}R_1}{(r_1 - 1)}, \\
v_{R2} &= \text{MS}R_2 = \frac{\text{SS}R_2}{(r_2 - 1)}, \\
v_{R1R2} &= \text{MS}R_{12} = \frac{\text{SS}R_{12}}{(r_1 - 1)(r_2 - 1)}, \text{ and} \\
v_W &= \text{MSE} = \frac{\text{SSW}}{n - r_1 r_2}.
\end{aligned}$$

Then, the test statistics are defined as

$$\begin{aligned}
F_1^* &= \frac{v_{R1}}{v_W} = \frac{\text{MS}R_1}{\text{MSE}} \sim F(r_1 - 1, n - r_1 r_2), \\
F_2^* &= \frac{v_{R2}}{v_W} = \frac{\text{MS}R_2}{\text{MSE}} \sim F(r_2 - 1, n - r_1 r_2), \text{ and} \\
F_{12}^* &= \frac{v_{R1R2}}{v_W} = \frac{\text{MS}R_{12}}{\text{MSE}} \sim F((r_1 - 1)(r_2 - 1), n - r_1 r_2).
\end{aligned}$$

As in the one-way case, a null hypothesis is rejected if the corresponding F statistic is significantly larger than 1. An ANOVA allows one to determine if

there is a significant difference between the means of groupings defined by a categorical variable, or the interaction between two categorical variables in the two-way case. In the case where a categorical variable has more than two levels, the ANOVA does not however specify, for which of the groupings the significant difference arose. To this end, one can use a post-hoc analysis as will be introduced in the next section.

2.6.2 Tukey's honest significant difference test

In this section, a post-hoc analysis to the ANOVA procedure is introduced, which is a pairwise comparison technique to determine which levels of a categorical variable are statistically significantly different from one another. The following description of the method is based on the information given by Montgomery (2017) and Abdi & Williams (2010). Tukey's honest significant difference (HSD) is the smallest amount by which means must differ from each other to be classed as "truly" different. The test utilises the distribution of the *studentised range statistic*,

$$q = \frac{\bar{y}_{max} - \bar{y}_{min}}{\sqrt{\mathrm{MSE}/n}}$$

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

$$T_\alpha = q_\alpha(a, f)\sqrt{\frac{\mathrm{MSE}}{n}},$$

for significance level α, where a is the number of observations in each group, n is the total number of observations and f is the dof associated with the MSE. The value $q_\alpha(a, f)$ can be obtained from a studentised range distribution table. Then, denoting by $\bar{g}$ and $\bar{g}'$, the means of two groups of some factor G, there is a significant difference between these means at the level α, if

$$|\bar{g} - \bar{g}'| \geq T_\alpha.$$

This test can be conducted for each pair of groups defined by a factor in the ANOVA procedure, to determine the cause of a significant effect.

2.6.3 Principal component analysis

In this section, another statistical technique is discussed, namely principal component analysis (PCA). PCA is a method of summarising, visualising and reducing the dimensionality of a multi-dimensional, highly correlated dataset. The aim is to be able to recast a p-dimensional dataset into p so called *principal components* (PCs), which are linear combinations of the original variables but where some of these PCs hold more variability than a single original variable. In this way one can more easily visualise, by plotting the data transformed to the PCs, different groupings in the dataset. An intuitive explanation of the procedure for finding the principal components is given here, based on the information given by Chatfield & Collins (1981) and Kassambara (2017).

Let $\mathbf{X} = [X_1, \ldots, X_p]$ be a p-dimensional dataset with mean μ and covariance matrix $\mathbf{\Sigma}$. The aim is to transform this dataset to p principal components $Y_1, \ldots, Y_p$ which are uncorrelated and where the variance decreases from Y_1 to Y_p. Each Y_j should be a linear combination of the X's, where

$$\begin{aligned} Y_j &= a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p \\ &= \mathbf{a}_j^T\mathbf{X}, \end{aligned}$$

and $\mathbf{a}_j^T = [a_{1j}, \ldots, a_{pj}]$ is a vector of constants. To find the first PC, $Y_1 = \mathbf{a}_1^T\mathbf{X}$, one must find the constants $\mathbf{a}_1^T$ such that $\mathrm{Var}(\mathbf{a}_1^T\mathbf{X})$ is maximal. A constraint, $\mathbf{a}_j^T\mathbf{a}_j = 1$ is used to ensure that the variance doesn't become unbounded and a similar constraint is used for each further PC. It turns out, by using the method of Lagrange multipliers, that $\mathbf{a}_1$ should be the eigenvector of the sample covariance matrix $\mathbf{\Sigma}$, corresponding to the largest eigenvalue. Similarly, the kth PC is defined by $Y_k = \mathbf{a}_k^T\mathbf{X}$, where $\mathbf{a}_k^T$ is the eigenvector of the sample covariance matrix corresponding to the kth largest eigenvalue.

By using this method to find the PCs, it is often the case that from the p PCs, there are only a few which hold more variability than any single original

variable, and hence these PCs become the focus when looking for further trends in the data.

Chapter 3

A stochastic model of receptor-ligand competition dynamics

Receptors, which often reside on the surface of a cell, play a significant role in determining the fate of cells through the initiation of intracellular phosphorylation (signalling) cascades (Chen *et al.*, 2015; Ernst & Jenkins, 2004; Hackel *et al.*, 1999; Katoh & Katoh, 2006; Santos *et al.*, 2007). Many receptors span the cellular membrane and are comprised of an extracellular domain, a trans-membrane domain and an intracellular domain (Lemmon & Schlessinger, 2010; Pierce *et al.*, 2002). For such receptors, signalling cascades can be initiated by the binding of a ligand molecule, in the extracellular medium, to the extracellular domain of the receptor, forming a receptor-ligand complex. Ligand binding can induce phosphorylation of tyrosine residues on the intracellular tail of the bound receptor, in turn allowing the receptor to interact with downstream signalling proteins in the cellular cytoplasm. The purpose of these signalling cascades can be to determine the fate of the cell, for example cell division, growth, death or migration, and the eventuality which occurs is determined by the specific receptor-ligand interaction. The formation of a single receptor-ligand complex is often not sufficient to push a cell to its fate and in fact it is the strength of the signalling that determines cellular fate, where in some systems this strength can be assumed to be proportional to the number of receptor-ligand complexes formed (Starbuck &

Lauffenburger, 1992). It is therefore important to quantify the number, and time scales of formation, of different receptor-ligand complexes on the cell membrane.

Often, ligand molecules are capable of binding to more than one type of receptor (Chen *et al.*, 2015; Weddell & Imoukhuede, 2017), resulting in an element of natural competition for this shared resource. For example, a ligand which is shared by two receptors is the vascular endothelial growth factor A (VEGF-A), which can bind two different vascular endothelial growth factor receptors (VEGFRs), namely VEGFR1 and VEGFR2 (Cross *et al.*, 2003). Although both receptors bind the same ligand, they have been shown to have different functions in normal and tumour vasculature (Dikov *et al.*, 2005; Shibuya, 2006). Shibuya (2006) explains how, during embryogenesis, these receptors have opposing functions upon VEGF-A stimulation; VEGFR1 is a negative regulator for angiogenesis (the formation of new blood cells), whereas VEGFR2 is a positive regulator. In the same reference, it is argued that the receptors have differing roles in pathology and the author suggests that it may be useful, in the control of disease, to be able to inhibit the receptors individually. In the development of such inhibitors it is crucial to be able to quantify the number, and time scales to formation, of receptor-ligand complexes.

In this chapter, a stochastic model for the competition between two receptors for a common ligand is introduced, in terms of a continuous-time birth-and-death Markov chain. Stochastic models can be relevant for some ligand molecules, such as pharmaceuticals or naturally produced cytokines (for example, IL-2), which can produce a cellular response at a very low dose (Gurevich *et al.*, 2003). A stochastic model for the competition dynamics between two species for a shared resource was first introduced and analysed by Reuter (1961) in the area of Mathematical Ecology. Iglehart *et al.* (1964) further generalised this idea to a multivariate competition process with two or more variables (*i.e.* species). The reactions between two receptors with a common ligand lead to a bi-variate stochastic process of the kind considered by Reuter (1961), and the aim of this chapter is to compute the steady state distribution of the number of receptor-ligand complexes of each type, and to quantify the time scales of formation of a particular type of receptor-ligand complex. Such properties of a Markov chain are here collectively referred to as *stochastic descriptors*.

In Sections 3.4.1 and 3.5.1, previously developed methods, namely matrix-analytic methods (Latouche *et al.*, 1999; López-García *et al.*, 2018), are introduced and used to compute the stochastic descriptors of interest. The main limitation of these methods however, is their practical computational implementation. The state space of a bi-variate stochastic process increases with the number of molecules in the system, where large numbers are typical in experimental and physiological settings (Cross *et al.*, 2003; Shibuya, 2006). When more than two receptor species are present, or when the intracellular dynamics are also considered in the mathematical model, the dimensionality and number of states are further increased. Thus, in Sections 3.4.2, 3.4.3, 3.5.2 and 3.5.3, analytical approximations for the stochastic descriptors are introduced. These approximations, for the two-dimensional system of interest (*i.e.* two kinds of receptor competing for a common ligand), are motivated by the fact that, under excess of ligand, the processes representing receptor-ligand complex formation for the two different receptors are approximately independent. The accuracy of these approximations is assessed by means of numerical comparison to the matrix-analytic results, for a wide range of biologically feasible parameter values and numbers of molecules.

3.1 A stochastic competition model

In this section, a stochastic mathematical model for the competition between two receptors for a common ligand is introduced. There are two receptor types, R_1 and R_2, assumed to be residing across the cell membrane, and a shared ligand, L, in the extracellular medium. Both receptors can associate with the ligand to form a receptor-ligand complex, denoted by M_1 or M_2, depending on the receptor type involved. For brevity, receptor-ligand complexes will be referred to only as complexes for the remainder of this chapter. It is assumed that there is a spatially homogeneous distribution of molecules on the cell membrane and mass action kinetics apply (Steinfeld *et al.*, 1989). The complex formation reactions occur with forward association rates $k_{f,1}$ and $k_{f,2}$, respectively. The complexes, M_1 and M_2, are allowed to dissociate into their constituent receptor and ligand, with rates $k_{r,1}$ and $k_{r,2}$, respectively (see Figure 3.1).

3. A STOCHASTIC MODEL OF RECEPTOR-LIGAND COMPETITION DYNAMICS

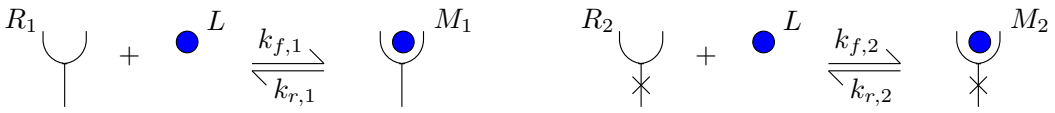

Figure 3.1: A depiction of the molecular reactions underlying the stochastic mathematical model. Two different types of receptor molecule can bind, reversibly, with a shared ligand to form two different complex types.

The dimensionality of the process can be reduced, and hence the model simplified, by considering the total number of receptors available in the system, denoted by $n_{R,1}$ and $n_{R,2}$, and the total number of ligands available, denoted by n_L. It is clear that $n_{R,1} = R_1(t) + M_1(t)$ and $n_{R,2} = R_2(t) + M_2(t)$, for all $t \geq 0$, where $R_1(t)$ and $R_2(t)$ represent the number of free receptors of type 1 and 2, respectively, at time t, and $M_1(t)$ and $M_2(t)$ represent the number of complexes of type 1 and 2, respectively, at time t. Then, the process can be described as a bi-variate continuous-time Markov chain (CTMC) (see Section 2.2.1 for details) $\mathcal{X} = \{\mathbf{X}(t) = (M_1(t), M_2(t)) : t \geq 0\}$, with $M_1(t), M_2(t) \geq 0$ for all $t \geq 0$, and the process $\mathcal{X}$ evolves over the state space $\mathcal{S}_\mathcal{X} = \{(m_1, m_2) \in (\mathbb{N} \cup \{0\})^2 : \quad m_1 \leq n_{R,1}, \quad m_2 \leq n_{R,2}, \quad m_1 + m_2 \leq n_L\}$, where m_1 and m_2 are the number of type 1 and type 2 complexes, respectively, at any given time.

The dynamics of complex formation and dissociation are then represented by *jumps*, or transitions, between states in $\mathcal{S}_\mathcal{X}$, where the notation $(m_1, m_2) \to (m_1', m_2')$ implies a transition in one step from state (m_1, m_2) to state (m_1', m_2'). The transition diagram is shown in Figure 3.2 and the infinitesimal transition rate from state (m_1, m_2) to state (m_1', m_2'), introduced in Definition 17, by assuming mass action kinetics, is given by

$$q_{(m_1,m_2),(m_1',m_2')} = \begin{cases} k_{f,1}(n_{R,1} - m_1)(n_L - m_1 - m_2), & \text{if } (m_1', m_2') = (m_1 + 1, m_2), \\ k_{r,1} m_1, & \text{if } (m_1', m_2') = (m_1 - 1, m_2), \\ k_{f,2}(n_{R,2} - m_2)(n_L - m_1 - m_2), & \text{if } (m_1', m_2') = (m_1, m_2 + 1), \\ k_{r,2} m_2, & \text{if } (m_1', m_2') = (m_1, m_2 - 1), \\ 0, & \text{otherwise.} \end{cases}$$

The number of states in $\mathcal{S}_{\mathcal{X}}$ is

$$\begin{aligned}\#\mathcal{S}_{\mathcal{X}} &= \sum_{k=0}^{N_2}(min(N_1, n_L - k) + 1) \\ &= \sum_{k=0}^{N_2}(min(n_{R,1}, n_L - k) + 1),\end{aligned}$$

where $N_1 = min(n_{R,1}, n_L)$ and $N_2 = min(n_{R,2}, n_L)$. Explicit formulae for the number of states in $\mathcal{S}_{\mathcal{X}}$ can be derived and depend on the values of $n_{R,1}$, $n_{R,2}$ and n_L, where

$$\#\mathcal{S}_{\mathcal{X}} = \begin{cases} (n_{R,1}+1)(n_{R,2}+1), & \text{if } n_{R,1}+n_{R,2} \leq n_L, \\ \frac{2(n_L-n_{R,1}+1)(n_{R,1}+1)+(n_{R,1}+n_{R,2}-n_L)(n_{R,1}+n_L-n_{R,2}+1)}{2}, & \text{if } n_{R,1} \leq n_{R,2} \leq n_L \\ & \text{and } n_{R,1}+n_{R,2} > n_L, \\ \frac{2(n_L-n_{R,2}+1)(n_{R,2}+1)+(n_{R,1}+n_{R,2}-n_L)(n_{R,2}+n_L-n_{R,1}+1)}{2}, & \text{if } n_{R,2} \leq n_{R,1} \leq n_L \\ & \text{and } n_{R,1}+n_{R,2} > n_L, \\ \frac{(n_{R,2}+1)(2n_L-n_{R,2}+2)}{2}, & \text{if } n_{R,1} > n_L \\ & \text{and } n_{R,2} < n_L, \\ \frac{(n_{R,1}+1)(2n_L-n_{R,1}+2)}{2}, & \text{if } n_{R,1} < n_L \\ & \text{and } n_{R,2} > n_L, \\ \frac{(n_L+1)(n_L+2)}{2}, & \text{if } n_{R,1} \geq n_L \\ & \text{and } n_{R,2} \geq n_L. \end{cases}$$

The model depicted in Figure 3.1 can be analysed by considering the probabilities $p_{(m_1,m_2)}(t)$ (see Section 2.2.2) for the process $\mathcal{X}$, where for initial state $(M_1(0), M_2(0))$,

$$p_{(m_1,m_2)}(t) = \mathbb{P}(M_1(t) = m_1, M_2(t) = m_2) \qquad \forall (m_1, m_2) \in \mathcal{S}_{\mathcal{X}},\ t \geq 0.$$

A differential equation with respect to t, for these probabilities can be written

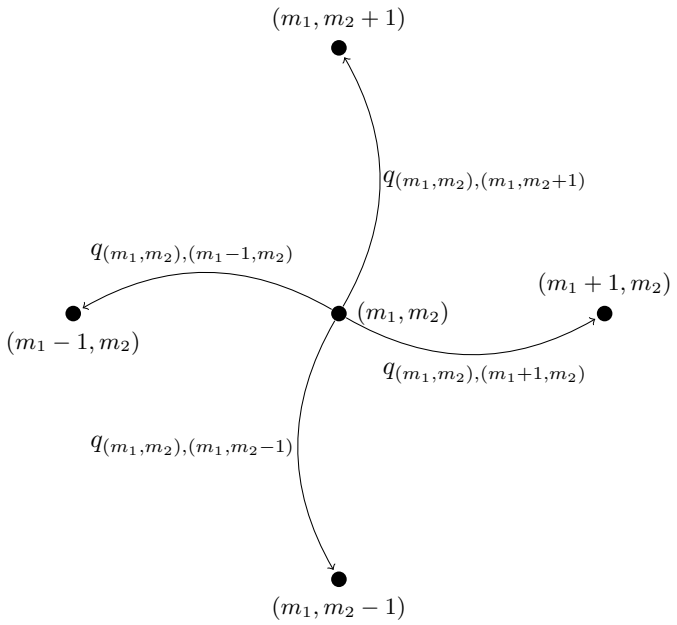

Figure 3.2: Transition diagram for the process $\mathcal{X}$, showing the possible states which the process can move to from a general state (m_1, m_2) and the transition rates with which these state moves occur.

as

$$\begin{aligned}\frac{dp_{(m_1,m_2)}}{dt} &= k_{f,1}(n_{R,1}-m_1+1)(n_L-m_1-m_2+1)p_{(m_1-1,m_2)} + k_{r,1}(m_1+1)p_{(m_1+1,m_2)} \\ &+ k_{f,2}(n_{R,2}-m_2+1)(n_L-m_1-m_2+1)p_{(m_1,m_2-1)} + k_{r,2}(m_2+1)p_{(m_1,m_2+1)} \\ &- [k_{f,1}(n_{R,1}-m_1)(n_L-m_1-m_2) + k_{r,1}m_1 \\ &+ k_{f,2}(n_{R,2}-m_1)(n_L-m_1-m_2) + k_{r,2}m_2]p_{(m_1,m_2)} \end{aligned} \tag{3.1}$$

by considering the infinitesimal transition rates $q_{(m_1,m_2),(m_1',m_2')}$, where t has been omitted from the notation for ease of reading. This equation is known as the chemical master equation (CME) (see Gillespie (1992) and Section 2.2.5), and is known to be challenging to solve analytically. Thus firstly, the model dynamics are explored here by means of stochastic (Gillespie) simulations, as introduced in Section 2.2.6, and the use of the linear noise approximation (also known as the system size expansion); see Section 2.2.6 and Van Kampen (1976, 1992).

3.2 Linear noise approximation

The linear noise approximation (LNA), introduced by Van Kampen (1976), is a method of approximately solving nonlinear CMEs such as Equation (3.1). It provides a second order approximation to the CME via a large volume expansion around the steady state. To begin, an operator $E_i^{\pm 1}$ is introduced which, when operating on an arbitrary function f of i, changes the number of i by ± 1 (Hayot & Jayaprakash, 2004):

$$E_i^{+1}[f(i)] = f(i+1), \qquad E_i^{-1}[f(i)] = f(i-1).$$

Equation (3.2) shows the CME written in terms of $E_i^{\pm 1}$,

$$\begin{aligned}\frac{dp_{(m_1,m_2)}}{dt} &= k_{f,1}(E_{m_1}^{-1}-1)[(n_{R,1}-m_1)(n_L-m_1-m_2)p_{(m_1,m_2)}] \\ &+ k_{r,1}(E_{m_1}^{+1}-1)[m_1 p_{(m_1,m_2)}] \\ &+ k_{f,2}(E_{m_2}^{-1}-1)[(n_{R,2}-m_2)(n_L-m_1-m_2)p_{(m_1,m_2)}] \\ &+ k_{r,2}(E_{m_2}^{+1}-1)[m_2 p_{(m_1,m_2)}]. \end{aligned} \tag{3.2}$$

In a specified volume, such as the volume inside a cell, the importance of studying the mesoscopic dynamics of a species population is dependent on the number of molecules of such a species. It is widely accepted that if the average size of a population is n, then the size of the fluctuations around n is $\sqrt{n}$ (Elf & Ehrenberg, 2003b). In order to study the stochastic fluctuations around the steady state, and to convert from a macroscopic analysis to a mesoscopic one, the CME is Taylor expanded in powers of $\sqrt{\Psi}^{-1}$, where Ψ is a parameter chosen to be the volume of the system. Firstly, the variables of the system are re-scaled so that they are written as a sum of the macroscopic concentration values and the mesoscopic fluctuations and in terms of the parameter Ψ. If m_1, m_2, $n_{R,1}$, $n_{R,2}$ and n_L are the numbers of molecules of complexes of type 1, complexes of type 2, total receptors of type 1, total receptors of type 2 and total ligands, respectively, then ϕ_1, ϕ_2, $N_{R,1}$, $N_{R,2}$ and N_L, denote the concentrations of the same species. The variables are then re-scaled as

$$m_1 = \Psi\phi_1 + \Psi^{\frac{1}{2}}\xi_1,$$

$$
\begin{aligned}
m_2 &= \Psi\phi_2 + \Psi^{\frac{1}{2}}\xi_2, \\
n_{R,1} &= \Psi N_{R,1}, \\
n_{R,2} &= \Psi N_{R,2}, \text{ and} \\
n_L &= \Psi N_L,
\end{aligned}
$$

where ξ_1 and ξ_2 represent the fluctuations in the number of type 1 and type 2 complexes, respectively. The forward rate constants $k_{f,1}$ and $k_{f,2}$ are also dependent on the volume of the system and hence are re-scaled as

$$
\begin{aligned}
k_{f,1} &= \Psi^{-1}\hat{k}_{f,1}, \text{ and} \\
k_{f,2} &= \Psi^{-1}\hat{k}_{f,2},
\end{aligned}
$$

so that they represent true rate constants. Then, transforming from the variables (m_1, m_2) to (ξ_1, ξ_2),

$$
p_{(m_1,m_2)} = p_{(\Psi\phi_1+\Psi^{\frac{1}{2}}\xi_1, \Psi\phi_2+\Psi^{\frac{1}{2}}\xi_2)} = \Pi_{(\xi_1,\xi_2)}.
$$

An equation can be found for the partial derivative of the transformed probability function, $\Pi_{(\xi_1,\xi_2)}$, with respect to time. To do this one must take the total time derivative of $\Pi_{(\xi_1,\xi_2)}$ since ξ_1 and ξ_2 depend on time, giving

$$
\frac{d}{dt}p_{(m_1,m_2)} = \frac{\partial}{\partial t}\Pi_{(\xi_1,\xi_2)} + \frac{\partial}{\partial \xi_1}\Pi_{(\xi_1,\xi_2)}\frac{d\xi_1}{dt} + \frac{\partial}{\partial \xi_2}\Pi_{(\xi_1,\xi_2)}\frac{d\xi_2}{dt}. \tag{3.3}
$$

Then, since the derivative $\frac{d}{dt}p_{(m_1,m_2)}$ is taken with fixed m_1 and m_2, it can be written that

$$
\begin{aligned}
0 = \frac{dm_1}{dt} &= \Psi\frac{d\phi_1}{dt} + \Psi^{\frac{1}{2}}\frac{d\xi_1}{dt}, \text{ and} \\
0 = \frac{dm_2}{dt} &= \Psi\frac{d\phi_2}{dt} + \Psi^{\frac{1}{2}}\frac{d\xi_2}{dt},
\end{aligned}
$$

and hence,

$$
\frac{d\xi_1}{dt} = -\Psi^{\frac{1}{2}}\frac{d\phi_1}{dt}, \text{ and} \tag{3.4}
$$

$$\frac{d\xi_2}{dt} = -\Psi^{\frac{1}{2}}\frac{d\phi_2}{dt}. \tag{3.5}$$

Substituting expressions (3.4) and (3.5) into Equation (3.3), reformulates the left hand side of Equation (3.2) in terms of the new variables ξ_1 and ξ_2 and the volume Ψ, giving

$$\frac{d}{dt}p_{(m_1,m_2)} = \frac{\partial}{\partial t}\Pi_{(\xi_1,\xi_2)} - \Psi^{\frac{1}{2}}\frac{d\phi_1}{dt}\frac{\partial}{\partial \xi_1}\Pi_{(\xi_1,\xi_2)} - \Psi^{\frac{1}{2}}\frac{d\phi_2}{dt}\frac{\partial}{\partial \xi_2}\Pi_{(\xi_1,\xi_2)}.$$

In order to reformulate the right hand side of Equation (3.2), it is convenient to approximate the step operator function $E_i^{\pm 1}$ as a Taylor expansion, where

$$E_i^{\pm 1} = 1 \pm \frac{\partial}{\partial i} + \frac{1}{2}\frac{\partial^2}{\partial i^2} \pm \cdots,$$

and then by the chain rule and using the fact that $\frac{\partial m_i}{\partial \xi_i} = \Psi^{\frac{1}{2}}$ for $i \in \{1, 2\}$,

$$E_{m_1}^{\pm 1} = 1 \pm \Psi^{-\frac{1}{2}}\frac{\partial}{\partial \xi_1} + \frac{1}{2}\Psi^{-1}\frac{\partial^2}{\partial \xi_1^2} + \cdots, \text{ and} \tag{3.6}$$

$$E_{m_2}^{\pm 1} = 1 \pm \Psi^{-\frac{1}{2}}\frac{\partial}{\partial \xi_2} + \frac{1}{2}\Psi^{-1}\frac{\partial^2}{\partial \xi_2^2} + \cdots. \tag{3.7}$$

Substituting Equations (3.6) and (3.7) into Equation (3.2) and rearranging gives,

$$\begin{aligned}
&\frac{\partial}{\partial t}\Pi_{(\xi_1,\xi_2)} - \Psi^{\frac{1}{2}}\frac{d\phi_1}{dt}\frac{\partial}{\partial \xi_1}\Pi_{(\xi_1,\xi_2)} - \Psi^{\frac{1}{2}}\frac{d\phi_2}{dt}\frac{\partial}{\partial \xi_2}\Pi_{(\xi_1,\xi_2)} \\
&\quad = \hat{k_{f,1}}\left(-\Psi^{-\frac{3}{2}}\frac{\partial}{\partial \xi_1} + \frac{1}{2}\Psi^{-2}\frac{\partial^2}{\partial \xi_1^2}\right)\left(\Psi N_{R,1} - \Psi\phi_1 - \Psi^{\frac{1}{2}}\xi_1\right) \\
&\qquad \left(\Psi N_L - \Psi\phi_1 - \Psi^{\frac{1}{2}}\xi_1 - \phi_2 - \Psi^{\frac{1}{2}}\xi_2\right)\Pi_{(\xi_1,\xi_2)} \\
&\qquad + k_{r,1}\left(\Psi^{-\frac{1}{2}}\frac{\partial}{\partial \xi_1} + \frac{1}{2}\Psi^{-1}\frac{\partial^2}{\partial \xi_1^2}\right)\left(\Psi\phi_1 + \Psi^{\frac{1}{2}}\xi_1\right)\Pi_{(\xi_1,\xi_2)} \\
&\qquad + \hat{k_{f,2}}\left(-\Psi^{-\frac{3}{2}}\frac{\partial}{\partial \xi_2} + \frac{1}{2}\Psi^{-2}\frac{\partial^2}{\partial \xi_2^2}\right)\left(\Psi N_{R,2} - \Psi\phi_2 - \Psi^{\frac{1}{2}}\xi_2\right) \\
&\qquad \left(\Psi N_L - \Psi\phi_1 - \Psi^{\frac{1}{2}}\xi_1 - \phi_2 - \Psi^{\frac{1}{2}}\xi_2\right)\Pi_{(\xi_1,\xi_2)} \\
&\qquad + k_{r,2}\left(\Psi^{-\frac{1}{2}}\frac{\partial}{\partial \xi_2} + \frac{1}{2}\Psi^{-1}\frac{\partial^2}{\partial \xi_2^2}\right)\left(\Psi\phi_2 + \Psi^{\frac{1}{2}}\xi_2\right)\Pi_{(\xi_1,\xi_2)}.
\end{aligned} \tag{3.8}$$

Collecting terms of order $\Psi^{\frac{1}{2}}$ from Equation (3.8) and equating coefficients yields the differential equations

$$\begin{aligned}\frac{d\phi_1}{dt} &= \hat{k}_{f,1}(N_{R,1}-\phi_1)(N_L-\phi_1-\phi_2)-k_{r,1}\phi_1, \text{ and}\\ \frac{d\phi_2}{dt} &= \hat{k}_{f,2}(N_{R,2}-\phi_2)(N_L-\phi_1-\phi_2)-k_{r,2}\phi_2,\end{aligned} \tag{3.9}$$

for the quantities ϕ_1 and ϕ_2 corresponding to the concentrations of complexes of type 1 and type 2 respectively. These are the differential equations that would be obtained by considering the system in Figure 3.1 deterministically, neglecting fluctuations. Collecting terms of order Ψ^0 from Equation (3.8), the linear Fokker-Planck equation

$$\frac{\partial \Pi_{(\xi_1,\xi_2)}}{\partial t} = -\sum_{i,j} A_{ij}\frac{\partial}{\partial \xi_i}\xi_j \Pi_{(\xi_1,\xi_2)} + \frac{1}{2}\sum_{i,j} B_{ij}\frac{\partial^2 \Pi_{(\xi_1,\xi_2)}}{\partial \xi_i \partial \xi_j} \tag{3.10}$$

is obtained, where

$$\begin{aligned}A_{11} &= -\hat{k}_{f,1}(N_{R,1}+N_L-2\phi_1^*-\phi_2^*)-k_{r,1},\\ A_{12} &= -\hat{k}_{f,1}(N_{R,1}-\phi_1^*),\\ A_{21} &= -\hat{k}_{f,2}(N_{R,2}-\phi_2^*),\\ A_{22} &= -\hat{k}_{f,2}(N_{R,2}+N_L-2\phi_2^*-\phi_1^*)-k_{r,2},\\ B_{11} &= -\hat{k}_{f,1}(N_{R,1}+N_L-2\phi_1^*-\phi_2^*)+k_{r,1}\phi_1^*,\\ B_{12} &= B_{21} = 0, \text{ and}\\ B_{22} &= -\hat{k}_{f,2}(N_{R,2}+N_L-2\phi_2^*-\phi_1^*)+k_{r,2}\phi_2^*,\end{aligned}$$

and (ϕ_1^*, ϕ_2^*) is the steady state of the system of Equations (3.9). ODEs for the first and second moments of the fluctuations can be found by multiplying Equation (3.10) by ξ_i for $i \in \{1,2\}$ and integrating. In particular,

$$\begin{aligned}\frac{d\langle \xi_k\rangle}{dt} &= \sum_j A_{kj}\langle \xi_j\rangle, \text{ and}\\ \frac{d\langle \xi_k\xi_l\rangle}{dt} &= \sum_i A_{ki}\langle \xi_i\xi_l\rangle + \sum_j A_{lj}\langle \xi_k\xi_j\rangle + B_{kl},\end{aligned} \tag{3.11}$$

where $\langle\cdot\rangle = \mathbb{E}[\cdot]$ and $k, l \in \{1, 2\}$.

3.3 Model dynamics

In this section, the dynamics of the model for the system in Figure 3.1 are explored for differing numbers of molecules and rate constants in order to see how these parameters effect the competition for complex formation. The LNA is used here by solving the system of Equations (3.11) with initial condition $\langle\xi_1\rangle = \langle\xi_2\rangle = \langle\xi_1\xi_1\rangle = \langle\xi_1\xi_2\rangle = \langle\xi_2\xi_2\rangle = 0$. In Figure 3.3, the stochastic model is simulated by means of the Gillespie algorithm (Section 2.2.9 and Gillespie (1976b)), for different numbers of molecules and K_d values of the reactions, where $K_{d,i} = \frac{k_{r,i}}{k_{f,i}}$ for $i \in \{1, 2\}$. As well as these parameter values, the number of available receptors of type 2, $n_{R,2}$, is varied, to see the effect on the competition dynamics. In all plots in Figure 3.3, $n_L = 10^2$, $k_{r,1} = k_{r,2} = 10^{-3}$ s^{-1} and $n_{R,1} = 10^2$. The values of the other parameters being varied are given as text at the side of a row of plots in Figure 3.3, or in the grey box in each subplot. The value chosen for the dissociation rate constants $k_{r,1}$ and $k_{r,2}$ is the common rate at which VEGFR1 and VEGFR2 dissociate their shared ligand, VEGF-A (Weddell & Imoukhuede, 2014, 2017). In each subplot, the LNA for the fluctuations ξ_1 and ξ_2 is also plotted. In particular, plotted as a dotted line are the deterministic solutions to the system of Equations (3.9) and as a shaded area is the deterministic solution plus and minus two standard deviations of ξ_1 and ξ_2. Plotted in this way, the LNA well captures the variability of the corresponding single stochastic realisation for the process.

In the first plots of each row in Figure 3.3, it can be seen that the process favours the formation of M_1 over M_2. In this case, receptor type 1 out-competes receptor type 2 for ligand binding. In the second plot of each row, the number of molecules and rate constants governing the formation of M_1 and M_2 are identical and hence the time courses for the two complexes are at a similar level, and are sometimes overlapping. The LNAs for the fluctuations corresponding to M_1 and M_2 in the second subplots of each row are identical, which is expected since the system is symmetric when the rate constants and numbers of molecules are

identical. Finally in the third plots of each row, the receptors of type 2 are outcompeting the receptors of type 1, and hence the time courses for M_2 are above those of M_1. It can also be observed from Figure 3.3 that in each of the subplots the process appears to reach steady state by the end of the time course. Thus, one of the aims of Section 3.4 is to examine the effect of the rate constants and the number of molecules on the steady state distribution of the process.

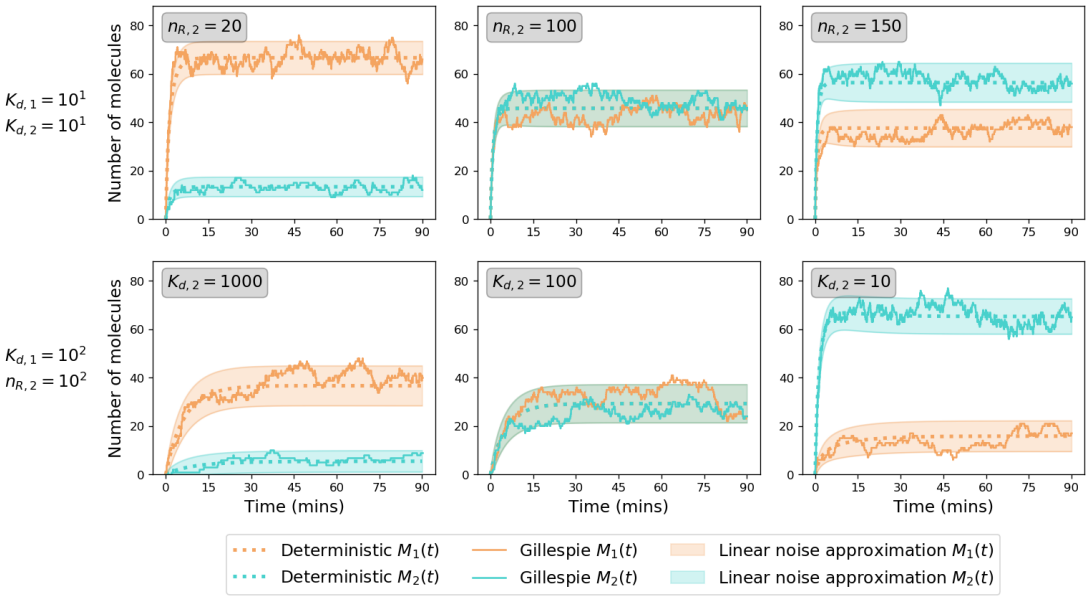

Figure 3.3: Gillespie simulation (solid lines), LNA (shaded areas) and deterministic solution (dotted lines) for the competition process showing the effect of varying rate constants and number of molecules on the time evolution of the complexes, $M_1(t)$ and $M_2(t)$. **Top row:** For equal K_d values, the number, $n_{R,2}$, of type 2 receptors is varied. **Bottom row:** For equal numbers of receptors, $K_{d,2}$ is varied.

3.4 Steady state distribution

In this section, the steady state distribution for the Markov chain defined in Section 3.1 is introduced. The steady state distribution for the model in Figure 3.1 is described in terms of the following probabilities (Pinsky & Karlin, 2010)

$$\pi_{(m_1,m_2)} = \lim_{t\to+\infty} \mathbb{P}((M_1(t), M_2(t)) = (m_1, m_2)), \quad (m_1, m_2) \in \mathcal{S}_{\mathcal{X}},$$

which can be stored in a row vector $\boldsymbol{\pi} = (\pi_{(m_1,m_2)},\ (m_1, m_2) \in \mathcal{S}_\mathcal{X})$ for any given order of states in $\mathcal{S}_\mathcal{X}$. These probabilities correspond to the number of complexes, of type 1 and type 2, respectively, found on the cell surface at late times and do not depend on the initial state of the Markov chain. They are known to satisfy (Allen, 2010; Latouche *et al.*, 1999) the system of equations

$$\begin{aligned} \boldsymbol{\pi} \cdot \boldsymbol{Q} &= \mathbf{0}^T_{\#\mathcal{S}_\mathcal{X}}, \\ \boldsymbol{\pi} \cdot \mathbf{1}_{\#\mathcal{S}_\mathcal{X}} &= 1, \end{aligned} \tag{3.12}$$

where $\boldsymbol{Q}$ is the infinitesimal generator of the CTMC $\mathcal{X}$ (Section 2.2.3), $\#\mathcal{S}_\mathcal{X}$ is the number of states in the state space, $\mathbf{0}_a$ is a column vector of zeros with length a, $\mathbf{1}_a$ is a column vector of ones with length a and superscript T denotes the transpose of a vector.

In the following sections, several methods of computation of the steady state distribution are considered and the scope and limitations of each method are explored.

3.4.1 Exact matrix-analytic approach (EMA)

In this section, an exact analytic method of computing the steady state distribution is outlined which uses an algorithmic matrix approach. This method will be referred to as the exact matrix-analytic approach (EMA), and provides an exact solution based on solving efficiently Equation (3.12). The infinitesimal generator, $\boldsymbol{Q}$, introduced in the previous section, encodes the infinitesimal transition rates, and it is possible to order the states in $\mathcal{S}_\mathcal{X}$ so that $\boldsymbol{Q}$ is tridiagonal by blocks. In particular, the state space can be organised into levels, as follows,

$$\mathcal{S}_\mathcal{X} \ = \ \bigcup_{k=0}^{N_2} L(k), \quad L(k) \ = \ \{(m_1, m_2) \in \mathcal{S}_\mathcal{X} : m_2 = k\}, \quad 0 \leq k \leq N_2 \ .$$

Thus, level $L(k) = \{(0,k),(1,k),...,(min(n_{R,1}, n_L - k), k)\}$ contains the states in which k type 2 complexes are present on the cell surface at any given time. If, when constructing the matrix $\boldsymbol{Q}$, one orders the states by levels with $L(0) \prec L(1) \prec ... \prec L(N_2)$, and states within each level $L(k)$ as $(0,k) \prec (1,k) \prec \cdots \prec$

$(min(n_{R,1}, n_L - k), k)$, then it is clear that only transitions within a level or to adjacent levels are allowed. This means that from any state $(m_1, m_2) \in L(m_2)$, the next event in the Markov chain can take the process to either another state in level $L(m_2)$, a state in level $L(m_2+1)$, or a state in level $L(m_2-1)$, by means of an association or dissociation reaction involving M_1, an association to form M_2 or a dissociation of M_2, respectively. Thus, since the process only moves up or down by a maximum of one level after each transition, the CTMC $\mathcal{X}$ is a level-dependent quasi-birth-and-death (LD-QBD) process (see Section 2.2.8 and Kulkarni (2016)), and the infinitesimal generator matrix $\boldsymbol{Q}$ is tridiagonal by blocks. $\boldsymbol{Q}$ is given by

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{0,0} & \boldsymbol{Q}_{0,1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \boldsymbol{Q}_{1,0} & \boldsymbol{Q}_{1,1} & \boldsymbol{Q}_{1,2} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}_{2,1} & \boldsymbol{Q}_{2,2} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{Q}_{N_2-1,N_2-1} & \boldsymbol{Q}_{N_2-1,N_2} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{Q}_{N_2,N_2-1} & \boldsymbol{Q}_{N_2,N_2} \end{pmatrix}.$$

Each level $L(k)$ contains $J(k) = \#L(k) = min(n_{R,1}, n_L - k) + 1$ states, so that each sub-matrix $\boldsymbol{Q}_{k,k'}$ has dimensions $J(k) \times J(k')$. Dimensions are omitted from the sub-matrices containing only zeros, $\mathbf{0}$, for ease of notation, however these matrices have dimensions corresponding to the number of states in each level, similarly to the matrices $\boldsymbol{Q}_{k,k'}$. Sub-matrices $\boldsymbol{Q}_{k,k'}$ are given as follows:

- For $1 \leq k \leq N_2$,

$$(\boldsymbol{Q}_{k,k-1})_{i,j} = \begin{cases} k_{r,2}k, & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

  for $0 \leq i \leq J(k)$ and $0 \leq j \leq J(k-1)$.

- For $0 \leq k \leq N_2 - 1$,

$$(\boldsymbol{Q}_{k,k+1})_{i,j} = \begin{cases} k_{f,2}(n_{R,2} - k)(n_L - i - k), & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

  for $0 \leq i \leq J(k)$ and $0 \leq j \leq J(k+1)$.

- For $0 \leq k \leq N_2$,

$$(\boldsymbol{Q}_{k,k})_{i,j} = \begin{cases} k_{f,1}(n_{R,1}-i)(n_L-i-k), & \text{if } j = i+1, \\ k_{r,1}i, & \text{if } j = i-1, \\ -(k_{f,1}(n_{R,1}-i)(n_L-i-k) + k_{r,1}i & \\ +k_{f,2}(n_{R,2}-k)(n_L-i-k) + k_{r,2}k), & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

  for $0 \leq i \leq J(k)$ and $0 \leq j \leq J(k)$.

With the infinitesimal generator in the tridiagonal by blocks form, one can solve the steady state matrix equations, Equations (3.12), with Algorithm 6, where $\boldsymbol{\pi}$ is comprised of row vectors $\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{N_2}$ which contain the ordered probabilities $\pi_{(m_1,m_2)}$ for the states in the corresponding level,

$$\boldsymbol{\pi}_k = \left(\pi_{(0,k)}, \pi_{(1,k)}, \ldots, \pi_{(min(n_{R,1},n_L-k),k)}\right),$$

$0 \leq k \leq N_2$. Algorithm 6 is an adapted version of the linear level reduction algorithm (Gaver *et al.*, 1984), and the steady state distribution computed via the EMA using this algorithm will be denoted $\pi^{EMA}_{(m_1,m_2)}$.
Once these probabilities are in hand, one can compute the mean number of complexes at steady state as

$$\mathbb{E}^{EMA}[M_1^*] = \sum_{m_1=0}^{N_1} m_1 \left(\sum_{m_2=0}^{min(n_{R,2},n_L-m_1)} \pi^{EMA}_{(m_1,m_2)} \right),$$
$$\mathbb{E}^{EMA}[M_2^*] = \sum_{m_2=0}^{N_2} m_2 \left(\sum_{m_1=0}^{min(n_{R,1},n_L-m_2)} \pi^{EMA}_{(m_1,m_2)} \right).$$

The advantage of using this method to compute the steady state distribution is that it is an exact method. There are however, considerable limitations to this method, whereby although Algorithm 6 is computationally efficient for a relatively small state space, as the number of molecules in the system, and hence the state space, increases, the computational time and memory also increase. It is also possible to use this EMA when considering a larger system in which

Algorithm 6 Computation of $(\pi^{EMA}_{(m_1,m_2)},\ (m_1,m_2) \in \mathcal{S}_{\mathcal{X}})$ (Gaver *et al.*, 1984; Gómez-Corral & López-García, 2018).

$\boldsymbol{H}_0 = \boldsymbol{Q}_{0,0}$.
for $k = 1, \ldots, N_2$ **do**:
    $\boldsymbol{H}_k = \boldsymbol{Q}_{k,k} - \boldsymbol{Q}_{k,k-1}\boldsymbol{H}_{k-1}^{-1}\boldsymbol{Q}_{k-1,k}$.
end for
Evaluate $\boldsymbol{\pi}^*_{N_2}$ by solving $\boldsymbol{\pi}^*_{N_2}\boldsymbol{H}_{N_2} = \mathbf{0}^T_{J(N_2)}$ with $\boldsymbol{\pi}^*_{N_2}\mathbf{1}_{J(N_2)} = 1$.
for $k = N_2 - 1, \ldots, 0$ **do**:
    $\boldsymbol{\pi}^*_k = -\boldsymbol{\pi}^*_{k+1}\boldsymbol{Q}_{k+1,k}\boldsymbol{H}_k^{-1}$.
end for
for $k = 0, \ldots, N_2$ **do**:
    $\boldsymbol{\pi}_k = \frac{1}{\sum_{i=0}^{N_2}\boldsymbol{\pi}^*_i\mathbf{1}_{J(i)}}\boldsymbol{\pi}^*_k$.
end for
return $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \ldots, \boldsymbol{\pi}_{N_2})$. ▷ EMA steady state distribution

three receptor types (as opposed to the two considered here) are competing for a common ligand. In this scenario, the matrix forms comprising the infinitesimal generator become more complex and, since the state space is again larger, the computational expense increases. For any situation with more than three receptor types it would be challenging to write down the form of the infinitesimal generator and the computational expense may become infeasible. Given these drawbacks to the EMA, in the following sections, approximate methods are proposed to compute the steady state distribution which are considerably faster than the EMA and can also be extended easily to a system with N receptor types competing for a common ligand, where $N \geq 2$.

3.4.2 No competition approximation (NCA)

In this section, an approximate method of computing the steady state distribution is introduced which can be used when $n_L \to +\infty$, *i.e.* when there is no competition between the receptor types for the common ligand because the lig-

and is in huge excess. In the limit $n_L \to +\infty$, given that the time dynamics of the complexes are linked only by the number of ligands present, the steady state probabilities for the two-dimensional process will be precisely equal to the product of the steady state probabilities for the one-dimensional processes modelling $M_1(t)$ and $M_2(t)$. This product, in the limit $n_L \to +\infty$, comprises the no competition approximation (NCA). In particular, two independent Markov chains

$$\mathcal{X}_1 = \{M_1(t) : \ t \geq 0\}, \quad \text{and} \quad \mathcal{X}_2 = \{M_2(t) : \ t \geq 0\},$$

can be defined for the complexes M_1 and M_2, respectively, with state spaces

$$\mathcal{S}_{\mathcal{X}_1} = \{m_1 \in \mathbb{N}_0 = \mathbb{N} \cup \{0\} : \ m_1 \leq N_1\}, \quad \text{and} \quad \mathcal{S}_{\mathcal{X}_2} = \{m_2 \in \mathbb{N}_0 : \ m_2 \leq N_2\}.$$

Since the number of complexes in each CTMC can only increase or decrease by one unit in every transition (by means of complex formation or dissociation), each CTMC $\mathcal{X}_j$, for $j \in \{1, 2\}$, is a birth-and-death process (see Section 2.2.7) as depicted in Figure 3.4, with $\lambda_n = k_{f,j}(n_{R,j} - n)(n_L - n)$ and $\mu_n = k_{r,j}n$.

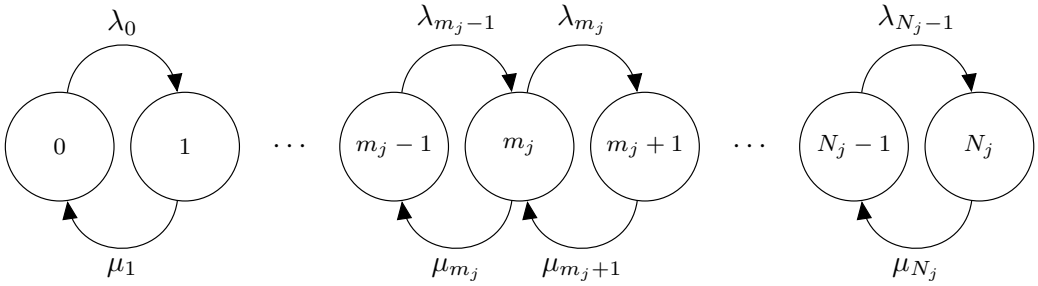

Figure 3.4: Diagram of the birth-and-death processes $\mathcal{X}_j$, $j \in \{1, 2\}$.

In the limit $n_L \to +\infty$, ligand depletion can be neglected, and then the binding rates become $\lambda_n = k_{f,j}n_L(n_{R,j} - n)$. For these rates, Equation (3.12) can be easily solved (Allen, 2010), giving

$$\pi_{m_1} = \lim_{t \to +\infty} \mathbb{P}(M_1(t) = m_1) \ = \ \binom{n_{R,1}}{m_1} \left(\frac{k_{f,1}n_L}{k_{r,1} + k_{f,1}n_L}\right)^{m_1} \left(\frac{k_{r,1}}{k_{r,1} + k_{f,1}n_L}\right)^{n_{R,1}-m_1}, \tag{3.13}$$

$$\pi_{m_2} = \lim_{t\to+\infty} \mathbb{P}(M_2(t) = m_2) \;=\; \binom{n_{R,2}}{m_2} \left(\frac{k_{f,2}n_L}{k_{r,2}+k_{f,2}n_L}\right)^{m_2} \left(\frac{k_{r,2}}{k_{r,2}+k_{f,2}n_L}\right)^{n_{R,2}-m_2}, \tag{3.14}$$

for $0 \leq m_1 \leq N_1$ and $0 \leq m_2 \leq N_2$, which can be identified as binomial distributions. For these one-dimensional birth-and-death processes, $\mathcal{X}_1$ and $\mathcal{X}_2$, one can determine the expected values of complexes in steady state as

$$\begin{aligned} \mathbb{E}^{NCA}[M_1^*] = \sum_{m_1=0}^{N_1} m_1\, \pi_{m_1} &= \frac{k_{f,1}n_{R,1}n_L}{k_{r,1}+k_{f,1}n_L}, \quad \text{and} \\ \mathbb{E}^{NCA}[M_2^*] = \sum_{m_2=0}^{N_2} m_2\, \pi_{m_2} &= \frac{k_{f,2}n_{R,2}n_L}{k_{r,2}+k_{f,2}n_L}. \end{aligned} \tag{3.15}$$

In the limit $n_L \to +\infty$, the steady state distribution of the process $\mathcal{X}$ is given by

$$\pi^{NCA}_{(m_1,m_2)} = \pi_{m_1} \times \pi_{m_2}, \quad (m_1,m_2) \in \mathcal{S}_{\mathcal{X}}. \tag{3.16}$$

Note that when $n_L \to +\infty$, one gets $N_1 = n_{R,1}$, $N_2 = n_{R,2}$ and $\mathcal{S}_{\mathcal{X}} = \{(m_1,m_2) \in \mathbb{N}_0^2 :\ 0 \leq m_1 \leq N_1,\ 0 \leq m_2 \leq N_2\} = \mathcal{S}_{\mathcal{X}_1} \times \mathcal{S}_{\mathcal{X}_2}$, so that Equations (3.13)-(3.16) are consistent.

Encouragingly, one can compare the expression found in Equation (3.15) for the expected number of complexes in steady state with the deterministic analogue and find that the resulting expressions have the same meaning in the large n_L limit. In particular, since the NCA assumes that $n_L \to \infty$, and hence there is no competition between the receptor types for the ligand, one can study the deterministic model of a single receptor type (R) binding the ligand. This model can be written as

$$R + L \underset{k_r}{\overset{k_f}{\rightleftharpoons}} M.$$

The deterministic system of ODEs for this model, assuming mass action kinetics can be written as

$$\begin{aligned} \frac{dR}{dt} &= -k_f RL + k_r M, \\ \frac{dL}{dt} &= -k_f RL + k_r M, \end{aligned} \tag{3.17}$$

$$\frac{dM}{dt} = k_f RL - k_r M.$$

The ODEs are valid for any time t, with $t \geq 0$, but time has been omitted in the species notation for ease of notation, where for example $R = R(t)$ for all $t \geq 0$. At steady state, one can set the right hand side of any of the three ODEs to 0 to obtain

$$k_f R^* L^* = k_r M^*, \tag{3.18}$$

where x^* denotes the steady state number of molecules of species x. Rearranging Equation (3.18) yields

$$R^* = \frac{K_d M^*}{L^*}, \tag{3.19}$$

where $K_d = \frac{k_r}{k_f}$. In this receptor ligand binding model, there is no synthesis or degradation of any species and hence one can write the total number of molecules of receptor, R^T, as

$$R^T = R + M, \tag{3.20}$$

for any time t. At steady state therefore, substituting the expression for R^* from Equation (3.19) into Equation (3.20) gives

$$R^T = \frac{K_d M^*}{L^*} + M^*,$$

and rearranging results in

$$\frac{M^*}{R^T} = \frac{L^*}{K_d + L^*}, \tag{3.21}$$

which is an expression in terms of L^* and K_d for the fraction of ligand-bound receptors at steady state. One can obtain a similar fraction in the stochastic model by rearranging the expressions in Equation (3.15). For example, considering the

expression for $\mathbb{E}^{NCA}[M_1^*]$, it can be seen that

$$\frac{\mathbb{E}^{NCA}[M_1^*]}{n_{R,1}} = \frac{n_L}{K_{d,1} + n_L}, \tag{3.22}$$

which is exactly equivalent in the stochastic sense to Equation (3.21) from the deterministic treatment when $n_L \to \infty$ and hence $L^* = n_L$. Clearly, in the deterministic model when an initial number of ligands, n_L, is considered which is relatively small, and so $L^* < n_L$,

$$\frac{M^*}{R^T} = \frac{L^*}{K_d + L^*} < \frac{n_L}{K_{d,1} + n_L} = \frac{\mathbb{E}^{NCA}[M_1^*]}{n_{R,1}}. \tag{3.23}$$

Given that R^T and $n_{R,1}$ have exactly the same meaning (just different notation for the deterministic and stochastic models), Equation (3.23) implies that in the steady state, $M^* < \mathbb{E}^{NCA}[M_1^*]$. From Equation (3.21) (or Equation (3.22) in the stochastic sense), one can derive the well-known (Hulme & Trevethick, 2010) number of ligand molecules required for the number of ligand-bound receptors to be half maximal. In this case, when $\frac{M^*}{R^T} = \frac{1}{2}$ it can be easily shown that $L^* = K_d$, *i.e.* the number of ligands should be equal to the dissociation constant. In the stochastic model this would imply that n_L should be equal to K_d for half of the total number of receptors to be bound by ligand in steady state.

In order to see how the approximate distribution computed via the NCA compares with the exact distribution computed via the EMA, the distributions are plotted in Figure 3.5, in columns 1 and 2, respectively, where each state is coloured according to its steady state probability. The third column of the figure shows the absolute difference $|\pi^{EMA}_{(m_1,m_2)} - \pi^{NCA}_{(m_1,m_2)}|$, between the distributions. Each row in the figure uses a different number of ligands, n_L, as stated in the text on the left hand side of the row, and the other parameter values and numbers of molecules used are stated in the figure caption. In the second column of the figure, the steady state number of complexes M^* is plotted as a green star, found by numerically solving the ODE system (3.17) to steady state. It can be seen that, as expected (by analysis of Equation (3.23)), the green stars underestimate the expected numbers of complexes found under the stochastic model, since the number of ligands used for each row of the figure is not infinite and hence $L^* < n_L$. For example, for

the largest number of ligands, $n_L = 2500$, in the top row of the figure, from Equation (3.21) it is found that $M^* = 70.8$, whereas from Equation (3.22) one finds $\mathbb{E}^{NCA}[M_1^*] = 71.4$. For the association and dissociation rate constants used in Figure 3.5, one can compute the K_d values as $K_{d,i} = \frac{k_{r,i}}{k_{f,i}} = \frac{10^{-3}}{10^{-6}} = 1000$, for $i \in \{1, 2\}$. By the analysis of Equation (3.22), when $n_L = K_{d,i} = 1000$ for $i \in \{1, 2\}$, the expected number of complexes of each type should be half maximal under the NCA. Indeed, from the second row of subplots in Figure 3.5 where $n_L = 1000$, it can be seen that $\mathbb{E}^{NCA}[M_1^*] = \mathbb{E}^{NCA}[M_2^*] = 50$ which is expected since $n_{R,1} = n_{R,2} = 10^2$ and $\frac{50}{100} = \frac{1}{2}$. In general, Figure 3.5 also provides some justification for the choice to study a stochastic model instead of a deterministic. It can be seen that the steady state distributions in the first two columns are rather spread out from their centres and this information is lacking from a deterministic model. A general rule of thumb (as mentioned in Section 3.2) is that if the average size of a population is n, then the size of the fluctuations around n is $\sqrt{n}$. One can then compute the percentage deviations from n as $\frac{\sqrt{n}}{n} \times 100$ and hence for large values such as $n = 10^4$ the percentage deviations from n will be only 1%, whereas for $n = 10^2$ the percentage deviations will be 10%. For values as large as 10% it is certainly worthwhile to use a stochastic model in order to see the range of feasible model outputs.

It can also be seen from Figure 3.5 that when n_L is large, and hence the competition is low, the NCA well approximates the EMA distribution where the absolute difference between the distributions is very small for all states in the state space. The expected number of complexes of each type is also reasonably well captured here. As the number of ligands decreases however, and in particular when $n_L = 10^2 < n_{R,1} + n_{R,2}$, the distributions become less similar, where the NCA overestimates the probabilities of states with larger values of m_1 and m_2 in the state space. This is because, when using the NCA, the assumption is that the Markov chains $\mathcal{X}_1$ and $\mathcal{X}_2$ are independent, which is clearly not the case when $n_L = 10^2$.

Given that the NCA works well only in scenarios when $n_L \gg n_{R,1} + n_{R,2}$, an alternative method is required for scenarios in which n_L is comparable to $n_{R,1} + n_{R,2}$. Thus, in the next section, an extension to the NCA method is proposed which can be used in such moderate competition scenarios.

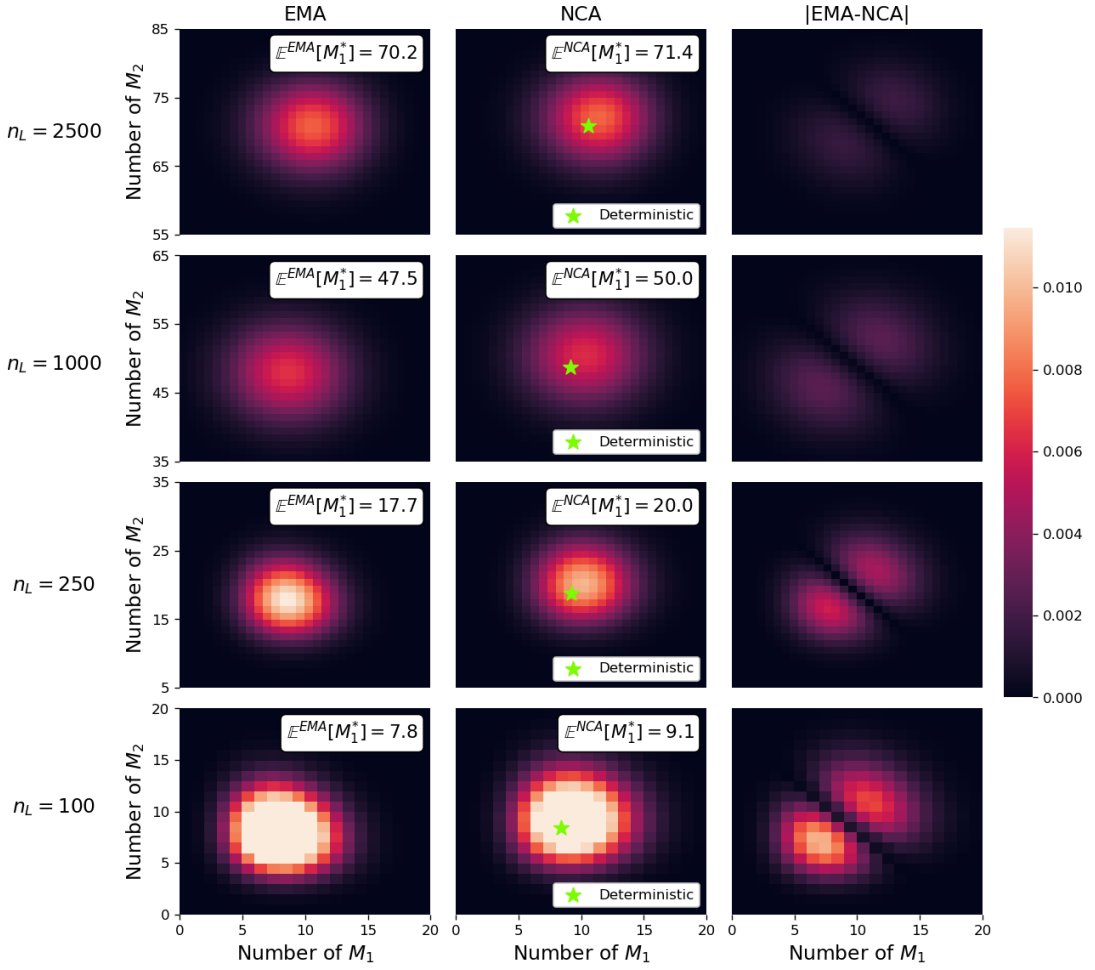

Figure 3.5: Comparison between the EMA and NCA steady state distributions, where the colour of a pixel in the first two columns indicates the steady state probability of that state, given by the colour bar. In the third column, the colour of a pixel indicates the absolute difference between the EMA and NCA derived steady state probability. For all distribution subplots, $n_{R,1} = n_{R,2} = 10^2$, $k_{r,1} = k_{r,2} = 10^{-3}$ s^{-1} and $k_{f,1} = k_{f,2} = 10^{-6}$ s^{-1}. Since the distributions are symmetric, $\mathbb{E}^{EMA}[M_1^*] = \mathbb{E}^{EMA}[M_2^*]$ and $\mathbb{E}^{NCA}[M_1^*] = \mathbb{E}^{NCA}[M_2^*]$. In the second column, a green star represents the value M^* as found by numerically solving the ODE system (3.17) to steady state.

3.4.3 Moderate competition approximation (MCA)

In this section, the moderate competition approximation (MCA) is introduced, whereby Equation (3.16) is again ultimately used to compute the two-dimensional

steady state distribution, but where the one-dimensional steady state probabilities (3.13) and (3.14) are computed using an *effective* number of ligands n_L^*. The idea is that, in each independent Markov chain, the competition from the other receptor type is accounted for by considering $n_L^* < n_L$, so that some of the available ligands are "used up" by the competing process. Using this smaller number of ligands in Equations (3.13) and (3.14) will yield smaller probabilities for the states with larger numbers of complexes, and hence when the product is taken in Equation (3.16) the probabilities will also be smaller for the states with larger numbers of m_1 and m_2. It is clear that the extent to which the distribution computed using n_L^* will differ from that using n_L depends on how small n_L^* is in comparison to n_L. Here, Algorithm 7 is proposed which uses an iterative scheme to reduce n_L to an appropriate value.

In Algorithm 7, the total number of ligands available in each independent birth-and-death process $\mathfrak{X}_1$ and $\mathfrak{X}_2$ is iteratively reduced, by subtracting from n_L the expected number of each complex type in steady state computed from the previous iteration. In each iteration i, the mean number of complexes in steady state under the NCA approximation, $\mathbb{E}^{NCA,(i)}[M_1^*]$ and $\mathbb{E}^{NCA,(i)}[M_2^*]$, are computed, and they are used to compute an effective number of free ligands available $n_L^{(i+1)}$, at iteration $i+1$. The iterative scheme stops once these mean values are close enough for two consecutive iterations, as determined by a threshold value ε. The final number of ligands considered, $n_L^* = n_L^{(i)}$, in the last iteration, can be seen as the *effective* number of ligands that reflects the competition for shared resources.

The parameter α ($\alpha \in [0,1]$) is a tuning parameter required in situations where $n_L < n_{R,1} + n_{R,2}$, to modulate the convergence speed of the algorithm. More specifically, from line 5 of Algorithm 7 and where $i = 0$, it can be seen that $n_L - \mathbb{E}^{NCA,(0)}[M_1^*] - \mathbb{E}^{NCA,(0)}[M_2^*]$ will be negative or equal to zero (and hence α required) when

$$n_L - \frac{k_{f,1} n_{R,1} n_L}{k_{r,1} + k_{f,1} n_L} - \frac{k_{f,2} n_{R,2} n_L}{k_{r,2} + k_{f,2} n_L} \leq 0,$$

from Equations (3.15). Rearranging this expression and dividing by n_L gives that

Algorithm 7 Iterative approximation for the steady state distribution, $(\pi^{MCA}_{(m_1,m_2)},\ (m_1,m_2) \in \mathcal{S}_{\mathcal{X}})$.

1: $i = 0$, $n_L^{(i)} = n_L$.

2: Compute $\mathbb{E}^{NCA,(i)}[M_1^*]$ and $\mathbb{E}^{NCA,(i)}[M_2^*]$ from Equation (3.15), using $n_L^{(i)}$ as the number of ligands in the system, instead of n_L.

3: **while** $\mid \mathbb{E}^{NCA,(i)}[M_1^*] - \mathbb{E}^{NCA,(i-1)}[M_1^*] \mid > \varepsilon$ or $\mid \mathbb{E}^{NCA,(i)}[M_2^*] - \mathbb{E}^{NCA,(i-1)}[M_2^*] \mid > \varepsilon$ or $i < 1$ **do**:

4: $\quad i = i + 1$.

5: $\quad n_L^{(i)} = (n_L - \mathbb{E}^{NCA,(i-1)}[M_1^*] - \mathbb{E}^{NCA,(i-1)}[M_2^*])\alpha + n_L^{(i-1)}(1-\alpha)$.

6: $\quad$ Compute $\mathbb{E}^{NCA,(i)}[M_1^*]$ and $\mathbb{E}^{NCA,(i)}[M_2^*]$ from Equation (3.15), using $n_L^{(i)}$ instead of n_L.

7: **end while**

8: Compute $\pi^{NCA,(i)}_{m_1}$, for $0 \leq m_1 \leq N_1$, and $\pi^{NCA,(i)}_{m_2}$, for $0 \leq m_2 \leq N_2$, from Equations (3.13)-(3.14) using $n_L^{(i)}$ instead of n_L.

9: Compute $\pi^{MCA}_{(m_1,m_2)} = \pi^{NCA,(i)}_{m_1} \times \pi^{NCA,(i)}_{m_2}$, for all $(m_1,m_2) \in \mathcal{S}_{\mathcal{X}}$.

10: **return** $\pi^{MCA}_{(m_1,m_2)} = \frac{\pi^{MCA}_{(m_1,m_2)}}{\sum_{(m_1',m_2') \in \mathcal{S}_{\mathcal{X}}} \pi^{MCA}_{(m_1',m_2')}}$, for all $(m_1,m_2) \in \mathcal{S}_{\mathcal{X}}$.

11: **return** $\mathbb{E}^{MCA}[M_1^*] = \sum_{m_1=0}^{N_1} m_1 \left(\sum_{m_2=0}^{min(n_{R,2}, n_L - m_1)} \pi^{MCA}_{(m_1,m_2)} \right)$.

12: **return** $\mathbb{E}^{MCA}[M_2^*] = \sum_{m_2=0}^{N_2} m_2 \left(\sum_{m_1=0}^{min(n_{R,1}, n_L - m_2)} \pi^{MCA}_{(m_1,m_2)} \right)$.

the parameter α will be required when

$$1 \leq \frac{n_{R,1}}{K_{d,1} + n_L} + \frac{n_{R,2}}{K_{d,2} + n_L}.$$

One can set $\alpha = 1$ when

$$1 > \frac{n_{R,1}}{K_{d,1} + n_L} + \frac{n_{R,2}}{K_{d,2} + n_L},$$

since in this case it is impossible for $n_L - \mathbb{E}^{NCA,(i-1)}[M_1^*] - \mathbb{E}^{NCA,(i-1)}[M_2^*]$ to be negative at any iteration. Otherwise, the parameter α is chosen between 0 and 1 such that the effective number of ligands at every iteration in the algorithm is positive. Numerical exploration indicates that the the choice of α does not affect the resulting steady state probabilities $\pi^{MCA}_{(m_1,m_2)}$. When $\alpha = 1$, the expected number of complexes and ligands in steady state exhibit damped oscillations as they converge. An example of this can be found in the top left plot of Figure 3.6. When $\alpha < 1$, for values of α for which the algorithm converges (smaller values of α), the expected number of complexes and ligands in steady state either exhibit damped oscillations as they converge (for larger values of α, see top right plot of Figure 3.6), or they decrease monotonically in convergence (for smaller values of α). Examples of monotonic convergence can be seen in the bottom row of Figure 3.6. For values of α for which there is monotonic convergence, the smaller the value of α, the more iterations the algorithm requires to converge. This effect can be seen when comparing the left hand plot of the bottom row of Figure 3.6 in which $\alpha = 0.3$, with the right hand plot of the same figure where $\alpha = 0.1$. In Figure 3.6, the mean number of free ligands in steady state, as predicted by the EMA, is given by $\mathbb{E}^{EMA}[L^*] = n_L - \mathbb{E}^{EMA}[M_1^*] - \mathbb{E}^{EMA}[M_2^*]$. Interestingly, the values of $n_L^{(i)}$ tend, in the limit $i \to +\infty$, to this mean number of free ligands in steady state.

In Algorithm 7, a natural choice is for α to vary between 0 and 1, and this is implemented in a linear fashion whereby α multiplies the first summand on the right hand side of the expression on line 5 and $(1 - \alpha)$ multiplies the second summand. Other non-linear choices have also been explored such as quadratic, cubic and quartic methods, whereby the first summand would be multiplied by α^n and the second by $(1 - \alpha^n)$ for $n \in \{2, 3, 4\}$. It was found that, in general, the larger the value of n, the greater the value of α could be in order for the algorithm still to converge. The trade-off for this widened range of feasible α values however, was that the algorithm took many more iterations to converge. For all values of α, and all linear and non-linear methods used, the accuracy of the algorithm was very similar. The accuracy was determined by computing

$$accuracy = |\mathbb{E}^{MCA}[M_1^*] - \mathbb{E}^{EMA}[M_1^*]| + |\mathbb{E}^{MCA}[M_2^*] - \mathbb{E}^{EMA}[M_2^*]|. \quad (3.24)$$

An example of this trade-off behaviour for α can be seen in Figure 3.7. Figure 3.7 uses the same parameter values and numbers of molecules as Figure 3.6 and each subplot shows a different linear or non-linear method for α as given in the subplot title. Each subplot shows a scatter plot with values of α between 0.1 and 1 on the x-axis and the number of iterations required for the algorithm to converge on the y-axis. The colour of a marker represents how accurate the algorithm was according to the definition given by Equation (3.24), where larger values in the colour bar therefore represent lesser accuracy. Finally, in each subplot there are two marker types, one for each of two ligand values, $n_L = 40$ and $n_L = 50$. In the case where $n_L = 50$, the parameter α is not required and hence it is seen that the algorithm converges in each subplot for all α values. As expected, the outcome of the MCA is more accurate for a larger number of ligands, however within each ligand number case, there is no noticeable difference in the accuracy depending on the value of α. When $n_L = 40$, for some larger values of α, the algorithm does not converge (or the values n_L^* still become negative at some iterations) and it can be seen from Figure 3.7 that this non-convergence happens more so for smaller values of n in α^n, *i.e.* the more non-linear the method, the larger α can be for convergence. Within an individual subplot one can note that as α increases the number of required iterations decreases up to a certain value of α at which the number of required iterations begins to rise again. It is also the case that the more non-linear the method chosen, the more iterations of the algorithm are required for convergence in general. Therefore, in order to keep the number of iterations to a minimum, the linear method was chosen in Algorithm 7. Since the algorithm using the linear method converges for only smaller values of α, it is recommended that a small value such as $\alpha = 0.05$ is used. This choice seems appropriate since, even for a rather high competition scenario whereby the parameter values are $k_{r,i} = 10^{-3}$, $K_{d,i} = 10^{1}$, $n_{R,i} = 100$ and $n_L = 10$ for $i \in \{1, 2\}$, the algorithm still converges with good accuracy under $\alpha = 0.05$. Upon non-convergence of the algorithm with the linear α method however, one could easily modify the algorithm to a non-linear method and obtain the MCA steady state distribution.

In Section 3.4.4 a thorough comparison between the EMA and MCA methods to compute the steady state distribution is carried out in order to validate the goodness of the approximate method and to find areas of the parameter space

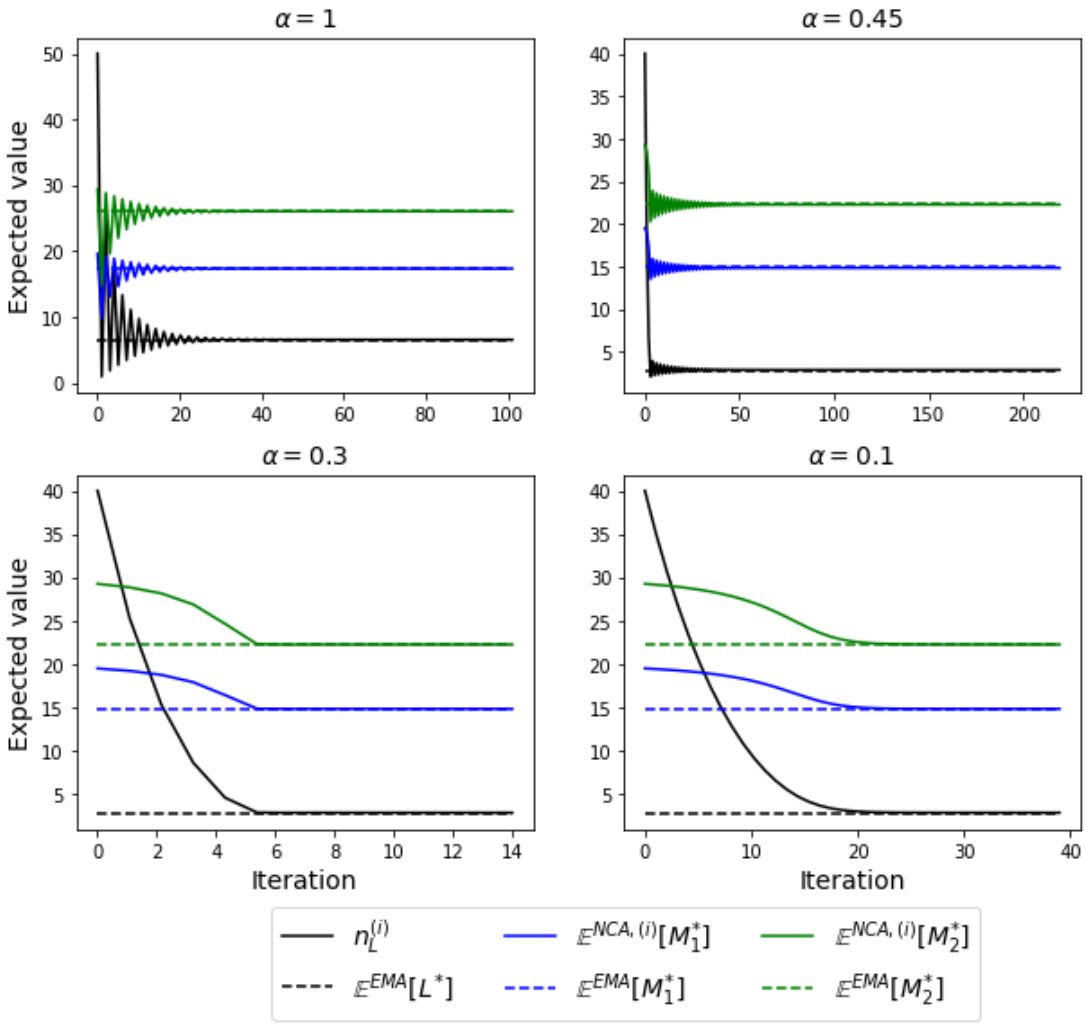

Figure 3.6: Examples of how Algorithm 7 converges when $n_L \geq n_{R,1} + n_{R,2}$ (**top left subplot**, $n_L = 50$) and when $n_L < n_{R,1} + n_{R,2}$ (**all other subplots**, $n_L = 40$). Iterative mean values $\mathbb{E}^{NCA,(i)}[M_1^*]$, $\mathbb{E}^{NCA,(i)}[M_2^*]$ and $n_L^{(i)}$ converge to the exact values $\mathbb{E}^{EMA}[M_1^*]$, $\mathbb{E}^{EMA}[M_2^*]$ and $\mathbb{E}^{EMA}[L^*]$. In all subplots, $\varepsilon = 10^{-5}$, $n_{R,1} = 20$, $n_{R,2} = 30$, $k_{r,1} = k_{r,2} = k_{f,1} = k_{f,2} = 1$ s^{-1}.

for which the MCA is particularly reliable. As a starting point, in Figure 3.8 the EMA is compared with both the NCA and MCA using the same parameter values as in Figure 3.5. It can be seen, from the third and fifth columns of the figure that the MCA performs much better than the NCA for the smaller values of n_L, where the MCA only decreases in accuracy very slightly as n_L decreases.

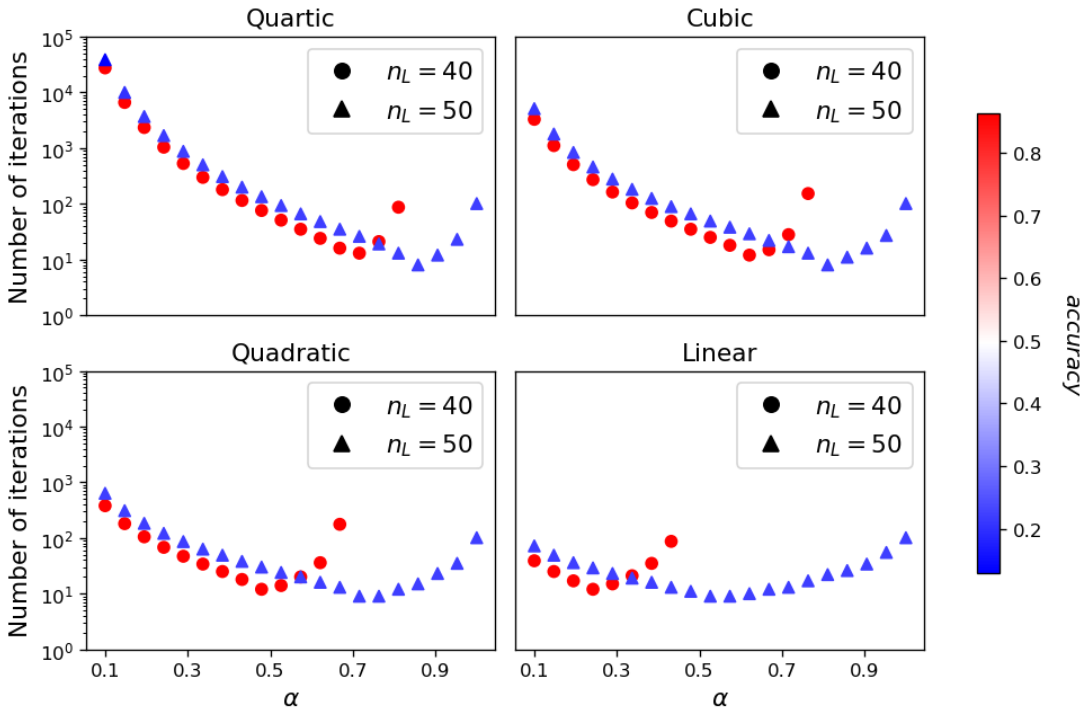

Figure 3.7: Scatter plots of the number of iterations required by, and accuracy of, Algorithm 7 for different linear and non-linear methods and varying values of $\alpha \in [0.1, 1]$. The accuracy of the algorithm is represented by the colour of a point and is computed via Equation (3.24). In all subplots, $\varepsilon = 10^{-5}$, $n_{R,1} = 20$, $n_{R,2} = 30$, $k_{r,1} = k_{r,2} = k_{f,1} = k_{f,2} = 1$ s^{-1} and two values of n_L are used (given in the figure legend). The parameter values and numbers of molecules used are the same as those in Figure 3.6.

It is also notable that for all four values of n_L, the MCA is able to predict to 1 decimal place, the expected number of type 1 and 2 complexes in steady state. In the following section therefore, as well as comparing the whole distributions as computed via the EMA and MCA, the expected numbers of complexes of each type in steady state will also be compared.

3.4.4 Numerical validation

In this section, a thorough numerical comparison is carried out between the EMA and MCA steady state distributions. Given that the NCA is a first step in proposing the MCA approximation, and that the MCA is expected to always perform

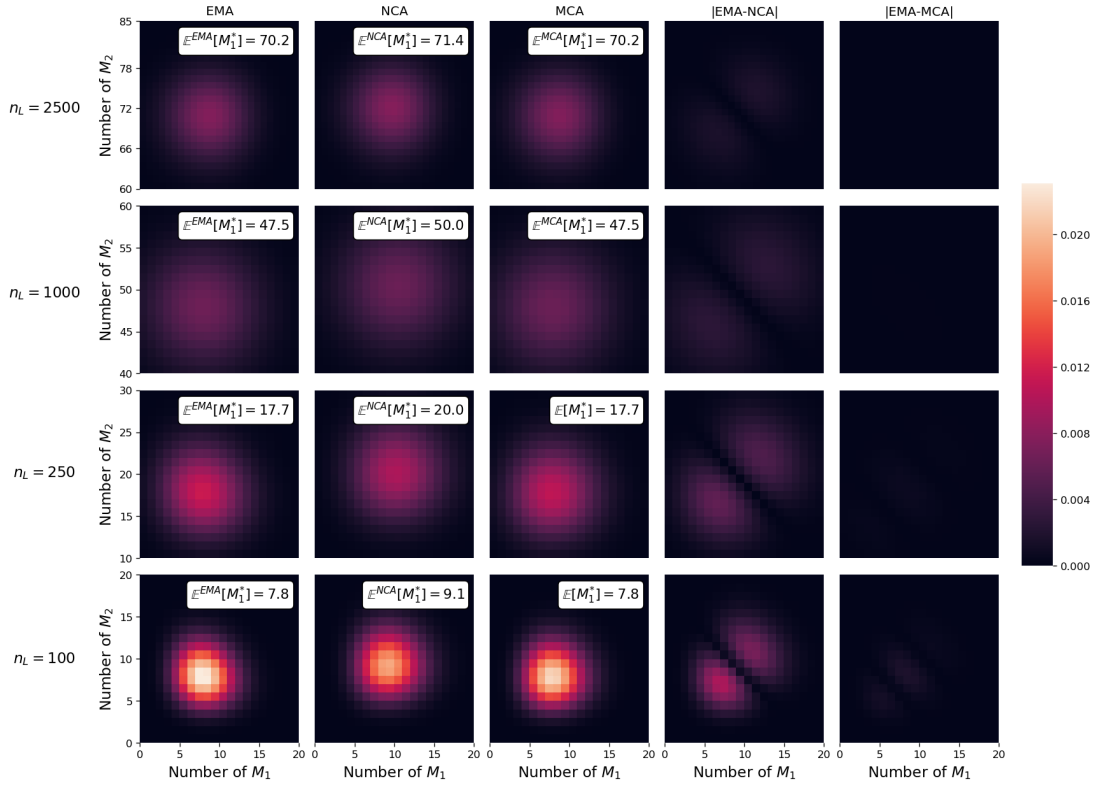

Figure 3.8: Comparison between the EMA, NCA and MCA steady state distributions, where the colour of a pixel in the first three columns indicates the steady state probability of that state, given by the colour bar. In the fourth and fifth columns, the colour of a pixel indicates the absolute difference between the EMA and NCA, and the EMA and MCA, derived steady state probabilities, respectively. For all distribution subplots, $n_{R,1} = n_{R,2} = 10^2$, $k_{r,1} = k_{r,2} = 10^{-3}$ s^{-1} and $k_{f,1} = k_{f,2} = 10^{-6}$ s^{-1}. Since the distributions are symmetric, $\mathbb{E}^{EMA}[M_1^*] = \mathbb{E}^{EMA}[M_2^*]$, $\mathbb{E}^{NCA}[M_1^*] = \mathbb{E}^{NCA}[M_2^*]$ and $\mathbb{E}^{MCA}[M_1^*] = \mathbb{E}^{MCA}[M_2^*]$.

better than the NCA, the goodness of the NCA compared to the EMA is not considered here. It is important to explore different areas of the parameter space which are biologically feasible, in order to see where the approximation performs best and is a viable alternative to the EMA. Different metrics are considered so that the accuracy of the MCA can be quantified when considering both the whole distribution and the expected number of complexes of each type in steady state.

In order to reduce the complexity of the numerical exploration, only the K_d value for each reaction in Figure 3.1, is considered, where $K_{d,i} = k_{r,i}/k_{f,i}$ for

$i \in \{1, 2\}$, and not the individual forward and backward rate constants, by fixing $k_{r,1} = k_{r,2} = 10^{-3}$ s^{-1} (Weddell & Imoukhuede, 2014, 2017) as in Figure 3.3. The K_d values and numbers of each receptor type are varied within the ranges given in Table 3.1, inspired from López-García *et al.* (2016) (Table 3), and which correspond to VEGFR1, VEGFR2 and VEGF-A. Finally, the number of ligands, n_L, is varied to consider different competition regimes between the receptors, where if $n_L \approx n_{R,1} + n_{R,2}$ the competition is reasonably high, whereas if $n_L \gg n_{R,1} + n_{R,2}$ then the competition is low.

Parameter	Range	Unit
$K_{d,1}, K_{d,2}$	$[1 \times 10^1 - 1 \times 10^4]$	molecules
$n_{R,1}, n_{R,2}$	$[2 \times 10^1 - 2 \times 10^2]$	molecules

Table 3.1: Ranges for the parameter values and numbers of molecules in the receptor-ligand competition model.

In order to compare the EMA with the MCA steady state distribution, the *Hellinger distance*

$$H(\{\pi^{EMA}_{(m_1,m_2)}\}_{(m_1,m_2)\in\mathcal{S}_\mathcal{X}}, \{\pi^{MCA}_{(m_1,m_2)}\}_{(m_1,m_2)\in\mathcal{S}_\mathcal{X}}) = \frac{1}{\sqrt{2}}\sqrt{\sum_{(m_1,m_2)\in\mathcal{S}_\mathcal{X}} \left(\sqrt{\pi^{EMA}_{(m_1,m_2)}} - \sqrt{\pi^{MCA}_{(m_1,m_2)}}\right)^2},$$

is used, which is a way to measure the similarity between two discrete probability distributions. The Hellinger distance (HD) can take values between 0 and 1, where lower values indicate that the probability distributions are more similar. Using the HD, one can identify regions of the parameter space for which the MCA performs well (HD closer to 0) and regions where the MCA is not so accurate (HD closer to 1). Thus in Figure 3.9, the HD is plotted as the colour on the heatmap, for each pair of parameters, where the parameters are varied within the ranges given in Table 3.1 and n_L is varied between 10^2 and 10^3 molecules. The default values of the parameters when they are not being varied (*i.e.* do not appear on either axis of the subplot) are $n_{R,1} = 10^2$, $n_{R,2} = 10^2$, $K_{d,1} = 10^3$, $K_{d,2} = 10^3$ and $n_L = 250$, so as to represent moderate competition.

It can be seen that the HD is lower than 0.16 for all of the scenarios considered, even though in some of them, the number of ligands is significantly smaller than the total number of receptors (*e.g.* $n_L = 250 < 400 = 200 + 200 = n_{R,1} + n_{R,2}$). The number of molecules in the system does not, by itself, explain the goodness of the MCA; the dissociation constants $K_{d,1}$ and $K_{d,2}$ need also to be taken into account. Settings where both $K_{d,1}$ and $K_{d,2}$ are large (indicating the lowest affinity for ligand) correspond to low competition, since a small number of complexes of each type is formed, and then the baseline number of ligands considered, $n_L = 250$, is sufficient for the competition to be negligible. On the other hand, scenarios where for example, $K_{d,1}$ is small, can still lead to low competition if n_L is sufficiently increased, or alternatively if $n_{R,1}$ is decreased.

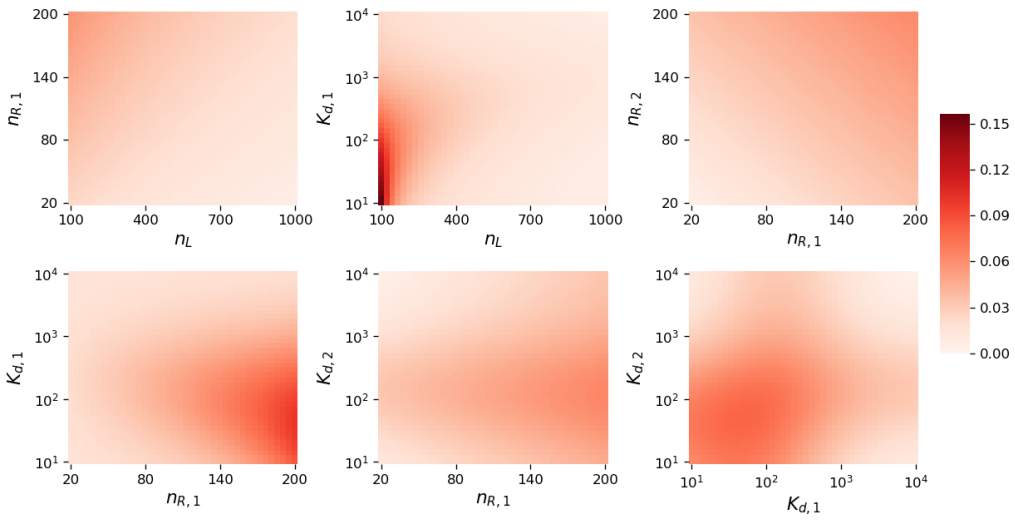

Figure 3.9: HD between the steady state distributions $\{\pi^{EMA}_{(m_1,m_2)}\}_{(m_1,m_2)\in\mathcal{S}_\mathcal{X}}$ and $\{\pi^{MCA}_{(m_1,m_2)}\}_{(m_1,m_2)\in\mathcal{S}_\mathcal{X}}$. Baseline parameter values (that is, the value chosen in each plot for any parameter that has been fixed) are $n_{R,1} = 10^2$, $n_{R,2} = 10^2$, $K_{d,1} = 10^3$, $K_{d,2} = 10^3$ and $n_L = 250$. The threshold parameter $\varepsilon = 10^{-5}$ is used for the MCA.

Although Figure 3.9 seems to indicate that the MCA is a good alternative to the EMA for almost all parameter regimes considered, this is not a thorough exploration since the baseline parameter values are fixed in each subplot. Hence, in order to summarise competition in fewer parameters, *competing strength* pa-

rameters are now introduced, in particular these parameters are $n_{R,j}/K_{d,j}$ for $j \in \{1,2\}$. The summary parameter $n_{R,j}/K_{d,j}$ is an indicator of the competing strength of receptor type j, which will be large whenever there are many of these receptors in the system and/or they have high affinity for the common ligand. Hence larger values of $n_{R,j}/K_{d,j}$ indicate high competition for the ligand from receptor j and smaller values indicate low competition. To use these competing strength parameters, 10^3 parameter sets $(K_{d,1}, K_{d,2}, n_{R,1}, n_{R,2})$ were sampled within the ranges of Table 3.1, from uniform distributions $n_{R,j} \sim Unif(20, 200)$ and $K_{d,j} = 10^x$ with $x \sim Unif(1, 4)$, for $j \in \{1,2\}$. The HD between the EMA and MCA steady state distributions was then calculated for each sampled parameter set and each of three values of $n_L \in \{100, 250, 500\}$, and the results are plotted as scatter plots in Figure 3.10.

It can be observed from Figure 3.10 that, as expected, the HD decreases with increasing n_L, so that in the third plot where $n_L = 500$, the HD is relatively small, even for the largest sampled values of $n_{R,1}/K_{d,1}$ and $n_{R,2}/K_{d,2}$. From the first plot in which $n_L = 100$, representing a high competition scenario, larger values of the HD are observed, especially when $n_{R,1}/K_{d,1}$ and $n_{R,2}/K_{d,2}$ are large. On the other hand, in the $n_L = 250$ plot, smaller values of the competition strength parameters still lead to relatively small distances, indicating that even when n_L is small comparable to $n_{R,1} + n_{R,2}$, the approximation can still be good if $n_{R,1}/K_{d,1}$ and $n_{R,2}/K_{d,2}$ are relatively small.

Although it is clear that there is some disparity between the probability distributions, especially when the competition is the highest, interestingly, the expected numbers of complexes in steady state are almost identical between the two methods of computation. In particular, for all the sets $(K_{d,1}, K_{d,2}, n_{R,1}, n_{R,2}, n_L)$ of parameter values considered in Figure 3.10, the percentage error $100 \cdot |1 - \mathbb{E}^{EMA}[M_j^*]/\mathbb{E}^{MCA}[M_j^*]|$, $j \in \{1,2\}$, when computing the mean values through the MCA, instead of the EMA, is less than 5% in all 10^3 cases. In fact, the overwhelming majority of parameter choices (991 out of 10^3) lead to the same mean values to integer precision, when choosing $\varepsilon = 10^{-5}$ in Algorithm 7, even if the resulting HD is relatively large.

Finally, it is of interest to compare the EMA and MCA derived steady state distributions for example parameter sets where the HD is both high and low. Fig-

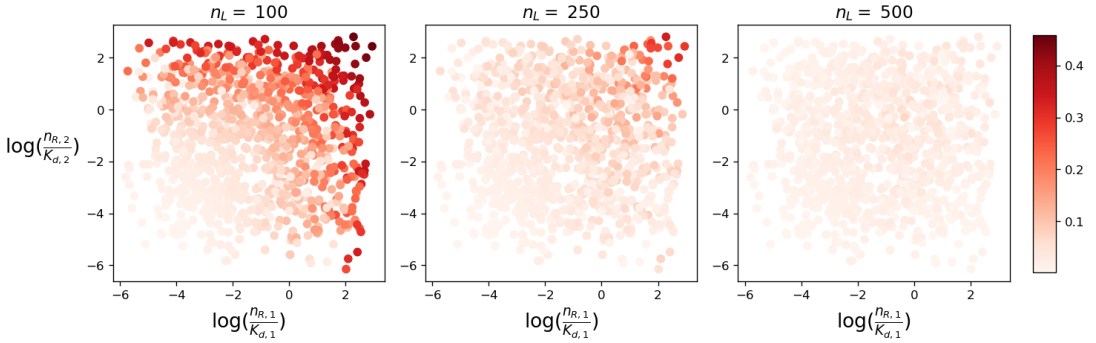

Figure 3.10: HD between the steady state distributions $\{\pi^{EMA}_{(m_1,m_2)}\}_{(m_1,m_2)\in\mathcal{S}_\mathcal{X}}$ and $\{\pi^{MCA}_{(m_1,m_2)}\}_{(m_1,m_2)\in\mathcal{S}_\mathcal{X}}$ plotted for sampled values $n_{R,j} \sim Unif(20, 200)$ and $K_{d,j} = 10^x$ with $x \sim Unif(1, 4)$, for $j \in \{1, 2\}$, and for different numbers of ligand, $n_L \in \{100, 250, 500\}$. The threshold parameter $\varepsilon = 10^{-5}$ is used for the MCA.

ure 3.11 shows the distributions computed via the two methods for one specific set of parameter values, $(K_{d,1}, K_{d,2}, n_{R,1}, n_{R,2})$, and for two values of n_L. The colour of a state represents its steady state probability as indicated by the colour bar. For the smaller value, $n_L = 100$, the HD is 0.24 and for the larger value, $n_L = 500$, the HD is 0.02, indicating that the distributions are very similar. The expected number of type 1 complexes in steady state is stated in each subplot, and since here the numbers of molecules and parameter values are identical for each receptor, $\mathbb{E}^{EMA}[M_1^*] = \mathbb{E}^{EMA}[M_2^*]$ and $\mathbb{E}^{MCA}[M_1^*] = \mathbb{E}^{MCA}[M_2^*]$. Although the HD is relatively large for the smaller value of n_L it is worth noting that the expected number of each complex type in steady state is still very well approximated by the MCA. Also plotted on top of the EMA steady state distributions in this figure are confidence ellipses (Schelp, 2018; Tucker, 2014) generated via the LNA. In particular, by integrating numerically the system of Equations (3.11), the variance and covariance of the fluctuations can be found, *i.e.* $\text{Var}(\xi_i) = \langle\xi_i^2\rangle - \langle\xi_i\rangle^2$ for $i \in \{1, 2\}$ and $\text{Cov}(\xi_1, \xi_2) = \langle\xi_1\xi_2\rangle - \langle\xi_1\rangle\langle\xi_2\rangle$. Using these quantities as elements in the covariance matrix for the fluctuations, the confidence ellipses in Figure 3.11 can be drawn, where the dot-dashed line represents a 1 standard deviation ellipse, the dashed line a 2 standard deviation ellipse and the dotted line a 3 standard deviation ellipse. This approximation appears to capture the distribution very well and is useful as a method of visualisation of the distribution, however it does

not give approximate probabilities for each individual state and hence it is still useful to have the MCA approximation.

3.5 Time scales of complex formation

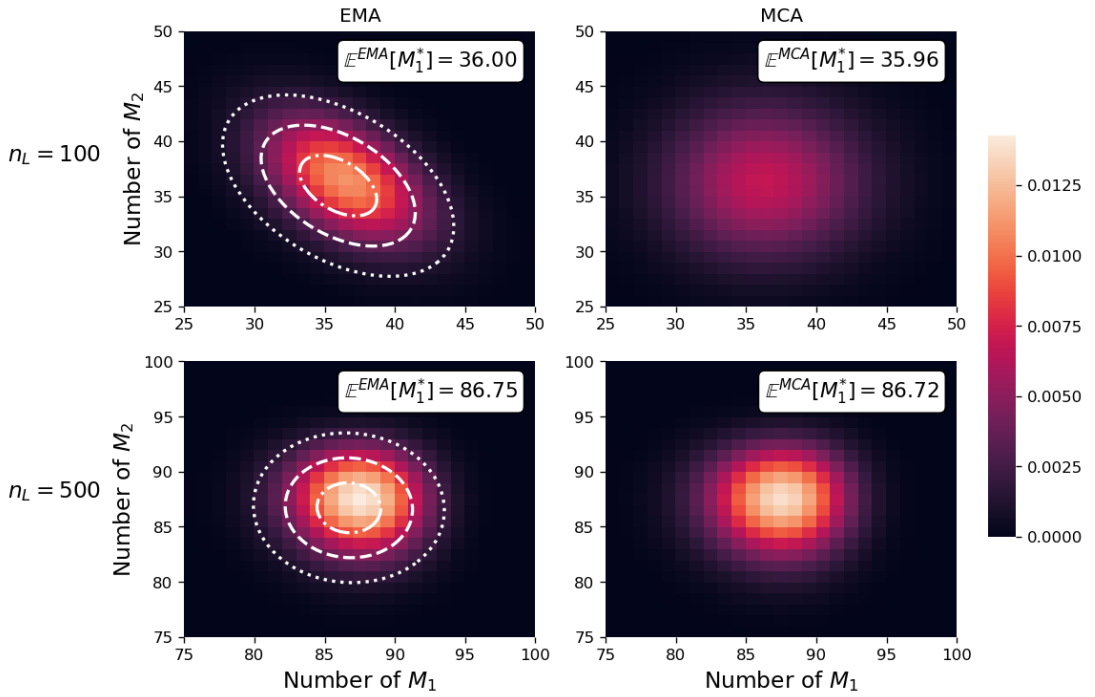

Figure 3.11: Comparison between the steady state distributions computed using the EMA, MCA and LNA, where the colour of a pixel indicates the steady state probability of that state, given by the colour bar. For all subplots the numbers of receptors are $n_{R,1} = n_{R,2} = 100$, and $K_{d,1} = K_{d,2} = 50$.

In this section, the time scales of complex formation are analysed, in particular, the time taken to reach a state in the process in which there are N complexes of one type. For the bi-variate process $\mathcal{X}$, $T_{(m_1,m_2)}(N)$ is defined as the time to reach N complexes of type 2 (without loss of generality, since the arguments presented here would similarly apply to type 1 complexes) given the initial state $(m_1, m_2) \in \mathcal{S}_{\mathcal{X}}$, *i.e.*

$$T_{(m_1,m_2)}(N) \quad = \quad inf\{t \geq 0 : \ M_2(t) = N \mid (M_1(0), M_2(0)) = (m_1, m_2)\},$$

for $(m_1, m_2) \in \mathcal{S}_{\mathcal{X}}$. Here N can be any arbitrary value $M_2(0) < N \leq N_2$, since $T_{(M_1(0),M_2(0))}(M_2(0)) \equiv 0$. Similarly to Section 3.4, this stochastic descriptor is first analysed in Section 3.5.1 via an exact, but computationally expensive, matrix-analytic approach. In Sections 3.5.2 and 3.5.3, alternative, computationally feasible methods of computation are presented, again based on the use of one-dimensional Markov chains to approximate the two-dimensional process $\mathcal{X}$.

3.5.1 Exact matrix-analytic approach (EMA)

In this section, it is shown how one can efficiently analyse $T_{(m_1,m_2)}(N)$, for $(m_1, m_2) \in \mathcal{S}_{\mathcal{X}}$ with $m_2 \leq N$, by means of first-step arguments. In particular, the moment of order l of the time to reach N complexes of type 2 can be computed, starting from any initial state $(m_1, m_2) \in \mathcal{S}_{\mathcal{X}}$. These moments can be obtained from the Laplace-Stieltjes transform of $T_{(m_1,m_2)}(N)$ (see Section 2.1.6), given by

$$\phi^N_{(m_1,m_2)}(z) \quad = \quad \mathbb{E}[e^{-zT_{(m_1,m_2)}(N)}] \, , \quad \mathrm{Re}(z) \geq 0 \, .$$

The index N will be omitted from now on to simplify notation. Using a first-step argument (*i.e.* considering only the first step that the process can take given that it starts in state (m_1, m_2), see Pinsky & Karlin (2010)), it can be written that

$$\begin{aligned} \phi_{(m_1,m_2)}(z) &= \mathbb{E}[e^{-zT_{(m_1,m_2)}}] \\ &= \sum_{(m_1',m_2')} \mathbb{E}[e^{-zT_{(m_1,m_2)}} | (m_1, m_2) \to (m_1', m_2')] \cdot \\ &\qquad \mathbb{P}\left((m_1, m_2) \to (m_1', m_2')\right) , \end{aligned}$$

where (m_1', m_2') is any state that can be possibly reached in one jump of the process, as seen in the transition diagram in Figure 3.2. The random variable $T_{(m_1,m_2)}$ can be split into two parts, $T_{(m_1,m_2)} = t_{(m_1,m_2)\to(m_1',m_2')} + T_{(m_1',m_2')}$, where $t_{(m_1,m_2)\to(m_1',m_2')}$ denotes the time taken for the process to move from state (m_1, m_2) to state (m_1', m_2') in one step, and hence,

$$\phi_{(m_1,m_2)}(z) = \sum_{(m_1',m_2')} \mathbb{E}[e^{-zt_{(m_1,m_2)\to(m_1',m_2')}} | (m_1, m_2) \to (m_1', m_2')] \cdot$$

$$\mathbb{E}[e^{-zT_{(m_1',m_2')}}] \cdot \mathbb{P}\left((m_1,m_2) \to (m_1',m_2')\right),$$

where the expectation of the product becomes the product of the expectations due to the Markov property implying that these two random times are independent. The second expectation in the sum is no longer conditional which is also due to the Markov property. Then given that

$$\mathbb{P}\left((m_1,m_2) \to (m_1',m_2')\right) = \frac{q_{(m_1,m_2),(m_1',m_2')}}{\Delta_{(m_1,m_2)}},$$

where $\Delta_{(m_1,m_2)} = k_{f,1}(n_{R,1}-m_1)(n_L-m_1-m_2)+k_{r,1}m_1+k_{f,2}(n_{R,2}-m_2)(n_L-m_1-m_2)+k_{r,2}m_2$ and that

$$\mathbb{E}[e^{-zt_{(m_1,m_2)\to(m_1',m_2')}}|(m_1,m_2)\to(m_1',m_2')] = \frac{\Delta_{(m_1,m_2)}}{\Delta_{(m_1,m_2)}+z},$$

since $t_{(m_1,m_2)\to(m_1',m_2')}|(m_1,m_2)\to(m_1',m_2')$ is exponentially distributed with rate $\Delta_{(m_1,m_2)}$ and $\mathbb{E}[e^{-zX}] = \frac{a}{a+z}$ if $X \sim Exp(a)$, one arrives at the system of linear equations

$$\begin{aligned}\phi_{(m_1,m_2)}(z) = &\frac{k_{f,1}(n_{R,1}-m_1)(n_L-m_1-m_2)}{z+\Delta_{(m_1,m_2)}}\phi_{(m_1+1,m_2)}(z) + \frac{k_{r,1}m_1}{z+\Delta_{(m_1,m_2)}}\phi_{(m_1-1,m_2)}(z)\\ &+\frac{k_{f,2}(n_{R,2}-m_2)(n_L-m_1-m_2)}{z+\Delta_{(m_1,m_2)}}\phi_{(m_1,m_2+1)}(z)\frac{k_{r,2}m_2}{z+\Delta_{(m_1,m_2)}}\phi_{(m_1,m_2-1)}(z).\end{aligned}$$

Boundary conditions are given by $\phi_{(m_1,m_2)}(z) \equiv 1$ if $m_2 = N$, since $T_{(m_1,m_2)}(m_2) \equiv 0$. The lth order moment of $T_{(m_1,m_2)}$ can then be found by differentiating $\phi_{(m_1,m_2)}(z)$ l times and multiplying by $(-1)^l$, *i.e.*

$$\mathbb{E}[T^l_{(m_1,m_2)}] = (-1)^l \left.\frac{d^l}{dz^l}\phi_{(m_1,m_2)}(z)\right|_{z=0}, \quad l \geq 1.$$

One can then compute the lth order moment of $T_{(m_1,m_2)}$, $\mathbb{E}[T^l_{(m_1,m_2)}]$, by solving the system of linear equations:

$$\begin{aligned}\Delta_{(m_1,m_2)}\mathbb{E}[T^l_{(m_1,m_2)}] = &k_{f,1}(n_{R,1}-m_1)(n_L-m_1-m_2)\mathbb{E}[T^l_{(m_1+1,m_2)}]\\ &+k_{r,1}m_1\mathbb{E}[T^l_{(m_1-1,m_2)}]\end{aligned}$$

$$
\begin{aligned}
&+ k_{f,2}(n_{R,2} - m_2)(n_L - m_1 - m_2)\mathbb{E}[T^l_{(m_1,m_2+1)}] \\
&+ k_{r,2} m_2 \mathbb{E}[T^l_{(m_1,m_2-1)}] \\
&+ l\ \mathbb{E}[T^{l-1}_{(m_1,m_2)}].
\end{aligned} \tag{3.25}
$$

Moreover, by arranging the states in $\mathcal{S}_\chi$ into levels as in Section 3.4.1, each equation in the system (3.25) corresponds to an initial state $(m_1, m_2) \in \cup_{k=0}^{N-1} L(k)$, so that one can rewrite Equation (3.25) in matrix form as follows:

$$
\boldsymbol{m}^{(l)} = \boldsymbol{A}\boldsymbol{m}^{(l)} + \boldsymbol{b}^{(l)} \tag{3.26}
$$

where

$$
\boldsymbol{m}^{(l)} = \begin{pmatrix} \boldsymbol{m}_0^{(l)} \\ \boldsymbol{m}_1^{(l)} \\ \vdots \\ \boldsymbol{m}_{N-1}^{(l)} \end{pmatrix}, \quad \boldsymbol{m}_k^{(l)} = \begin{pmatrix} \mathbb{E}[T^l_{(0,k)}] \\ \mathbb{E}[T^l_{(1,k)}] \\ \vdots \\ \mathbb{E}[T^l_{(min(n_{R,1}, n_L - k), k)}] \end{pmatrix}, \ 0 \le k \le N-1.
$$

Vector $\boldsymbol{b}^{(l)}$ is also organised in sub-vectors as

$$
\boldsymbol{b}^{(l)} = \begin{pmatrix} \boldsymbol{b}_0^{(l)} \\ \boldsymbol{b}_1^{(l)} \\ \vdots \\ \boldsymbol{b}_{N-1}^{(l)} \end{pmatrix}
$$

which are obtained from the $(l-1)$th order moments, as

$$
(\boldsymbol{b}_k^{(l)})_{(m_1,m_2)} = \frac{l}{\Delta_{(m_1,m_2)}} (\boldsymbol{m}_k^{(l-1)})_{(m_1,m_2)},
$$

for all $(m_1, m_2) \in L(k)$; note that sub-vectors $\boldsymbol{m}_k^{(0)}$ are just column vectors of ones. The matrix $\boldsymbol{A}$ does not depend on l and is given by

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{0,0} & \boldsymbol{A}_{0,1} & \boldsymbol{0} & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{A}_{1,0} & \boldsymbol{A}_{1,1} & \boldsymbol{A}_{1,2} & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{2,1} & \boldsymbol{A}_{2,2} & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{A}_{N-2,N-2} & \boldsymbol{A}_{N-2,N-1} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{A}_{N-1,N-2} & \boldsymbol{A}_{N-1,N-1} \end{pmatrix}.$$

The sub-matrices $\boldsymbol{A}_{k,k'}$ are defined as follows:

- For $1 \leq k \leq N-1$,

$$(\boldsymbol{A}_{k,k-1})_{i,j} = \begin{cases} \frac{k_{r,2}k}{\Delta_{(i,k)}}, & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

  where $0 \leq i \leq J(k)$ and $0 \leq j \leq J(k-1)$.

- For $0 \leq k \leq N-2$,

$$(\boldsymbol{A}_{k,k+1})_{i,j} = \begin{cases} \frac{k_{f,2}(n_{R,2}-k)(n_L-i-k)}{\Delta_{(i,k)}}, & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

  where $0 \leq i \leq J(k)$ and $0 \leq j \leq J(k+1)$.

- For $0 \leq k \leq N-1$,

$$(\boldsymbol{A}_{k,k})_{i,j} = \begin{cases} \frac{k_{f,1}(n_{R,1}-i)(n_L-i-k)}{\Delta_{(i,k)}}, & \text{if } j = i+1, \\ \frac{k_{r,1}i}{\Delta_{(i,k)}}, & \text{if } j = i-1, \\ 0, & \text{otherwise,} \end{cases}$$

  where $0 \leq i \leq J(k)$ and $0 \leq j \leq J(k)$.

Equation (3.26) can then be solved efficiently using Algorithm 8 to obtain the lth order moments of the random variable $T_{(m_1,m_2)}$. In this algorithm, $\boldsymbol{I}_a$ denotes the $a \times a$ identity matrix.

Algorithm 8 Moments of the random variable $T_{(m_1,m_2)}$.

$(\boldsymbol{b}_k^{(1)})_{(m_1,m_2)} = \frac{1}{\Delta_{(m_1,m_2)}}$ for all $(m_1, m_2) \in L(k)$, for $k = 0, \ldots, N-1$.
$\boldsymbol{R}_0 = \boldsymbol{I}_{J(0)} - \boldsymbol{A}_{0,0}$.
for $p = 1, \ldots, l$ **do**:
    $\boldsymbol{S}_0 = \boldsymbol{R}_0^{-1} \cdot \boldsymbol{b}_0^{(p)}$.
    for $k = 1, \ldots N-1$ **do**:
        $\boldsymbol{R}_k = \boldsymbol{I}_{J(k)} - \boldsymbol{A}_{k,k} - \boldsymbol{A}_{k,k-1}\boldsymbol{R}_{k-1}^{-1}\boldsymbol{A}_{k-1,k}$.
        $\boldsymbol{S}_k = \boldsymbol{R}_k^{-1}\boldsymbol{A}_{k,k-1}\boldsymbol{S}_{k-1} + \boldsymbol{R}_k^{-1}\boldsymbol{b}_k^{(p)}$.
    end for
    $\boldsymbol{m}_{N-1}^{(p)} = \boldsymbol{S}_{N-1}$.
    for $k = N-2, \ldots, 0$ **do**:
        $\boldsymbol{m}_k^{(p)} = \boldsymbol{S}_k + \boldsymbol{R}_k^{-1}\boldsymbol{A}_{k,k+1}\boldsymbol{m}_{k+1}^{(p)}$.
    end for
    $(\boldsymbol{b}_k^{(p+1)})_{(m_1,m_2)} = \frac{p+1}{\Delta_{(m_1,m_2)}}(\boldsymbol{m}_k^{(p)})_{(m_1,m_2)}$, for all $(m_1, m_2) \in L(k)$, for $k = 0, \ldots, N-1$.
end for
return $\boldsymbol{m}^{(l)} = (\boldsymbol{m}_0^T, \ldots, \boldsymbol{m}_{N-1}^T)^T$. ▷ lth order moment of $T_{(m_1,m_2)}$

With Algorithm 8 and $l = 1$, one can then compute the mean time to reach N type 2 complexes in an exact way, which is denoted by $\mathbb{E}^{EMA}[T_{(m_1,m_2)}(N)]$, making use of the exact matrix-analytic methodology. As in the case of the steady state distribution, the drawback of this method is that it is computationally expensive as it involves the inversion of matrices, which increase in size as the state space increases. Thus an alternative method is required which is feasible for larger numbers of molecules $n_{R,1}$, $n_{R,2}$ and n_L. Such methods are proposed in the following sections, drawing again from the theory of one-dimensional Markov birth-and-death processes.

3.5.2 No competition approximation (NCA)

In this section, an approximation to the mean time to reach N type 2 complexes is presented, which is a good alternative in the large n_L limit. Specifically, one can analyse this mean time for the independent one-dimensional birth-and-death process $\mathcal{X}_2$, *i.e.* $T_{m_2}(N) = inf\{t \geq 0 : M_2(t) = N \mid M_2(0) = m_2\}$. Since, in an experimental setting, it is often the starting point of the experiment to stimulate cells with ligand, it is reasonable to focus on the mean value $\mathbb{E}[T_{m_2}(N)]$ for $m_2 = 0$ (that is, no type 2 complexes initially in the system), although the arguments presented here can be easily modified for different initial states $0 < m_2 < N$, or for higher order moments of $T_{m_2}(N)$. It is clear that, for $N = 1$, $\mathbb{E}[T_0(1)] = \frac{1}{\lambda_0}$, where $\lambda_{m_2} = k_{f,2}(n_{R,2} - m_2)(n_L - m_2)$ and $\mu_{m_2} = k_{r,2}m_2$.

Then, using a first-step argument, one can compute $\mathbb{E}[T_i(i+1)]$ (Allen, 2010). Firstly, given that from state i, the process can move only to either state $i+1$ or state $i-1$,

$$\begin{aligned}\mathbb{E}[T_i(i+1)] &= \mathbb{P}(i \to i+1) \cdot \mathbb{E}[T_i(i+1)|i \to i+1] \\ &\quad + \mathbb{P}(i \to i-1) \cdot \mathbb{E}[T_{i-1}(i+1)|i \to i-1].\end{aligned}$$

Then, since the interevent times in a Markov chain are exponentially distributed, $\mathbb{E}[T_i(j)|i \to j] = \frac{1}{\lambda_i+\mu_i}$ for $j \in \{i-1, i+1\}$, and hence,

$$\begin{aligned}\mathbb{E}[T_i(i+1)] &= \mathbb{P}(i \to i+1) \cdot \mathbb{E}[T_i(i+1)|i \to i+1] \\ &\quad + \mathbb{P}(i \to i-1) \cdot \mathbb{E}[T_{i-1}(i+1)|i \to i-1] \\ &= \frac{\lambda_i}{\lambda_i+\mu_i}\left(\frac{1}{\lambda_i+\mu_i}\right) \\ &\quad + \frac{\mu_i}{\lambda_i+\mu_i}\left(\frac{1}{\lambda_i+\mu_i} + \mathbb{E}[T_{i-1}(i)] + \mathbb{E}[T_i(i+1)]\right) \\ &= \frac{1}{\lambda_i+\mu_i} + \frac{\mu_i}{\lambda_i+\mu_i}\left(\mathbb{E}[T_{i-1}(i)] + \mathbb{E}[T_i(i+1)]\right),\end{aligned}$$

and then rearranging yields,

$$\mathbb{E}[T_i(i+1)] = \frac{1}{\lambda_i} + \frac{\mu_i}{\lambda_i}\mathbb{E}[T_{i-1}(i)].$$

This allows one to recursively obtain

$$\mathbb{E}[T_0(N)] \quad = \quad \frac{1}{\lambda_0} + \sum_{j=1}^{N-1}\left[\frac{\mu_1\ldots\mu_j}{\lambda_0\ldots\lambda_j}\left(\sum_{i=1}^{j}\frac{\lambda_0\ldots\lambda_{i-1}}{\mu_1\ldots\mu_i}+1\right)\right],$$

which is similar to the well-known extinction time expressions for a one-dimensional birth-and-death process (Allen, 2010). Thus, the following approximation is proposed,

$$\mathbb{E}^{NCA}[T_{(0,0)}(N)] \quad = \quad \frac{1}{\lambda_0} + \sum_{j=1}^{N-1}\left[\frac{\mu_1\ldots\mu_j}{\lambda_0\ldots\lambda_j}\left(\sum_{i=1}^{j}\frac{\lambda_0\ldots\lambda_{i-1}}{\mu_1\ldots\mu_i}+1\right)\right], \quad (3.27)$$

for situations in which $n_L \gg n_{R,1} + n_{R,2}$.

When n_L is comparable to $n_{R,1} + n_{R,2}$, it is clear that $\mathbb{E}^{NCA}[T_{(0,0)}(N)]$ will underestimate the exact mean time to reach N type 2 complexes, $\mathbb{E}^{EMA}[T_{(0,0)}(N)]$, since it does not take into account the ligands being depleted by the formation of complexes in the competing Markov chain. Thus, in Section 3.5.3 an alternative method is introduced (MCA) to calculate $\mathbb{E}[T_{(0,0)}(N)]$, which can be used under low-to-moderate competition scenarios (*i.e.* when $n_L \approx n_{R,1} + n_{R,2}$).

3.5.3 Moderate competition approximation (MCA)

Since the NCA underestimates the time to reach N type 2 complexes, in particular in situations in which there is high competition from the receptors of type 1, in this section the MCA is defined, in order to approximately account for the depletion of ligand. In the MCA, one should compute $\mathbb{E}^{MCA}[T_{(0,0)}(N)]$ by implementing Equation (3.27), but with an effective number of free ligands

$$n_L^* = n_L - \mathbb{E}[M_1^*], \quad (3.28)$$

where $\mathbb{E}[M_1^*]$ is the mean number of type 1 complexes in steady state. Ideally, one would use $\mathbb{E}^{EMA}[M_1^*]$ in Equation (3.28). Alternatively, if this is computationally not feasible, one can use its approximation $\mathbb{E}^{MCA}[M_1^*]$ instead. From a practical perspective this does not seem crucial, since as stated in Section 3.4.4, $\mathbb{E}^{EMA}[M_1^*]$ and $\mathbb{E}^{MCA}[M_1^*]$ are almost equal (to integer value) in most of the scenarios considered. The idea behind this effective number of free ligands is that by subtracting from n_L, the expected number of type 1 complexes in steady state, one removes from the beginning of the calculation, the expected number of ligands which will be occupied by the competing receptors in the long run. It is clear that this approximation will work best when the complexes of type 1 form rapidly, in which case it is reasonable to remove the ligands from the beginning of the calculation for $\mathbb{E}^{MCA}[T_{(0,0)}(N)]$. This corresponds to situations in which $k_{f,1}$ is large, in comparison to $k_{f,2}$ and $k_{r,1}$. In the case where the complexes of type 1 form more slowly, it is expected that this approximation may not perform as well. Unlike in the case of the steady state (Sections 3.4.1 - 3.4.3), where the NCA was just a necessary step to propose the MCA, which always behaved better than the NCA for the steady state computation, it is expected that for the mean times, the EMA will be approximated best by the NCA or the MCA depending on the number of molecules $n_{R,1}$, $n_{R,2}$ and n_L and the parameter values $k_{f,1}$, $k_{f,2}$, $k_{r,1}$ and $k_{r,2}$.

To give a first insight into how each of the approximations compare with the exact result, in Figure 3.12 the mean time to reach N complexes of type 2 is plotted, where N is represented as a percentage P of the steady state number of type 2 complexes, $N = \frac{P}{100}\mathbb{E}^{EMA}[M_2^*]$, by means of the EMA, NCA and MCA. In low competition scenarios (*e.g.* large n_L or when $K_{d,1} \gg K_{d,2}$ so that the dynamics of formation of type 1 complexes does not significantly affect the time scales of formation of type 2 complexes), the three approaches lead to almost indistinguishable results. For low-to-moderate competition settings, the MCA leads to a reasonable approximation. It is worth noting that the MCA and NCA seem to lead to lower and upper bounds for the mean time under analysis: the NCA always underestimates the time scales of type 2 complex formation, by considering that there are n_L ligands available, neglecting ligand depletion due to type 1 complex formation competition; while the MCA tends to overestimate the

time scales of type 2 complex formation because one considers only $n_L - \mathbb{E}[M_1^*]$ ligands available (that is, all the ligands occupied by steady state type 1 complexes are removed from the beginning, even if these complexes may take some time to form). Since this steady state amount of type 1 complexes takes some time to form, some of these ligands might still be able to contribute to the formation of type 2 complexes during early times, leading to the observed overestimation. Still, it is striking how well both approximations work when $K_{d,1} \gg K_{d,2}$, even when the number of ligands is less than the total number of receptors available in the system ($n_L = 100 < 200 = n_{R,1} + n_{R,2}$). The relationship between the EMA, NCA and MCA is explored more thoroughly in terms of the parameter values and numbers of molecules in Section 3.5.4.

Plotted also in the insets of the subplots in Figure 3.12 are again the EMA mean times (green lines) and also the times that are predicted from the deterministic (ODE) competition model derived via the LNA in Equations (3.9) (black dashed lines). In this case the deterministic model was numerically solved (for numbers of molecules rather than concentrations) and the first time at which the number of complexes of type 2 exceeded $N = \frac{P}{100}\mathbb{E}^{EMA}[M_2^*]$ was recorded and is plotted. In general, it can be seen that the deterministic time is always greater than or equal to the stochastic EMA mean time, especially for larger values of P. In order to further compare the deterministic result with the stochastic result, one could also compute higher order moments of the times, such as the variance of the times, using Algorithm 8 with $l = 2$.

3.5.4 Numerical validation

In this section, the accuracy of the NCA and MCA for analysing the expected time to reach N complexes of type 2 is assessed, by means of numerical comparison with the EMA. Similarly to the analysis carried out in Figure 3.9 for the steady state distributions, in Figure 3.13 the parameters $(K_{d,1}, K_{d,2}, n_L, n_{R,1}, n_{R,2})$ are varied in pairs, and heatmaps of the relative differences,

$$1 - \frac{\mathbb{E}^{NCA}[T_{(0,0)}(N)]}{\mathbb{E}^{EMA}[T_{(0,0)}(N)]} \quad \text{and} \quad 1 - \frac{\mathbb{E}^{MCA}[T_{(0,0)}(N)]}{\mathbb{E}^{EMA}[T_{(0,0)}(N)]},$$

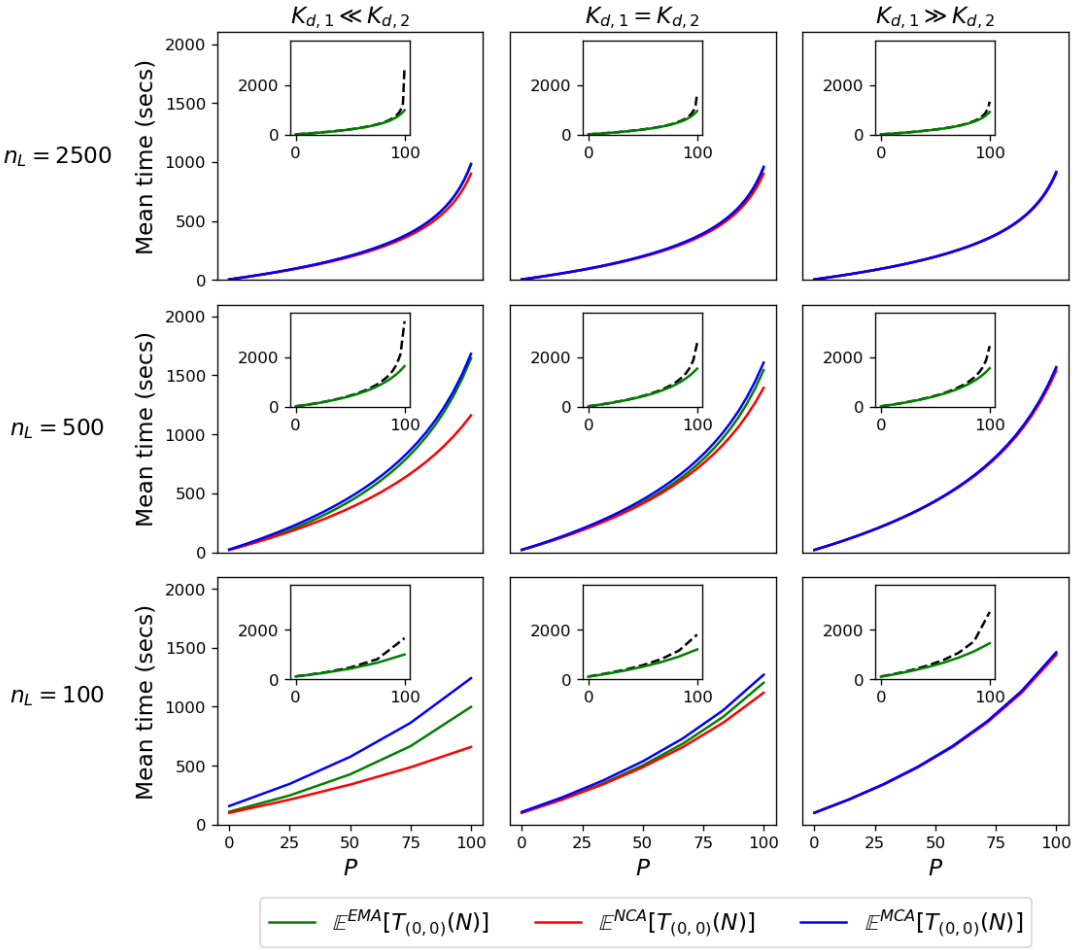

Figure 3.12: Comparison between the expected times to reach N complexes of type 2, computed using the EMA, NCA and the MCA, for $K_{d,2} = 10^3$, different values of $n_L \in \{100, 500, 2500\}$ and $K_{d,1} \in \{10^2, 10^3, 10^4\}$ (*i.e.* $K_{d,1} \ll K_{d,2}$, $K_{d,1} = K_{d,2}$ or $K_{d,1} \gg K_{d,2}$). For all subplots the number of receptors are $n_{R,1} = n_{R,2} = 10^2$, and the initial state is $(m_1, m_2) = (0, 0)$. P represents a percentage of the mean steady state value, so that $N = \frac{P}{100}\mathbb{E}^{EMA}[M_2^*]$. The insets of each subplot show the EMA mean times (green lines) as well as the corresponding times found from solving the deterministic competition model (dashed black lines).

are plotted, for $N = \mathbb{E}^{EMA}[M_2^*]$, so that $T_{(0,0)}(N)$ encodes the time scales for type 2 complexes to reach steady state. It can be seen that the MCA seems to better approximate the exact mean time in almost all parameter regimes explored, but it is expected, that the NCA might perform better in some scenarios, for example, when N is small. Still, it can be seen from Figure 3.13 (left) that

the NCA performs relatively well in parameter regimes representing low natural competition from R_1, for example when $K_{d,1}$ is relatively large (indicative of low affinity of R_1 for L) and/or $n_{R,1}$ is small. In Figure 3.13 (right), this effect of competition from receptors of type 1 is reduced, and the MCA in general performs much better thanks to the effective number of ligands considered in Equation (3.28), which takes into account ligand depletion due to receptor competition. The MCA and the NCA lead to very similar, and almost exact, results under very low competition scenarios, since when $\mathbb{E}[M_1^*] \approx 0$, $n_L^* \approx n_L$ in Equation (3.28) and thus, the two approximations are effectively the same.

As the analysis in Figure 3.13 uses fixed baseline parameters and hence does not explore the whole parameter space, the EMA and NCA/MCA can be compared by looking again at the competition strength parameters $n_{R,1}/K_{d,1}$ and $n_{R,2}/K_{d,2}$ as introduced in Section 3.4.4. Similar scatter plots to those in Figure 3.10 are plotted in Figure 3.14, where now the relative differences between the mean times are plotted instead of the HD. From the left hand subplot in the top row of Figure 3.14 (where the top row corresponds to the comparison between the EMA and the NCA), it can be seen that the relative difference increases with the competing strength parameter $\frac{n_{R,1}}{K_{d,1}}$. However, similarly to the case of the steady state distributions, as n_L increases, the competition from R_1 decreases and the NCA improves. For the sampled parameter sets, in general it seems that the MCA outperforms once again the NCA. Similarly to the NCA, the performance of the MCA tends to improve with large enough values of n_L, where in the right hand plot of the bottom row of Figure 3.14 (where the bottom row corresponds to the comparison between the EMA and the MCA) in which $n_L = 500$, more than half of the sampled parameter sets lead to an approximation with a relative error smaller than 5%. It is clear that the MCA will behave well in situations where n_L is large enough and, in particular, of a different order of magnitude to $n_{R,1}+n_{R,2}$; for example, for $n_L = 2000$ (data not shown here; note that $n_{R,1}+n_{R,2}$ ranges from 40 to 400), an overwhelming majority of scenarios (more than 90%) in the bottom row of Figure 3.14 lead to MCA predictions with relative errors smaller than 5%. Encouragingly also, in the top row of Figure 3.14, all of the sampled points have relative differences greater than 0, which implies that the NCA always underestimates the EMA. This is expected given that the formation

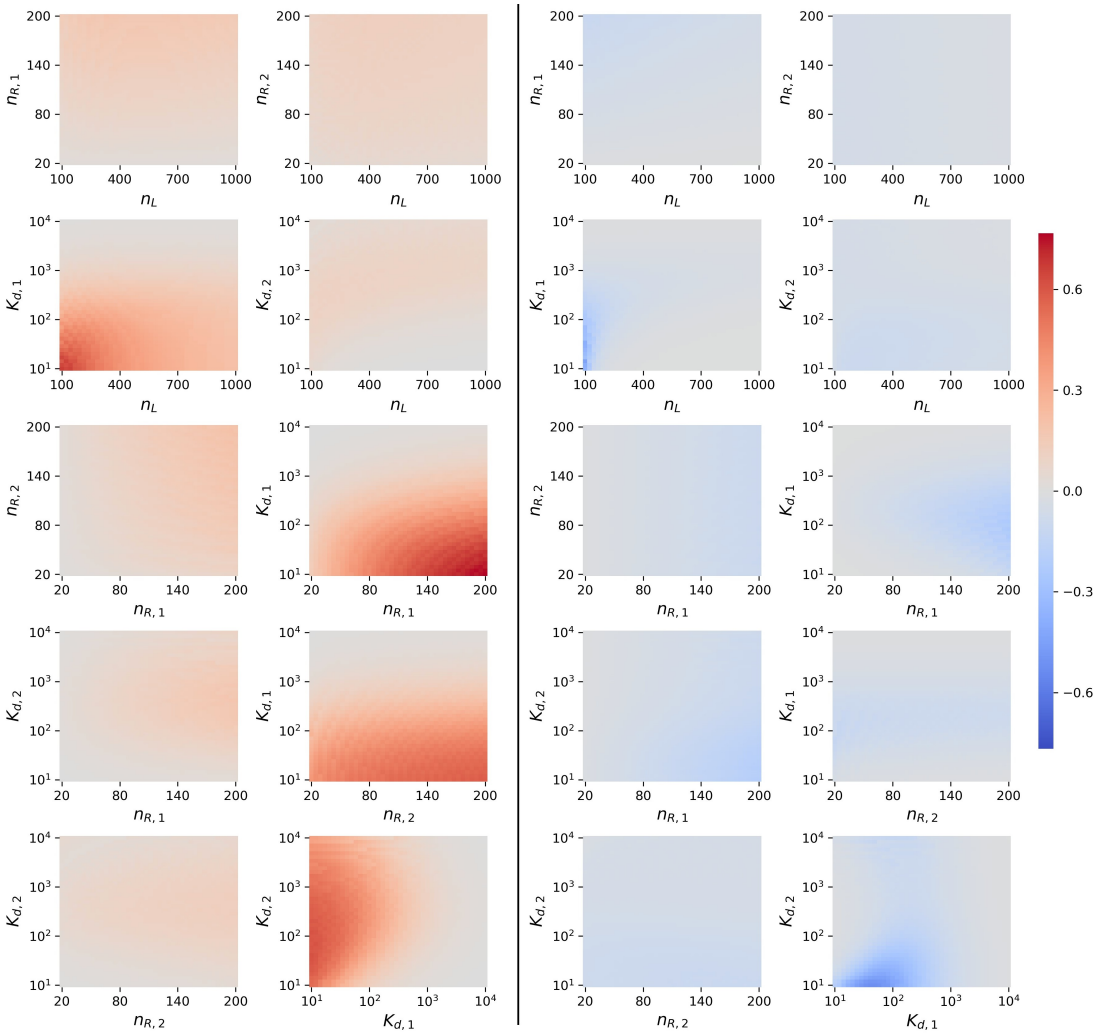

Figure 3.13: Left: Relative difference $1 - \frac{\mathbb{E}^{NCA}[T_{(0,0)}(N)]}{\mathbb{E}^{EMA}[T_{(0,0)}(N)]}$ between the EMA and NCA derived expected times to reach N complexes of type 2. **Right:** Relative difference $1 - \frac{\mathbb{E}^{MCA}[T_{(0,0)}(N)]}{\mathbb{E}^{EMA}[T_{(0,0)}(N)]}$ between the EMA and MCA derived expected times to reach N complexes of type 2. The colour of a pixel indicates the relative difference for this combination of parameters, as given by the colour bar. Baseline parameter values are $n_{R,1} = 10^2$, $n_{R,2} = 10^2$, $K_{d,1} = 10^3$, $K_{d,2} = 10^3$ and $n_L = 250$.

of complexes of type 2 will happen faster if there is no competition from another receptor, as assumed in the NCA. A similar pattern can be seen in the bottom row of Figure 3.14, in which the MCA tends to overestimate the EMA, leading to negative values of the relative difference. This is true for the majority of the

points sampled. However, it is not always true as 6 of the 10^3 points sampled lead to very slightly positive relative differences (for $n_L = 100$ and $n_L = 250$ only).

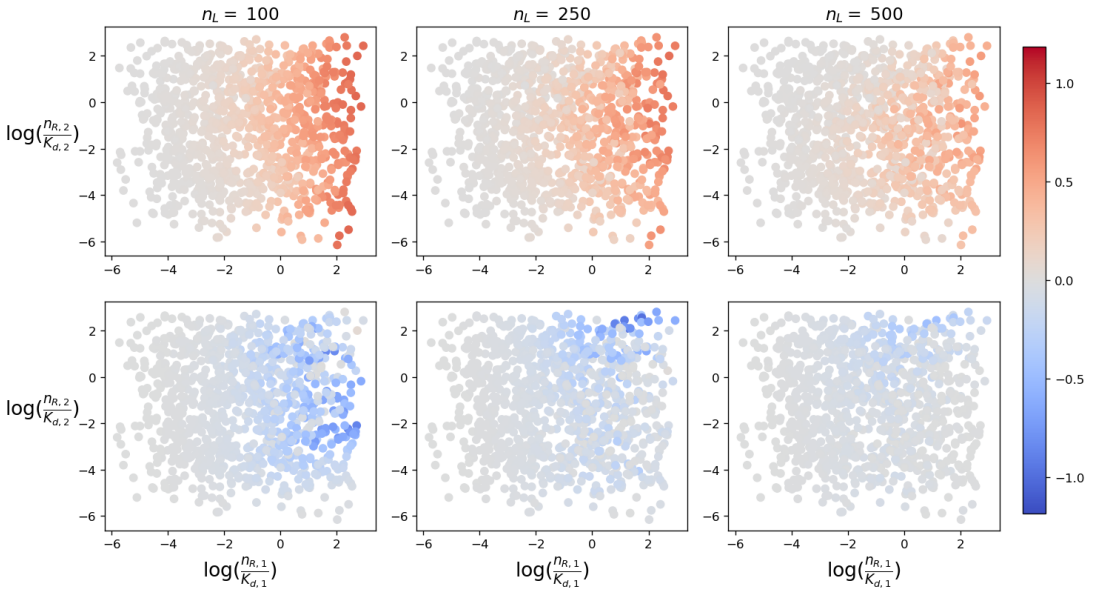

Figure 3.14: Top row: Relative difference $1 - \frac{\mathbb{E}^{NCA}[T_{(0,0)}(N)]}{\mathbb{E}^{EMA}[T_{(0,0)}(N)]}$ between the EMA and NCA mean times to reach N complexes of type 2. **Bottom row:** Relative difference $1 - \frac{\mathbb{E}^{MCA}[T_{(0,0)}(N)]}{\mathbb{E}^{EMA}[T_{(0,0)}(N)]}$ between the EMA and MCA mean times to reach N complexes of type 2. Both rows of the figure are plotted for the 10^3 sampled parameter values in Figure 3.10, and for $n_L \in \{100, 250, 500\}$.

Finally, it is to be expected that the NCA may behave better than the MCA in situations where the mean steady state number of type 1 complexes is perhaps non-negligible, but the time scales of type 1 complex formation are significantly slower than the time scales of type 2 complex formation. In these situations, considering an effective number of ligands $n_L - \mathbb{E}[M_1^*]$ in Equation (3.28) might lead to worse predictions, since this ligand depletion might take a long time to occur, and should, thus, be neglected, so that the NCA would prevail. For example, the NCA seems to lead to better predictions than the MCA in Figure 3.15 (left), where the formation of type 2 complexes occurs more rapidly than that of type 1 complexes. In this figure, by carrying out a single stochastic (Gillespie) simulation of the process, a particular realisation of the time $T_{(0,0)}(N)$ to reach $N = \frac{P}{100}\mathbb{E}^{EMA}[M_2^*]$ type 2 complexes is produced, for different values of P; in

particular, for $P = 20$, 40, 60, 80 and 100% of the average number of complexes in steady state, one can produce the dots plotted in Figure 3.15 (top). On the other hand, in situations where the type 1 receptor is a strong competitor, such as in Figure 3.15 (right), the MCA outperforms the NCA.

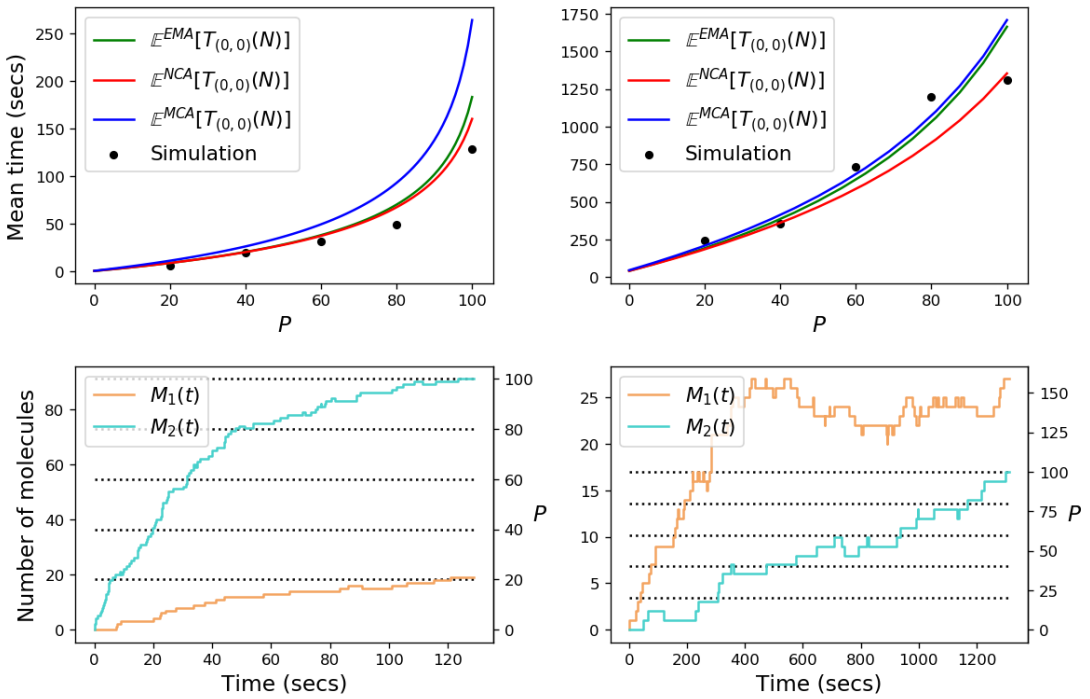

Figure 3.15: Top row: Comparison between the mean times computed using the EMA, NCA and MCA. P represents a percentage of the mean steady state value, as in Figure 3.12, so that $N = \frac{P}{100}\mathbb{E}^{EMA}[M_2^*]$. **Bottom row:** Stochastic realisations of the processes analysed in the top scenarios, leading to stochastic realisations (plotted as black dots in top plots) of random variable $T_{(0,0)}(N)$ for $N = \frac{P}{100}\mathbb{E}^{EMA}[M_2^*]$ and different values of P. **Left:** $n_{R,1} = n_{R,2} = 10^2$, $K_{d,1} = 10^2$ and $K_{d,2} = 10$. **Right:** $n_{R,1} = 50$, $n_{R,2} = 10^2$, $K_{d,1} = 10$ and $K_{d,2} = 10^3$. The number of ligands is $n_L = 250$ in each subplot.

3.5.5 Time scales of productive complex formation

Thus far, two stochastic descriptors have been analysed for the receptor competition system in Figure 3.1, the steady state distribution and the expected time to reach N complexes of type 2. In this section, an extension of the approximation

methodology is presented, for a third descriptor of interest. This descriptor is linked to signal initiation for some receptor tyrosine kinases (RTKs): the time to reach a threshold number N of *productive* complexes on the cell surface. Currie *et al.* (2012) and Castro *et al.* (2014), hypothesise that signal initiation of T cells through the T cell receptor (TCR) is determined by the time a threshold number, N, of productive complexes is reached. A productive complex is considered to be a receptor that remains bound to the ligand, for at least a dwell time τ. The authors compute the mean time, $T(N,\tau)$ to reach N productive complexes. Note that $T_{(0,0)}(N)$ represents the time to reach a threshold number N of simultaneously bound complexes in the system, while $T(N,\tau)$ does not require these events to be simultaneous, but instead requires the corresponding complexes to be productive, based on the hypothesis that "*counting devices are at work to allow signal accumulation, decoding and translation into biological responses*" (Currie *et al.*, 2012).

The model in Figure 3.1 is proposed by Currie *et al.* (2012) and when analysing $T(N,\tau)$ with a single receptor type (the TCR), under the approximation of ligand excess they find,

$$T(N,\tau) \;=\; \tau + \frac{1}{k_f n_L}\left(\frac{N'}{n_R} + \frac{1}{2!}\left(\frac{N'}{n_R}\right)^2 + \dots\right), \tag{3.29}$$

where k_f is the forward binding rate of the receptor, n_R is the number of receptors available in the system, and $N' = exp(k_r\tau)N$ represents the average number of binding events required for N productive ones to be reached. Recall that dissociation of a complex occurs after an exponentially distributed random time $Exp(k_r)$, where k_r is the dissociation rate of the receptor. Equation (3.29) is found by considering that $T(N,\tau) = \tau + \mathbb{E}[t_{N'}]$, where t_i is defined as the time that the ith complex is formed. Then, under excess of ligand, and considering that the number of bound receptors is much less than the total number of receptors in the system, $\mathbb{E}[t_{N'}] = \frac{N'}{k_f n_L n_R}$, since $t_{N'}$ is a sum of N' exponentially distributed random variables with mean $(k_f n_L n_R)^{-1}$. Hence,

$$T(N,\tau) = \tau + \frac{N'}{k_f n_L n_R}. \tag{3.30}$$

One can then consider a differential equation for $M(t)$, the number of bound receptors at time t, under the same assumptions (excess of ligand and mass action kinetics). Then the mean number of binding events up to time t can be found by the solution of the integral

$$C(t) = \int_0^t k_f n_L (n_R - M(s)) ds.$$

Finally, one can set $C(T - \tau) = N'$, since both expressions correspond to the mean number of binding events before the Nth productive one, and rearranging and substituting the expression found for $T(N, \tau)$ (Equation (3.30)), one arrives at Equation (3.29).

Currie *et al.* (2012) make use of experimental data to test two different hypotheses: that (a) the time scale of a T cell response correlates with the time it takes to have had N receptor-ligand complexes bound for at least a threshold dwell time, τ, each; or that (b) the time scale of a T cell response correlates with the time a threshold number, N, of TCRs must be occupied at equilibrium. Their conclusion is that experimental data supports hypotheses (a), but not (b). The descriptor $T(N, \tau)$ has been proposed by Currie *et al.* (2012) and Castro *et al.* (2014) for the T cell receptor, and a similar hypothesis has been considered for other RTKs (Alarcón & Page, 2006; López-García *et al.*, 2018). Here, similar methods as those presented for the previous descriptors, are used to compute $T(N, \tau)$ for a model as described in Figure 3.1, where the receptor of interest for signal initiation (receptor 1) has a competitor (receptor 2) for binding the ligand.

By following similar arguments to those for the other descriptors, the NCA approximation for $T(N, \tau)$ consists of applying Equation (3.29) with the total number of ligands in the system n_L, thus neglecting competition from receptor 2. On the other hand, as described in the MCA approximation, one may use Equation (3.29) with n_L replaced by the effective number of ligands $n_L^* = n_L - \mathbb{E}[M_2^*]$.

Figure 3.16 shows a comparison between the mean time $T(N, \tau)$, computed making use of the NCA and MCA approximations, and numerical simulations ($T^{SIM}(N, \tau)$). The matrix-analytic approach for this descriptor is not applicable since τ is a deterministic (*i.e.* fixed) dwell time. It is seen that the time $T(N, \tau)$

to reach N productive type 1 complexes is not well estimated by the NCA (using Equation (3.29) with n_L ligands) in situations where n_L is not large enough (*e.g.* $n_L = 100$ in Figure 3.16). The worst estimates are obtained in scenarios where the competing strength of receptor 1 is low, and that of receptor 2 is high, as one would expect. On the other hand, the MCA approach, using an effective number of ligands n_L^*, is better, even for a small number of ligands. In general, it can be observed that the MCA approximation gives good estimates of $T(N,\tau)$ over a wide range of parameter values: N and τ.

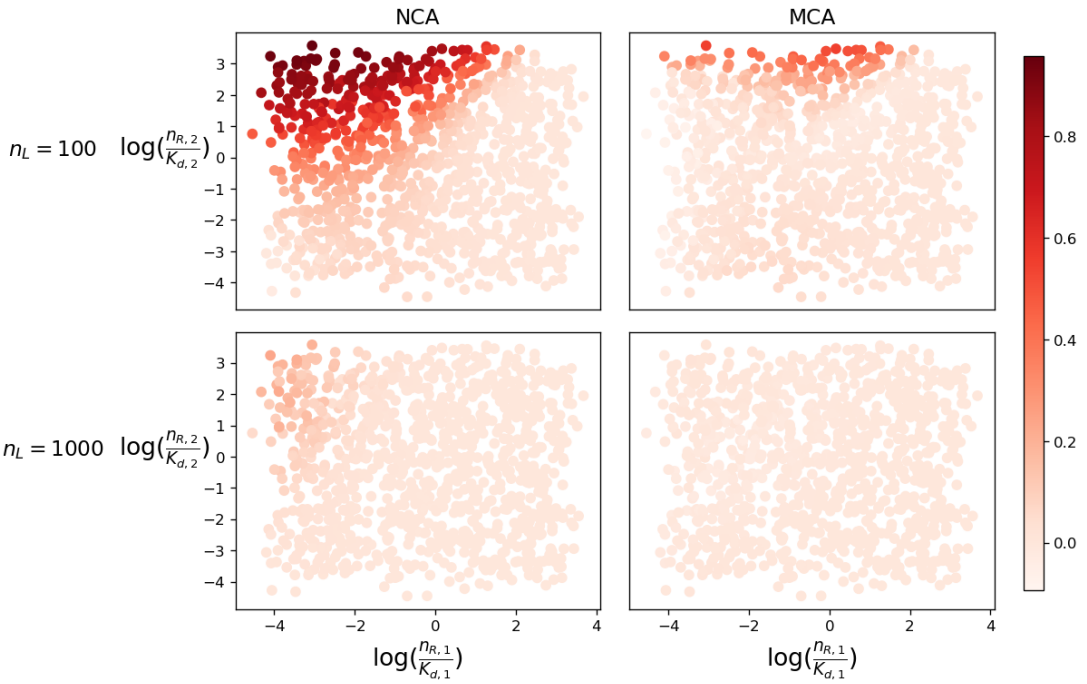

Figure 3.16: Relative difference $1 - \frac{T^j(N,\tau)}{T^{SIM}(N,\tau)}$, $j \in \{NCA, MCA\}$, between the mean time $T(N,\tau)$ computed through the NCA and MCA approaches, and the time computed through stochastic simulations, for $n_L \in \{10^2, 10^3\}$. In these scenarios, 10^3 parameter sets have been sampled by varying the number of receptors n_{R1} and n_{R2} between 100 and 400, $K_{d,1}$ and $K_{d,2}$ rates vary between 10^1 and 10^4, and setting $k_{r,1} = k_{r,2} = 10^{-3}$ s^{-1}. In these examples, $N = 10$ and $\tau = k_{r,1}^{-1}$, so that a complex is considered to be productive if it lasts for longer than its average lifetime.

This analysis of a third stochastic descriptor is an example of how the methodology in this chapter can be extended to other situations. It is expected that vari-

ations of the NCA and MCA approaches could be used to compute even further descriptors, or to compute similar descriptors in a higher dimensional scenario (for example with more than 2 receptor types competing for the common ligand). To this end, in the following sections, the EMA, NCA and MCA are defined and compared for a 3 receptor system.

3.6 Higher dimensional systems

In this section, a competition process in three variables is introduced, where the system in Figure 3.1 is extended by adding a third competing receptor type, R_3 as seen in Figure 3.17. For this larger system it is still possible to use the exact matrix methodology as described in Sections 3.4.1 and 3.5.1 to compute the two stochastic descriptors of interest in this chapter. The matrix forms and algorithms required for the EMA for the three dimensional system are defined in Sections 3.6.1 and 3.6.2. Due to the increased size of the state space caused by adding a third receptor type, the EMA becomes more computationally expensive for this three dimensional system and hence there is a greater requirement for the NCA and MCA, which are trivial to expand to a system of three receptors, and even to a system of N receptors. On the other hand, to expand the EMA to a system of more than three receptors would be complex, both in terms of the computational requirements and even to write down the forms of the matrices required in the algorithms used to compute the descriptors.

For the system depicted in Figure 3.17, the dimensionality can once again be reduced by considering the total number of receptors available in the system, $n_{R,1}$, $n_{R,2}$ and $n_{R,3}$ and the total number of ligands, n_L. As well as the conservation equations defined in Section 3.1, here the equation $n_{R,3} = R_3(t) + M_3(t)$ also applies and hence $R_3(t)$ can also be implicitly tracked, since $R_3(t) = n_{R,3} - M_3(t)$. Then, the process can be described as a multi-variate continuous-time Markov chain (CTMC) $\mathcal{Y} = \{\mathbf{Y}(t) = (M_1(t), M_2(t), M_3(t)) : t \geq 0\}$, with $M_1(t), M_2(t), M_3(t) \geq 0$ for all $t \geq 0$, and the process $\mathcal{Y}$ evolves over the state space $\mathcal{S}_{\mathcal{Y}} = \{(m_1, m_2, m_3) \in (\mathbb{N} \cup \{0\})^3 : \quad m_1 \leq n_{R,1}, \quad m_2 \leq n_{R,2}, \quad m_3 \leq n_{R,3}, \quad m_1 + m_2 + m_3 \leq n_L\}$, where m_1, m_2 and m_3 are the number of type 1, 2 and 3 complexes, respectively, at any given time.

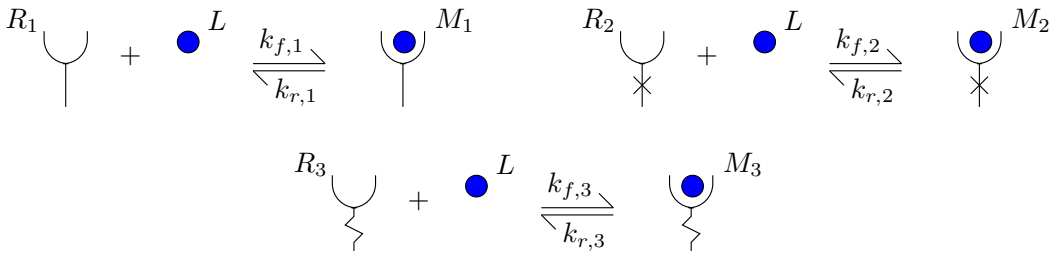

Figure 3.17: A depiction of the molecular reactions underlying the stochastic mathematical model for the formation of three different complexes with one shared ligand. Three different types of receptor molecule can bind, reversibly, with a shared ligand to form three different complex types.

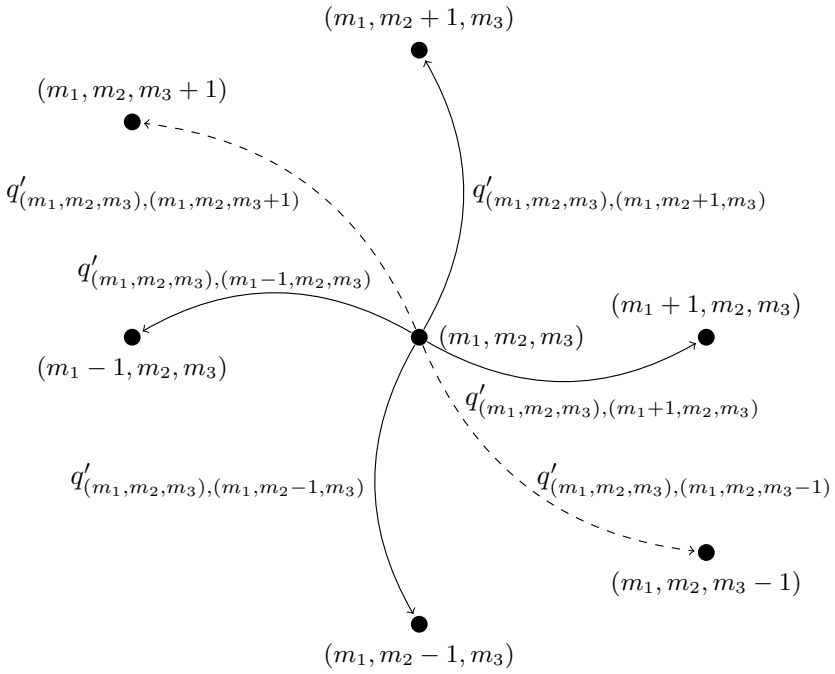

Figure 3.18: Transition diagram for the process $\mathcal{Y}$, showing the possible states which the process can move to from a general state (m_1, m_2, m_3) and the transition rates with which these state moves occur.

The dynamics of complex formation and dissociation are then represented by *jumps*, or transitions, between states in $\mathcal{S}_{\mathcal{Y}}$, $(m_1, m_2, m_3) \to (m_1', m_2', m_3')$. The transition diagram is shown in Figure 3.18 and the infinitesimal transition rate from state (m_1, m_2, m_3) to state (m_1', m_2', m_2'), by assuming mass action kinetics, is given by $q'_{(m_1,m_2,m_3),(m_1',m_2',m_3')} =$

$$\begin{cases} k_{f,1}(n_{R,1} - m_1)(n_L - m_1 - m_2 - m_3), & \text{if } (m_1', m_2', m_3') = (m_1 + 1, m_2, m_3), \\ k_{r,1}m_1, & \text{if } (m_1', m_2', m_3') = (m_1 - 1, m_2, m_3), \\ k_{f,2}(n_{R,2} - m_2)(n_L - m_1 - m_2 - m_3), & \text{if } (m_1', m_2', m_3') = (m_1, m_2 + 1, m_3), \\ k_{r,2}m_2, & \text{if } (m_1', m_2', m_3') = (m_1, m_2 - 1, m_3), \\ k_{f,3}(n_{R,3} - m_3)(n_L - m_1 - m_2 - m_3), & \text{if } (m_1', m_2', m_3') = (m_1, m_2, m_3 + 1), \\ k_{r,3}m_3, & \text{if } (m_1', m_2', m_3') = (m_1, m_2, m_3 - 1), \\ 0, & \text{otherwise.} \end{cases}$$

3.6.1 Steady state distribution

In this section, the steady state distribution for the process depicted in Figure 3.17 is analysed using an extension of the EMA introduced in Section 3.4.1. It is also shown how the NCA and MCA methods can be extended to a system of three dimensions, and again the methods are numerically compared. The steady state distribution for the model in Figure 3.17 is described in terms of the following probabilities

$$\pi^{EMA}_{(m_1,m_2,m_3)} = \lim_{t\to+\infty} \mathbb{P}((M_1(t), M_2(t), M_3(t)) = (m_1, m_2, m_3)), \quad (m_1, m_2, m_3) \in \mathcal{S}_{\mathcal{Y}},$$

which can be stored in a row vector $\boldsymbol{\pi}' = (\pi_{(m_1,m_2,m_3)},\ (m_1, m_2, m_3) \in \mathcal{S}_{\mathcal{Y}})$ for any given order of states in $\mathcal{S}_{\mathcal{Y}}$. These probabilities correspond to the number of complexes, of type 1, 2 and 3 respectively, found on the cellular surface at late times. To begin, the state space for the CTMC $\mathcal{Y}$ can be organised into levels as for the two-dimensional process $\mathcal{X}$. The state space is organised into levels $L'(k)$ where now $L'(k) = \{(m_1, m_2, m_3) : m_3 = k\}$ for $0 \leq k \leq N_3 = min(n_{R,3}, n_L)$ and $L'(0) \prec L'(1) \prec ... \prec L'(N_3)$. Then, one can define two further minimums, $h_1(k, r) = min(N_1, n_L - r - k) = min(n_{R,1}, n_L - r - k)$ and $h_2(k) = min(N_2, n_L - k) = min(n_{R,2}, n_L - k)$, where k and r will be omitted from the notation for h_1

and h_2 for simplicity. Having done this, the states within each level $L'(k)$ can be split into sub-levels $l(k;r)$ where $l(k;r) = \{(m_1, m_2, m_3) : m_2 = r, m_3 = k\}$ for $0 \leq r \leq h_2$ and $0 \leq k \leq N_3$. The states within a sub-level $l(k;r)$ can be ordered as $l(k;r) = \{(0,r,k),(1,r,k),\ldots,(h_1,r,k)\}$, and the state space contains,

$$\#\mathcal{S}_y = \sum_{k=0}^{N_3}\sum_{r=0}^{h_2}(h_1+1)$$

states, and can be expressed in terms of the levels $L'(k)$ as

$$\mathcal{S}_y = \bigcup_{k=0}^{N_3} L'(k),$$

where

$$L'(k) = \bigcup_{r=0}^{h_2} l(k;r).$$

Having ordered the states in this way, the infinitesimal generator $\boldsymbol{Q'}$ is once again tridiagonal by blocks where now each block $\boldsymbol{Q'}_{k,k'}$ has also a block tridiagonal or block diagonal structure. Matrices $\boldsymbol{Q'}_{k,k'}$ contain transition rates for transitions between states in level $L'(k)$ and level $L'(k')$. The sub-blocks within each block contain the transition rates for the transitions between sub-level $l(k;r)$ and $l(k';r')$. The infinitesimal generator $\boldsymbol{Q'}$ is given by,

$$\boldsymbol{Q'} = \begin{pmatrix} \boldsymbol{Q'}_{0,0} & \boldsymbol{Q'}_{0,1} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \boldsymbol{Q'}_{1,0} & \boldsymbol{Q'}_{1,1} & \boldsymbol{Q'}_{1,2} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q'}_{2,1} & \boldsymbol{Q'}_{2,2} & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{Q'}_{N_3-1,N_3-1} & \boldsymbol{Q'}_{N_3-1,N_3} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{Q'}_{N_3,N_3-1} & \boldsymbol{Q'}_{N_3,N_3} \end{pmatrix}.$$

Each level $L'(k)$ contains $J'(k) = \#L'(k) = h_2 + 1$ states, so that each sub-matrix $\boldsymbol{Q'}_{k,k'}$ has dimensions $J'(k) \times J'(k')$. Sub-matrices $\boldsymbol{Q'}_{k,k'}$ are given as follows:

- For $0 \leq k \leq N_3$,

$$\boldsymbol{Q'}_{k,k} = \begin{pmatrix} \boldsymbol{B}_{0,0}^{k,k} & \boldsymbol{B}_{0,1}^{k,k} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \boldsymbol{B}_{1,0}^{k,k} & \boldsymbol{B}_{1,1}^{k,k} & \boldsymbol{B}_{1,2}^{k,k} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B}_{2,1}^{k,k} & \boldsymbol{B}_{2,2}^{k,k} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{B}_{h_2-1,h_2-1}^{k,k} & \boldsymbol{B}_{h_2-1,h_2}^{k,k} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{B}_{h_2,h_2-1}^{k,k} & \boldsymbol{B}_{h_2,h_2}^{k,k} \end{pmatrix}.$$

- For $0 \leq k \leq N_3 - 1$,

$$\boldsymbol{Q'}_{k,k+1} = \begin{pmatrix} \boldsymbol{B}_{0,0}^{k,k+1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B}_{1,1}^{k,k+1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{B}_{2,2}^{k,k+1} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{B}_{h_2-1,h_2-1}^{k,k+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \boldsymbol{B}_{h_2,h_2}^{k,k+1} \end{pmatrix}.$$

- For $1 \leq k \leq N_3$,

$$\boldsymbol{Q'}_{k,k-1} = \begin{pmatrix} \boldsymbol{B}_{0,0}^{k,k-1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B}_{1,1}^{k,k-1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{B}_{2,2}^{k,k-1} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{B}_{h_2-1,h_2-1}^{k,k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \boldsymbol{B}_{h_2,h_2}^{k,k-1} \end{pmatrix}.$$

The sub-levels $l(k;r)$ each contain $j(k;r) = \#l(k;r) = h_1 + 1$ states, so that each sub-matrix $\boldsymbol{B}^{k,k'}_{r,r'}$ has dimensions $j(k;r) \times j(k';r')$. Sub-matrices $\boldsymbol{B}^{k,k'}_{r,r'}$ are given as follows:

- For $0 \leq r \leq h_2$ and $0 \leq k \leq N_3$,

$$(\boldsymbol{B}^{k,k}_{r,r})_{i,j} = \begin{cases} k_{f,1}(n_{R,1} - i)(n_L - i - r - k), & \text{if } j = i + 1, \\ k_{r,1} i, & \text{if } j = i - 1, \\ -\Delta_{(i,r,k)} & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

  where $0 \leq i \leq j(k;r)$, $0 \leq j \leq j(k;r)$ and $\Delta_{(i,r,k)} = k_{f,1}(n_{R,1} - i)(n_L - i - r - k) + k_{r,1} i + k_{f,2}(n_{R,2} - r)(n_L - i - r - k) + k_{r,2} r + k_{f,3}(n_{R,3} - k)(n_L - i - r - k) + k_{r,3} k$.

- For $0 \leq r \leq h_2 - 1$ and $0 \leq k \leq N_3$,

$$(\boldsymbol{B}^{k,k}_{r,r+1})_{i,j} = \begin{cases} k_{f,2}(n_{R,2} - r)(n_L - i - r - k), & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

  where $0 \leq i \leq j(k;r)$ and $0 \leq j \leq j(k;r+1)$.

- For $1 \leq r \leq h_2$ and $0 \leq k \leq N_3$,

$$(\boldsymbol{B}^{k,k}_{r,r-1})_{i,j} = \begin{cases} k_{r,2} r, & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

  where $0 \leq i \leq j(k;r)$ and $0 \leq j \leq j(k;r-1)$.

- For $0 \leq r \leq h_2$ and $0 \leq k \leq N_3 - 1$,

$$(\boldsymbol{B}^{k,k+1}_{r,r})_{i,j} = \begin{cases} k_{f,3}(n_{R,3} - k)(n_L - i - r - k), & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

  where $0 \leq i \leq j(k;r)$ and $0 \leq j \leq j(k+1;r)$.

- For $0 \leq r \leq h_2$ and $1 \leq k \leq N_3$,

$$(\boldsymbol{B}_{r,r}^{k,k-1})_{i,j} = \begin{cases} k_{r,3}k, & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

  where $0 \leq i \leq j(k;r)$ and $0 \leq j \leq j(k-1;r)$.

With the infinitesimal generator $\boldsymbol{Q'}$ in the tridiagonal by blocks form, the steady state distribution for this three-dimensional process can be computed using an algorithm similar to Algorithm 6. The matrices $\boldsymbol{Q}_{k,k}$ would be replaced by $\boldsymbol{Q'}_{k,k}$, $\boldsymbol{\pi}$ by $\boldsymbol{\pi'}$ and N_2 by N_3 in the algorithm. Having obtained the steady state probabilities, the mean number of complexes in steady state can be computed as

$$\mathbb{E}^{EMA'}[M_1^*] = \sum_{m_3=0}^{N_3} \sum_{m_2=0}^{min(n_{R,2},n_L-m_3)} \sum_{m_1=0}^{min(n_{R,1},n_L-m_2-m_3)} m_1 \cdot \pi_{(m_1,m_2,m_3)}^{EMA},$$

$$\mathbb{E}^{EMA'}[M_2^*] = \sum_{m_3=0}^{N_3} \sum_{m_2=0}^{min(n_{R,2},n_L-m_3)} \sum_{m_1=0}^{min(n_{R,1},n_L-m_2-m_3)} m_2 \cdot \pi_{(m_1,m_2,m_3)}^{EMA},$$

$$\mathbb{E}^{EMA'}[M_3^*] = \sum_{m_3=0}^{N_3} \sum_{m_2=0}^{min(n_{R,2},n_L-m_3)} \sum_{m_1=0}^{min(n_{R,1},n_L-m_2-m_3)} m_3 \cdot \pi_{(m_1,m_2,m_3)}^{EMA}.$$

From the matrix forms presented in this section, one can easily see the increase in complexity between the two receptor system and the three receptor system. In order to extend the method to even higher dimensions, the EMA would become infeasible since it would involve matrices with several sub-blocks and the computational expense would be too great. On the other hand, to extend the MCA to a system of three receptors is trivial as it only involves analysing a third CTMC $\mathcal{X}_3 = \{M_3(t) : t \geq 0\}$ defined over the state space $\mathcal{S}_{\mathcal{X}_3} = \{m_3 \in \mathbb{N}_0 : m_3 \leq N_3\}$, which is identical in form to $\mathcal{X}_1$ and $\mathcal{X}_2$ as introduced in Section 3.4.2. The steady state probabilities for this third birth-and-death process are

$$\pi_{m_3} = \lim_{t \to +\infty} \mathbb{P}(M_3(t) = m_3) = \binom{n_{R,3}}{m_3} \left(\frac{k_{f,3}n_L}{k_{r,3} + k_{f,3}n_L}\right)^{m_3} \left(\frac{k_{r,3}}{k_{r,3} + k_{f,3}n_L}\right)^{n_{R,3}-m_3},$$

for $0 \leq m_3 \leq N_3$, which again follow a binomial distribution. The expected

values of complexes in steady state can also be determined as

$$\mathbb{E}^{NCA}[M_3^*] = \sum_{m_3=0}^{N_3} m_3 \, \pi_{m_3} \quad = \quad \frac{k_{f,3} n_{R,3} n_L}{k_{r,3} + k_{f,3} n_L}.$$

In the limit $n_L \to +\infty$, the NCA steady state distribution of the process $\mathcal{Y}$ is given by

$$\pi^{NCA}_{(m_1,m_2,m_3)} = \pi_{m_1} \times \pi_{m_2} \times \pi_{m_3} \ , \quad (m_1, m_2, m_3) \in \mathcal{S}_{\mathcal{Y}} \ . \tag{3.31}$$

Finally, in order to compute the MCA steady state distribution for the three receptor process, one would use an adapted version of Algorithm 7, where the probabilities $\pi_{(m_1,m_2)}$ would be replaced throughout with $\pi_{(m_1,m_2,m_3)}$, $\mathcal{S}_{\mathcal{X}}$ replaced with $\mathcal{S}_{\mathcal{Y}}$ and the equation on line 9 of the algorithm would be replaced with Equation (3.31). Computation of the steady state probabilities and expected numbers of complexes in steady state for the one-dimensional process $\mathcal{X}_3$ would also feature in all the obvious places, for example, the while condition would also have the term $|\mathbb{E}^{NCA,(i)}[M_3^*] - \mathbb{E}^{NCA,(i-1)}[M_3^*]| > \varepsilon$ and the computation of $n_L^{(i)}$ would become $n_L^{(i)} = (n_L - \mathbb{E}^{NCA,(i-1)}[M_1^*] - \mathbb{E}^{NCA,(i-1)}[M_2^*] - \mathbb{E}^{NCA,(i-1)}[M_3^*])\alpha + n_L^{(i-1)}(1-\alpha)$. It is clear that this algorithm could be easily generalised to analyse the steady state of a system of N receptors, and this generalisation will not come with any notable increase in the computational cost.

In order to verify that the MCA is still a good alternative to the EMA for the computation of the steady state distribution for a three receptor system, in Figure 3.19, a similar scatter plot of the HD is seen to that in Figure 3.10. Here a third competition strength parameter, $\frac{n_{R,3}}{K_{d,3}}$, is introduced and each of the three competition strength parameters are plotted against each other in each row of the figure. The ranges for the receptor numbers are lower here than in the 2D case since, even by increasing the system by only one receptor type, the number of states in the state space becomes much larger and the EMA becomes computationally expensive. In the top row, $n_L = 30$ and in the bottom row $n_L = 100$ so that one can observe once again, that with increasing ligand numbers (and hence less competition), the approximation improves. The same pattern is

also seen within each subplot as in Figure 3.10, where for higher values of each of the competition parameters, the HD is greater.

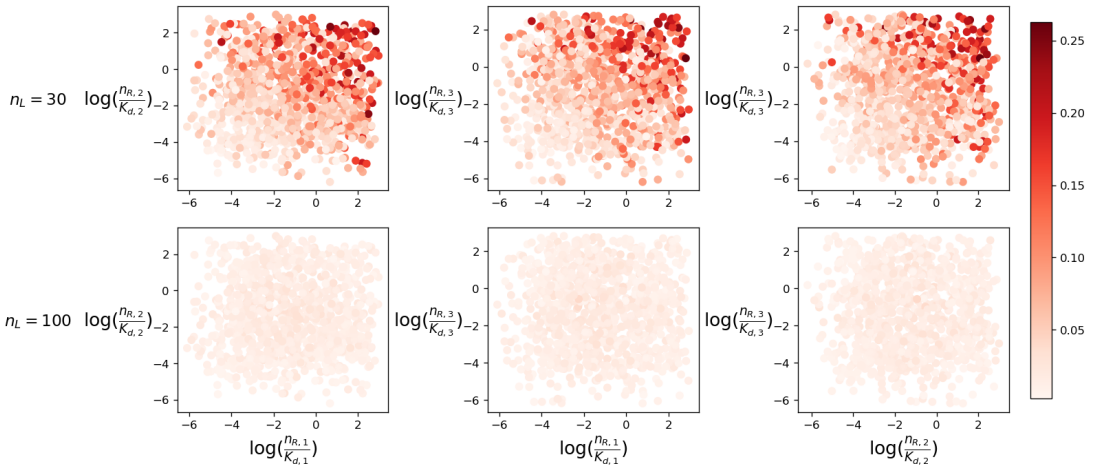

Figure 3.19: HD between the steady state distributions $\{\pi^{EMA}_{(m_1,m_2,m_3)}\}_{(m_1,m_2,m_3)\in\mathcal{S}_y}$ and $\{\pi^{MCA}_{(m_1,m_2,m_3)}\}_{(m_1,m_2,m_3)\in\mathcal{S}_y}$ plotted for sampled values $n_{R,j} \sim Unif(2, 20)$ and $K_{d,j} = 10^x$ with $x \sim Unif(0, 3)$, for $j \in \{1, 2, 3\}$, and for different numbers of ligand $n_L \in \{30, 100\}$. The threshold parameter $\varepsilon = 10^{-5}$ is used for the MCA.

Here it is feasible to compare once again the EMA with the MCA via the HD, however if the system were to increase further (say adding a fourth receptor type), the EMA would become intractable. In this situation one could verify the accuracy of the MCA by comparing the expected number of complexes of each type in steady state with those obtained via stochastic simulations. An example of such a comparison is seen in Figure 3.20 where the MCA is compared with Gillespie simulations for a process with four receptor types. Here the HD is not considered as, in order to accurately capture the whole steady state distribution using Gillespie simulations, one would need to simulate the process a very large number of times. Instead the accuracy of the MCA is assessed by considering the absolute difference between the expected values of each receptor type in steady state as computed via the MCA and via an average of 10^6 Gillespie simulations. As expected, it is again seen that with increasing ligand numbers, the accuracy of the MCA increases. Strikingly however, even for the lower value of $n_L = 10^2$,

the maximal value of the absolute difference is approximately 4, *i.e.* the expected values are captured by the MCA roughly to integer value.

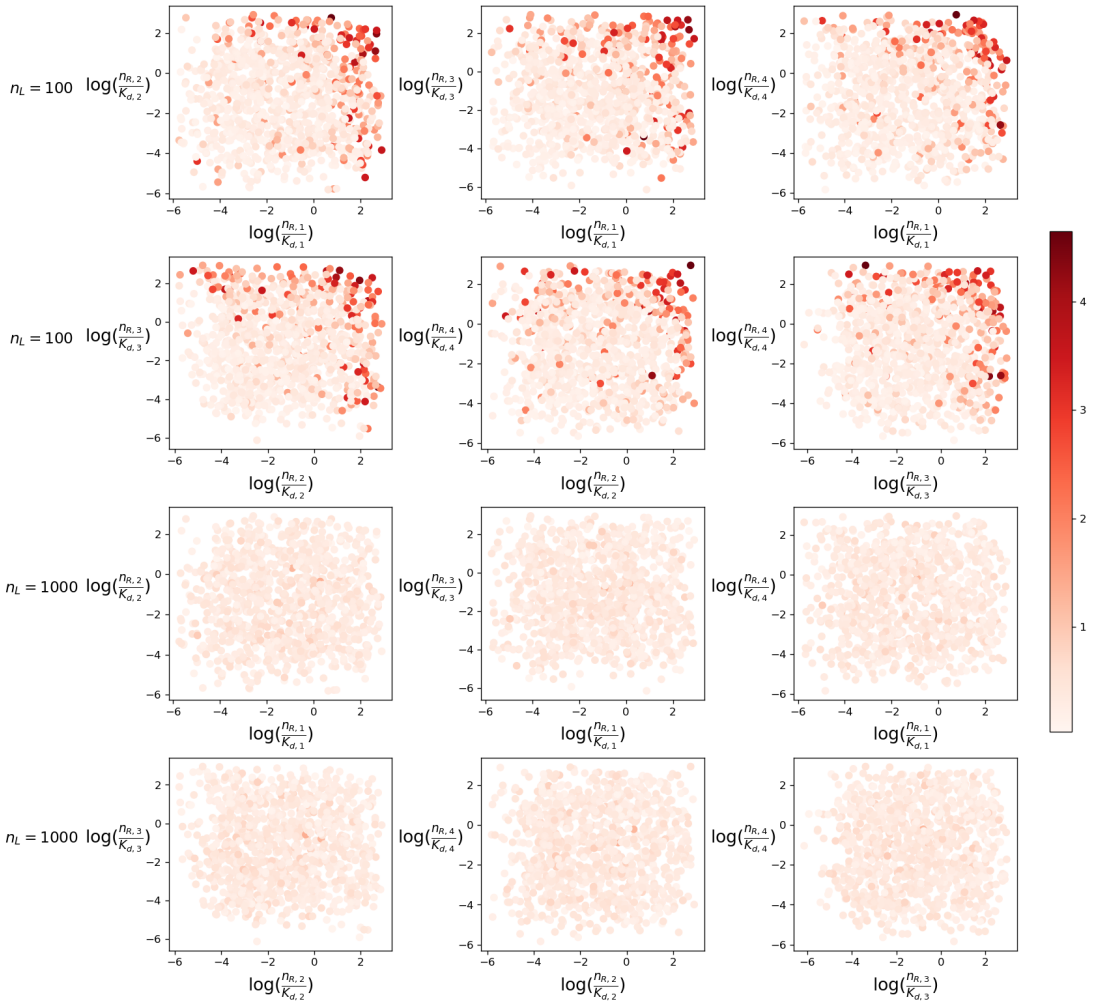

Figure 3.20: Absolute difference $\sum_{i=1}^{4} |\mathbb{E}^{MCA}[M_i^*] - \mathbb{E}^{SIM}[M_i^*]|$ between the mean number of complexes in steady state computed through Gillespie simulations and the MCA approach, for $n_L \in \{10^2, 10^3\}$. 10^3 parameter sets are sampled by considering the number of receptors $n_{R,j} \sim Unif(20, 200)$ and $K_{d,j} = 10^x$ with $x \sim Unif(1, 4)$, $j \in \{1, 2, 3, 4\}$. The unbinding rates are fixed at $k_{r,j} = 10^{-3}$ s^{-1} for $j \in \{1, 2, 3, 4\}$.

It has been shown in this section that the MCA for the steady state distribution can be easily extended to a higher dimensional system, where it is still an accurate approximation. In the next section, the NCA and MCA for the time

scales of complex formation are also extended to a higher dimensional system.

3.6.2 Time scales of complex formation

In this section, it is shown how the time scales of complex formation for the process depicted in Figure 3.17 can be analysed using an extension of the EMA introduced in Section 3.5.1. It is also shown how the NCA and MCA methods can be extended to a system of three dimensions for this descriptor, and again the methods are numerically compared. For the three receptor system, the expected time to reach N complexes of type 3 is now computed. For the multi-variate process $\mathcal{Y}$, $T_{(m_1,m_2,m_3)}(N)$ is defined as the time to reach N complexes of type 3 given the initial state $(m_1, m_2, m_3) \in \mathcal{S}_{\mathcal{Y}}$, *i.e.*

$$T_{(m_1,m_2,m_3)}(N) = inf\{t \geq 0 : \ M_3(t) = N \mid (M_1(0), M_2(0), M_3(0)) = (m_1, m_2, m_3)\},$$

for $(m_1, m_2, m_3) \in \mathcal{S}_{\mathcal{Y}}$. Here N can be any arbitrary value $M_3(0) < N \leq N_3$, since $T_{(M_1(0),M_2(0),M_3(0))}(M_3(0)) \equiv 0$. As in Section 3.5.1, moments of order l of the time to reach N complexes of type 3 can be computed, using the Laplace-Stieltjes transform of $T_{(m_1,m_2,m_3)}(N)$, given by

$$\phi^N_{(m_1,m_2,m_3)}(z) \quad = \quad \mathbb{E}[e^{-zT_{(m_1,m_2,m_3)}(N)}] \ , \quad \mathbb{R}(z) \geq 0 \ .$$

Using a first-step argument, one arrives at the system of linear equations

$$\begin{aligned}
\phi_{(m_1,m_2,m_3)}(z) = & \frac{k_{f,1}(n_{R,1} - m_1)(n_L - m_1 - m_2 - m_3)}{z + \Delta_{(m_1,m_2,m_3)}} \phi_{(m_1+1,m_2,m_3)}(z) \\
& + \frac{k_{r,1} m_1}{z + \Delta_{(m_1,m_2,m_3)}} \phi_{(m_1-1,m_2,m_3)}(z) \\
& + \frac{k_{f,2}(n_{R,2} - m_2)(n_L - m_1 - m_2 - m_3)}{z + \Delta_{(m_1,m_2,m_3)}} \phi_{(m_1,m_2+1,m_3)}(z) \\
& + \frac{k_{r,2} m_2}{z + \Delta_{(m_1,m_2,m_3)}} \phi_{(m_1,m_2-1,m_3)}(z) \\
& + \frac{k_{f,3}(n_{R,3} - m_3)(n_L - m_1 - m_2 - m_3)}{z + \Delta_{(m_1,m_2,m_3)}} \phi_{(m_1,m_2,m_3+1)}(z)
\end{aligned}$$

$$+\frac{k_{r,3}m_3}{z+\Delta_{(m_1,m_2,m_3)}}\phi_{(m_1,m_2,m_3-1)}(z),$$

where now $\Delta_{(m_1,m_2,m_3)} = k_{f,1}(n_{R,1}-m_1)(n_L-m_1-m_2-m_3)+k_{r,1}m_1+k_{f,2}(n_{R,2}-m_2)(n_L-m_1-m_2-m_3)+k_{r,2}m_2+k_{f,3}(n_{R,3}-m_3)(n_L-m_1-m_2-m_3)+k_{r,3}m_3$. Following the same methodology as in Section 3.5.1, The lth order moment of $T_{(m_1,m_2,m_3)}$ can then be found by differentiating $\phi_{(m_1,m_2,m_3)}(z)$ l times and multiplying by $(-1)^l$, resulting in the system of equations,

$$\begin{aligned}\Delta_{(m_1,m_2,m_3)}\mathbb{E}[T^l_{(m_1,m_2,m_3)}] &= k_{f,1}(n_{R,1}-m_1)(n_L-m_1-m_2-m_3)\mathbb{E}[T^l_{(m_1+1,m_2,m_3)}]\\ &+ k_{r,1}m_1\mathbb{E}[T^l_{(m_1-1,m_2,m_3)}]\\ &+ k_{f,2}(n_{R,2}-m_2)(n_L-m_1-m_2-m_3)\mathbb{E}[T^l_{(m_1,m_2+1,m_3)}]\\ &+ k_{r,2}m_2\mathbb{E}[T^l_{(m_1,m_2-1,m_3)}]\\ &+ k_{f,3}(n_{R,3}-m_3)(n_L-m_1-m_2-m_3)\mathbb{E}[T^l_{(m_1,m_2,m_3+1)}]\\ &+ k_{r,3}m_3\mathbb{E}[T^l_{(m_1,m_2,m_3-1)}]\\ &+ l\ \mathbb{E}[T^{l-1}_{(m_1,m_2,m_3)}]. \end{aligned} \tag{3.32}$$

Moreover, by arranging the states in $\mathcal{S}_{\mathcal{Y}}$ into levels as in Section 3.6.1, each equation in the system (3.32) corresponds to an initial state $(m_1,m_2,m_3)\in\cup_{k=0}^{N-1}L'(k)$, so that one can rewrite Equation (3.32) in matrix form as follows

$$\boldsymbol{n}^{(l)} = \boldsymbol{A'}\boldsymbol{n}^{(l)} + \boldsymbol{c}^{(l)} \tag{3.33}$$

where

$$\boldsymbol{n}^{(l)} = \begin{pmatrix} \boldsymbol{n}_0^{(l)} \\ \boldsymbol{n}_1^{(l)} \\ \vdots \\ \boldsymbol{n}_{N-1}^{(l)} \end{pmatrix}, \quad \boldsymbol{n}_k^{(l)} = \begin{pmatrix} \boldsymbol{n}_{0,k}^{(l)} \\ \boldsymbol{n}_{1,k}^{(l)} \\ \vdots \\ \boldsymbol{n}_{h_2,k}^{(l)} \end{pmatrix}, \quad \boldsymbol{n}_{r,k}^{(l)} = \begin{pmatrix} \mathbb{E}[T^l_{(0,r,k)}] \\ \mathbb{E}[T^l_{(1,r,k)}] \\ \vdots \\ \mathbb{E}[T^l_{(h_1,r,k)}] \end{pmatrix},$$

for $0 \leq k \leq N-1$ and $0 \leq r \leq h_2$. Vector $\boldsymbol{c}^{(l)}$ is also organised in sub-vectors as

$$\boldsymbol{c}^{(l)} = \begin{pmatrix} \boldsymbol{c}_0^{(l)} \\ \boldsymbol{c}_1^{(l)} \\ \vdots \\ \boldsymbol{c}_{N-1}^{(l)} \end{pmatrix}, \quad \boldsymbol{c}_k^{(l)} = \begin{pmatrix} \boldsymbol{c}_{0,k}^{(l)} \\ \boldsymbol{c}_{1,k}^{(l)} \\ \vdots \\ \boldsymbol{c}_{h_2,k}^{(l)} \end{pmatrix},$$

which are obtained from the $(l-1)$th order moments, as

$$(\boldsymbol{c}_{r,k}^{(l)})_{(m_1,m_2,m_3)} = \frac{l}{\Delta_{(m_1,m_2,m_3)}} (\boldsymbol{n}_{r,k}^{(l-1)})_{(m_1,m_2,m_3)},$$

for all $(m_1, m_2, m_3) \in l(k; r)$ and $0 \leq r \leq h_1$. The matrix $\boldsymbol{A'}$ is given by

$$\boldsymbol{A'} = \begin{pmatrix} \boldsymbol{A'}_{0,0} & \boldsymbol{A'}_{0,1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \boldsymbol{A'}_{1,0} & \boldsymbol{A'}_{1,1} & \boldsymbol{A'}_{1,2} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A'}_{2,1} & \boldsymbol{A'}_{2,2} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{A'}_{N-2,N-2} & \boldsymbol{A'}_{N-2,N-1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{A'}_{N-1,N-2} & \boldsymbol{A'}_{N-1,N-1} \end{pmatrix}.$$

Each level $L'(k)$ contains $J'(k) = \#L'(k) = h_2 + 1$ states, so that each sub-matrix $\boldsymbol{A'}_{k,k'}$ has dimensions $J'(k) \times J'(k')$. Sub-matrices $\boldsymbol{A'}_{k,k'}$ are given as follows:

- For $0 \leq k \leq N-1$,

$$\boldsymbol{A'}_{k,k} = \begin{pmatrix} \boldsymbol{D}_{0,0}^{k,k} & \boldsymbol{D}_{0,1}^{k,k} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \boldsymbol{D}_{1,0}^{k,k} & \boldsymbol{D}_{1,1}^{k,k} & \boldsymbol{D}_{1,2}^{k,k} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D}_{2,1}^{k,k} & \boldsymbol{D}_{2,2}^{k,k} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{D}_{h_2-1,h_2-1}^{k,k} & \boldsymbol{D}_{h_2-1,h_2}^{k,k} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{D}_{h_2,h_2-1}^{k,k} & \boldsymbol{D}_{h_2,h_2}^{k,k} \end{pmatrix}.$$

- For $0 \leq k \leq N-2$,

$$\boldsymbol{A'}_{k,k+1} = \begin{pmatrix} \boldsymbol{D}_{0,0}^{k,k+1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D}_{1,1}^{k,k+1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{D}_{2,2}^{k,k+1} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{D}_{h_2-1,h_2-1}^{k,k+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \boldsymbol{D}_{h_2,h_2}^{k,k+1} \end{pmatrix}.$$

- For $1 \leq k \leq N-1$,

$$\boldsymbol{A'}_{k,k-1} = \begin{pmatrix} \boldsymbol{D}_{0,0}^{k,k-1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D}_{1,1}^{k,k-1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{D}_{2,2}^{k,k-1} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{D}_{h_2-1,h_2-1}^{k,k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \boldsymbol{D}_{h_2,h_2}^{k,k-1} \end{pmatrix}.$$

The sub-levels $l(k;r)$ each contain $j(k;r) = \#l(k;r) = h_1 + 1$ states, so that each sub-matrix $\boldsymbol{D}_{r,r'}^{k,k'}$ has dimensions $j(k;r) \times j(k';r')$. Sub-matrices $\boldsymbol{D}_{r,r'}^{k,k'}$ are given as follows:

- For $0 \leq r \leq h_2$ and $0 \leq k \leq N-1$,

$$(\boldsymbol{D}_{r,r}^{k,k})_{i,j} = \begin{cases} \frac{k_{f,1}(n_{R,1}-i)(n_L-i-r-k)}{\Delta_{(i,r,k)}}, & \text{if } j = i+1, \\ \frac{k_{r,1} i}{\Delta_{(i,r,k)}}, & \text{if } j = i-1, \\ 0, & \text{otherwise,} \end{cases}$$

where $0 \leq i \leq j(k;r)$, $0 \leq j \leq j(k;r)$.

- For $0 \leq r \leq h_2 - 1$ and $0 \leq k \leq N-1$,

$$(\boldsymbol{D}^{k,k}_{r,r+1})_{i,j} = \begin{cases} \frac{k_{f,2}(n_{R,2}-r)(n_L-i-r-k)}{\Delta_{(i,r,k)}}, & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

where $0 \leq i \leq j(k;r)$ and $0 \leq j \leq j(k;r+1)$.

- For $1 \leq r \leq h_2$ and $0 \leq k \leq N-1$,

$$(\boldsymbol{D}^{k,k}_{r,r-1})_{i,j} = \begin{cases} \frac{k_{r,2}r}{\Delta_{(i,r,k)}}, & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

where $0 \leq i \leq j(k;r)$ and $0 \leq j \leq j(k;r-1)$.

- For $0 \leq r \leq h_2$ and $0 \leq k \leq N-2$,

$$(\boldsymbol{D}^{k,k+1}_{r,r})_{i,j} = \begin{cases} \frac{k_{f,3}(n_{R,3}-k)(n_L-i-r-k)}{\Delta_{(i,r,k)}}, & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

where $0 \leq i \leq j(k;r)$ and $0 \leq j \leq j(k+1;r)$.

- For $0 \leq r \leq h_2$ and $1 \leq k \leq N-1$,

$$(\boldsymbol{D}^{k,k-1}_{r,r})_{i,j} = \begin{cases} \frac{k_{r,3}k}{\Delta_{(i,r,k)}}, & \text{if } j = i, \\ 0, & \text{otherwise,} \end{cases}$$

where $0 \leq i \leq j(k;r)$ and $0 \leq j \leq j(k-1;r)$.

With the matrices in Equation (3.33) arranged in this way, the equation can then be solved efficiently using a similiar algorithm to Algorithm 8 to obtain the lth order moments of the random variable $T_{(m_1,m_2,m_3)}$. In the algorithm, the matrix blocks $\boldsymbol{A}$ would be replaced by $\boldsymbol{A'}$, $\boldsymbol{m}$ by $\boldsymbol{n}$ and $\boldsymbol{b}$ by $\boldsymbol{c}$. Similarly to the computation of the steady state distribution for the three receptor system, the computation for the time scales of complex formation has also increased

in complexity with the added receptor. To utilise the NCA for this stochastic descriptor however, one could use the same methodology as in Section 3.5.2, where now the expected time to reach N complexes of type 3, *i.e.* $T_{m_3}(N) = inf\{t \geq 0 : M_3(t) = N | M_3(0) = m_3\}$ is computed. Clearly the only change here is to replace λ_{m_2} and μ_{m_2} with $\lambda_{m_3} = k_{f,3}(n_{R,3} - m_3)(n_L - m_3)$ and $\mu_{m_3} = k_{r,3}m_3$ in Equation (3.27) where now the notation for the expected time is $\mathbb{E}^{NCA}[T_{(0,0,0)}(N)]$, since a third receptor type has been added, and initially it is assumed that no complexes of this type are present. Finally, to extend the MCA to this system, one would replace Equation (3.28) with

$$n_L^{**} = n_L - \mathbb{E}[M_1^*] - \mathbb{E}[M_2^*] \,, \tag{3.34}$$

i.e. subtracting from the initial number of ligands the expected number of type 1 and type 2 complexes in steady state. Equation (3.27) would then be used to compute $\mathbb{E}^{MCA}[T_{(0,0,0)}(N)]$ but using n_L^{**} in the calculation of the birth rates instead of n_L and with the alterations as described for the NCA in the paragraph above.

Extending the system from two receptor types to three receptor types will clearly have no effect on the computation time of the NCA for the expected time to reach N complexes of type 3. The only potential increase in computation time for the MCA is due to the increase in computation time when evaluating $\mathbb{E}[M_1^*]$ and $\mathbb{E}[M_2^*]$ to use in Equation (3.34). However, one could use the MCA for the steady state distribution to compute these expected values and then the increase in the computation time would not be significant.

The time saved when using either the NCA or MCA to compute this descriptor, as opposed to the EMA is even greater than for the two receptor system, but it remains to validate the accuracy of the approximations for this system. Hence in Figures 3.21 and 3.22 the relative difference between the times computed via the EMA and NCA and the EMA and MCA are plotted, respectively. Here the value of N is chosen as the expected number of complexes of type 3 in steady state, *i.e.* $N = \mathbb{E}[M_3^*]$ computed via the EMA. It can be seen that these two figures are very similar in trend to Figure 3.14, *i.e.* the approximations are best for low competition areas of the parameter space and when n_L is greatest.

Again it is seen that in general, the NCA underestimates the time and the MCA overestimates the time, as expected.

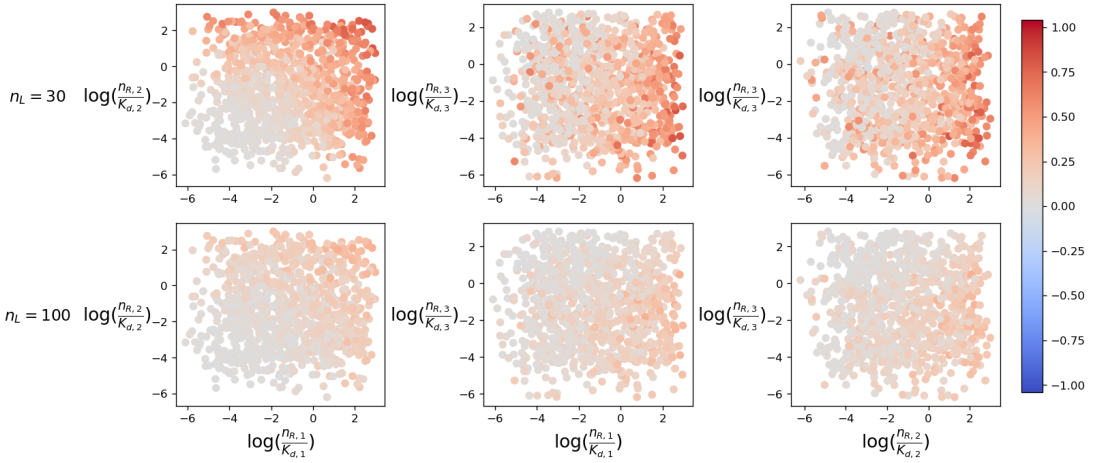

Figure 3.21: Relative difference $1 - \frac{\mathbb{E}^{NCA}[T_{(0,0,0)}(N)]}{\mathbb{E}^{EMA}[T_{(0,0,0)}(N)]}$ between the EMA and NCA mean times to reach N complexes of type 3, plotted for the 10^3 sampled parameter values in Figure 3.19, and for $n_L \in \{30, 100\}$.

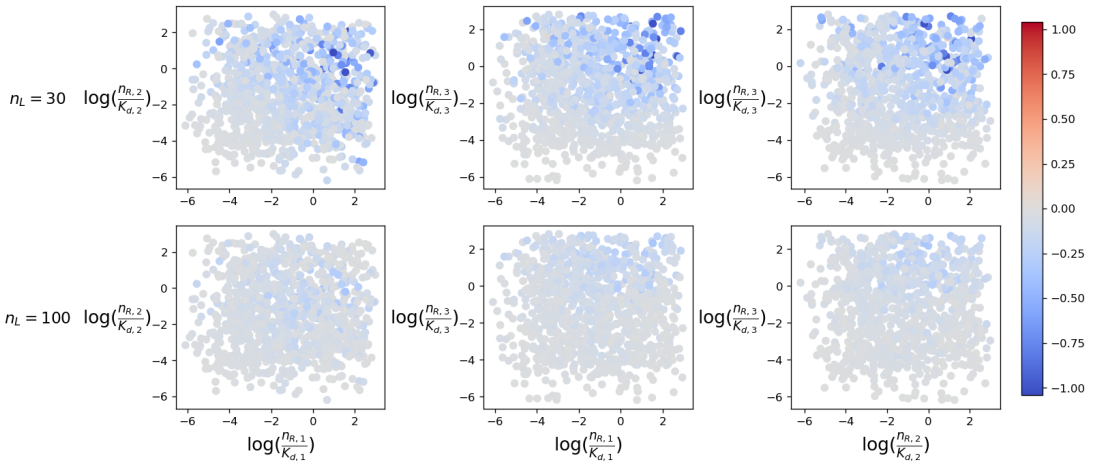

Figure 3.22: Relative difference $1 - \frac{\mathbb{E}^{MCA}[T_{(0,0,0)}(N)]}{\mathbb{E}^{EMA}[T_{(0,0,0)}(N)]}$ between the EMA and MCA mean times to reach N complexes of type 3, plotted for the 10^3 sampled parameter values in Figure 3.19, and for $n_L \in \{30, 100\}$.

As in the case of the steady state distribution, it is interesting to see how the approximations fare for a higher dimensional system, for example for the

four receptor system as discussed in Section 3.6.1. Again, the EMA would be too complex and computationally challenging to use to compute this stochastic descriptor in four dimensions and hence, stochastic simulations are again used to compare to the NCA and MCA methods. In Figures 3.23 and 3.24 the NCA and MCA (respectively) approaches to compute the expected times to reach N complexes of type 4 are compared with the same times computed via an average of 10^3 Gillespie simulations. Similarly to the three receptor case, the value of N here is chosen as the expected number of complexes of type 4 in steady state, *i.e.* $N = \mathbb{E}[M_4^*]$ computed now via the MCA, since the EMA is analytically intractable in this case.

In Figures 3.23 and 3.24 it can be seen that there are fairly large relative differences in the times for the lower number of ligands $n_L = 100$. In this four receptor case $n_L < n_{R,1} + n_{R,2} + n_{R,3} + n_{R,4}$ in almost all of the parameter sets sampled given that $n_{R,j} \sim Unif(20, 200)$ for $j \in \{1, 2, 3, 4\}$ and hence the competition here is very high so this result is expected. For $n_L = 1000$, many more of the relative differences are closer to zero for both the NCA and MCA. It is expected however, that as the size of the system increases (*i.e.* with adding more competing receptors), both approximations for the expected times to reach N complexes of a single type, will worsen. In the case of the NCA, this is because more receptor types means more competition and hence the assumption of no competition becomes less and less valid. In the case of the MCA, the reduction in accuracy is caused by the computation of the effective number of ligands, where this value is computed by subtracting from n_L, the expected number of complexes of each type (other than the type that the expected time to reach N of, is being computed) in steady state. It is likely that, when there are multiple receptor varieties, potentially all with different rates of binding and unbinding the ligand, the dynamics of formation of at least one receptor type may be slow enough that it is not reasonable to consider subtracting from the beginning of the calculation, the expected number of this complex type in steady state. Although the approximations generally decrease in accuracy with an increasing number of receptor types, for a large enough number of ligands n_L, both the NCA and MCA are still very accurate. For example, for the four receptor system, using $n_L = 10^4$ and the receptor numbers as varied in Figures 3.23 and 3.24, 85% of

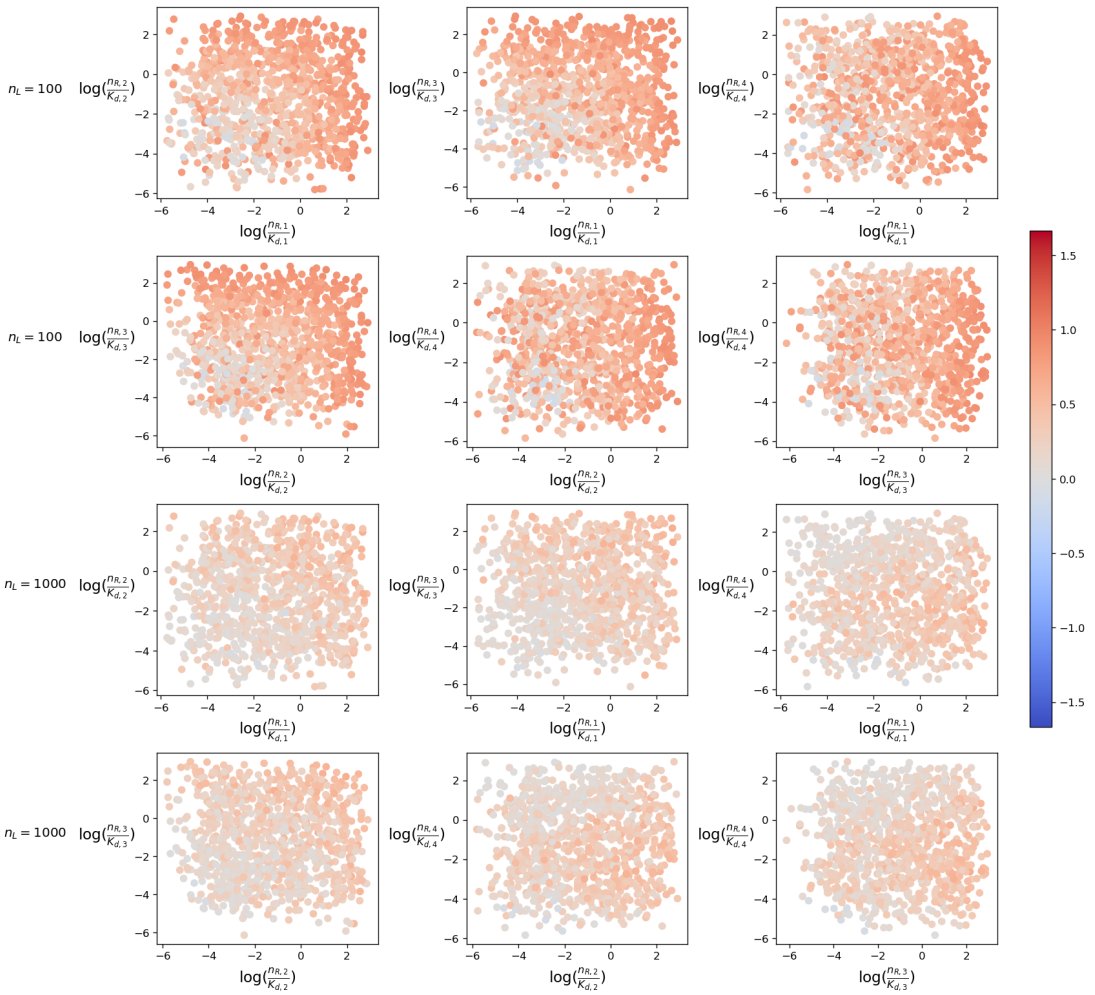

Figure 3.23: Relative difference $1 - \frac{\mathbb{E}^{NCA}[T_{(0,0,0,0)}(N)]}{\mathbb{E}^{EMA}[T_{(0,0,0,0)}(N)]}$ between the NCA and average of Gillespie simulations mean times to reach N complexes of type 4 for $n_L \in \{10^2, 10^3\}$. 10^3 parameter sets are sampled by considering the number of receptors $n_{R,j} \sim Unif(20, 200)$ and $K_{d,j} = 10^x$ with $x \sim Unif(1, 4)$, $j \in \{1, 2, 3, 4\}$. The unbinding rates are fixed at $k_{r,j} = 10^{-3}$ s^{-1} for $j \in \{1, 2, 3, 4\}$.

the percentage errors between the Gillespie derived time and the NCA derived time were less than 5% and this figure was 95% when comparing with the MCA. Finally it is also to be noted that the accuracy of the NCA and MCA as presented for the four receptor system is likely to increase in accuracy with a larger number of Gillespie simulations, and hence a better approximation of the expected time from simulation.

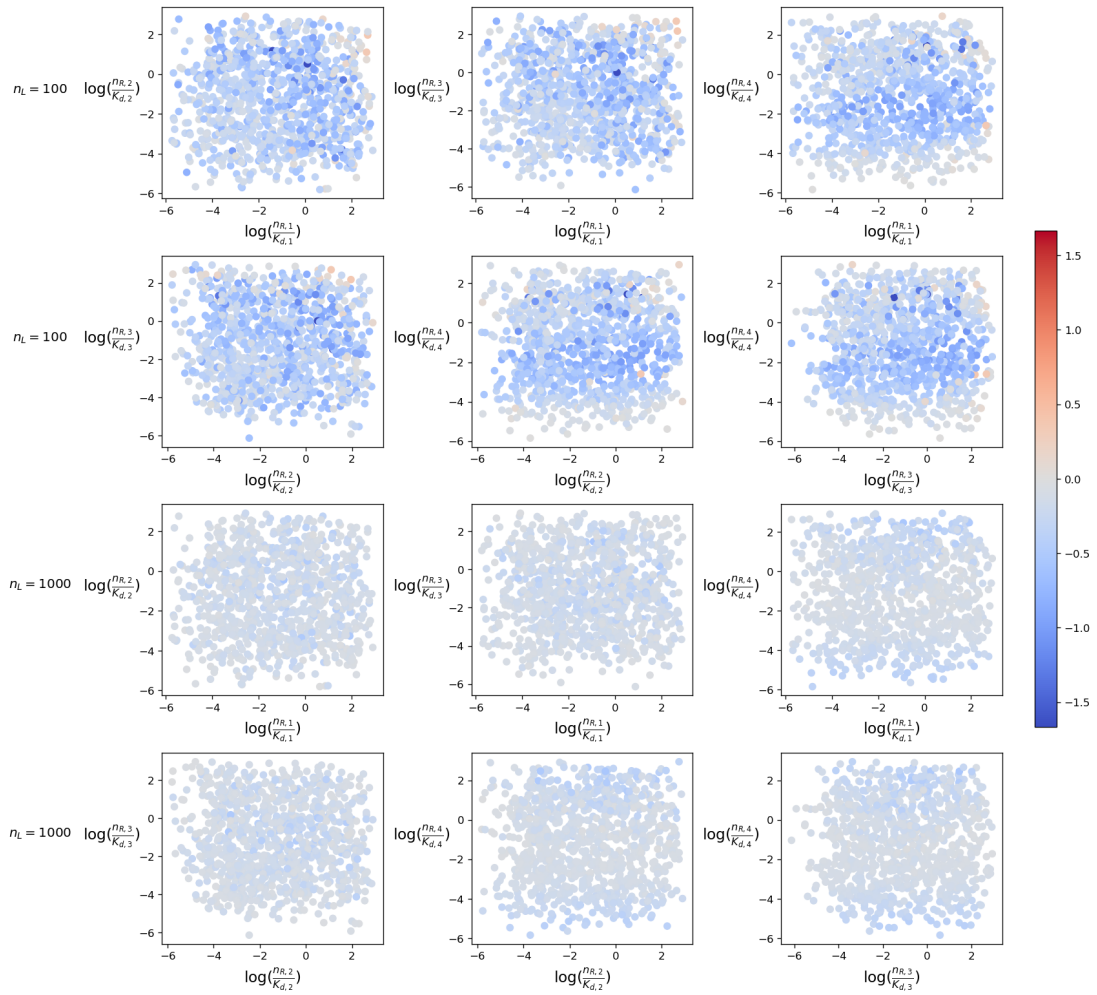

Figure 3.24: Relative difference $1 - \frac{\mathbb{E}^{MCA}[T_{(0,0,0,0)}(N)]}{\mathbb{E}^{EMA}[T_{(0,0,0,0)}(N)]}$ between the MCA and average of Gillespie simulations mean times to reach N complexes of type 4 for $n_L \in \{10^2, 10^3\}$. 10^3 parameter sets are sampled by considering the number of receptors $n_{R,j} \sim Unif(20, 200)$ and $K_{d,j} = 10^x$ with $x \sim Unif(1, 4)$, $j \in \{1, 2, 3, 4\}$. The unbinding rates are fixed at $k_{r,j} = 10^{-3}$ s^{-1} for $j \in \{1, 2, 3, 4\}$.

3.7 Discussion

In this chapter, mathematical models of receptor competition for a shared ligand have been introduced, expressed as multi-variate competition processes, as defined by Reuter (1961) and Iglehart *et al.* (1964) in the area of Mathematical Ecology. For any number of competing receptor types, this leads to the study

of a structured continuous-time Markov chain; in particular, a LD-QBD process. For these processes, the interest was in analysing two stochastic descriptors; the steady state distribution and the expected time to reach N complexes of a certain type, by implementing first-step arguments, leading to the study of systems of linear equations. Since the processes are LD-QBD processes, there are matrix-oriented algorithms available in the literature (Latouche *et al.*, 1999), which can be exploited in order to efficiently solve these systems of linear equations.

The main limitation of this matrix-oriented approach is that the number of equations increases with the number of states, which turns out to increase with the number of molecules in the system, typically large in *in vivo* and *in vitro* settings (López-García *et al.*, 2016). Thus, in this chapter, several approximations have been proposed based on the analysis of the process under low-to-moderate competition scenarios (*e.g.* when the number of competing receptors is small, there is excess of ligand, or the competing receptor has a relatively low affinity compared to the other receptor). These approximations lead to the analysis of independent one-dimensional birth-and-death processes, for which analogous computations can be carried out much more efficiently. Numerical comparison between the approximations and exact results, suggest that these approximations could be exploited in a wide range of parameter regimes, which have been explored inspired from values corresponding to the VEGF1 and VEGF2 receptors, which can both bind the shared ligand, VEGF-A.

A striking advantage of using the approximate methods to compute the stochastic descriptors presented in this chapter is that the computational cost is much less than the cost of computing the same descriptors in an exact fashion (EMA). In Figures 3.25 and 3.26, a simple computational comparison between the NCA/MCA and the EMA is presented for the two receptor system in Figure 3.1 (although similar comparisons could be carried out for higher dimensional systems). In particular, heatmaps of the central processing unit (CPU) times to run each method are plotted, for a range of numbers of the molecules $n_{R,1}$ and $n_{R,2}$. In both figures, $n_L = 10^4$ and $K_{d,1} = K_{d,2} = 10^3$, since for a large number of ligands such as $n_L = 10^4$, the computational cost of the EMA just depends on the values of $n_{R,1}$ and $n_{R,2}$, and is independent of $K_{d,1}$ and $K_{d,2}$. The colour bar on each figure represents the CPU time for the computation of the stochastic descriptor,

with units *minutes*. As expected, the EMA requires the largest amount of time to compute both descriptors since, especially when $n_{R,1}$ and $n_{R,2}$ are large, the method involves the computation of relatively large matrix inverses. There is a similar trend for both the NCA and MCA approaches to compute the steady state distribution, *i.e.* the larger $n_{R,1}$ and $n_{R,2}$, the greater the CPU time. It can be seen that, whereas the CPU time under the EMA takes up to a maximum of approximately 45 minutes, the NCA and MCA take less than 1 minute for all numbers of molecules considered. For the second descriptor, from Figure 3.26 it can be observed that the EMA takes roughly half the CPU time as the EMA for the steady state distribution in Figure 3.25. This is since the value of N has been set as $N = \frac{n_{R,2}}{2}$, for the expected times and hence only half of the matrix inverses are required compared to Figure 3.25. Again in Figure 3.26, the NCA and MCA are strikingly faster to compute than the EMA. As expected, the NCA in this case does not depend on the value of $n_{R,1}$ since this quantity does not feature in the calculation. For the MCA however, one must first compute $\mathbb{E}^{MCA}[M_1^*]$ using the MCA for the steady state distribution and hence the CPU times here *do* depend on $n_{R,1}$.

In Figure 3.27, specific pixels from Figures 3.25 and 3.26 are explored in more detail. In particular, 3 values of $n_{R,2}$ were chosen, corresponding to the 3 columns in Figure 3.27 and on the x-axis of each plot, the value of $n_{R,1}$ is varied. The top row of Figure 3.27 relates to the steady state distribution descriptor and hence the y-axis of the plots in this row is the HD. The second row in the figure relates to the expected times to reach N complexes of type 2 descriptor and hence the y-axis of the plots in this row is the relative difference as plotted in Figure 3.14. Finally the colour bar represents the CPU time saved with units *minutes*, by computing the descriptors using the approximations (NCA and MCA) instead of the analytic method (EMA). For the steady state descriptor in the first row it is seen that, as expected, the HD between the EMA and the NCA increases as $n_{R,1}$ increases, whereas the HD between the EMA and the MCA remains almost constant, indicating that this approximation works very well for these numbers of molecules. As $n_{R,1}$ and $n_{R,2}$ are increased, one finds that the time saved by using the NCA or MCA increases. In the second row, both the NCA and MCA worsen as approximations as $n_{R,1}$ is increased, but this worsening is only very slight (the

relative differences are always $\ll 1$). Again, as $n_{R,1}$ and $n_{R,2}$ are increased, the time saved by using the NCA or MCA increases. From this figure one can see that there is a trade-off between the HD or relative difference and the CPU time saved, but for the numbers of molecules and parameter values considered here, the time saved clearly outweighs the slight deviation from the EMA result caused by using the MCA for the steady state distribution and both the NCA and MCA for the expected times descriptor.

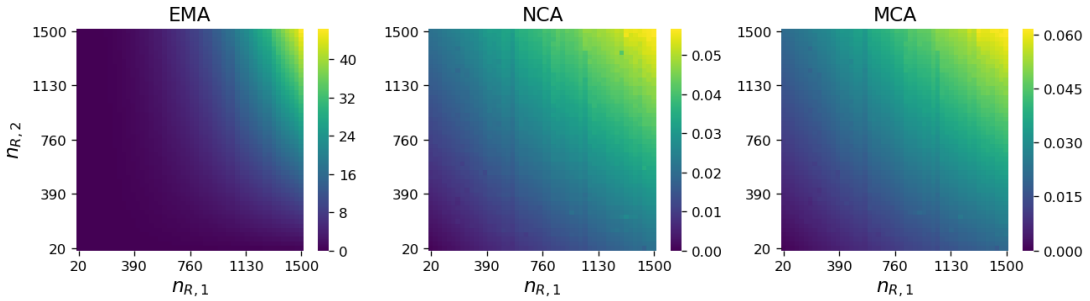

Figure 3.25: Comparison of the CPU time (in *minutes*) required to compute the steady state distribution for the EMA, NCA and MCA for different receptor numbers.

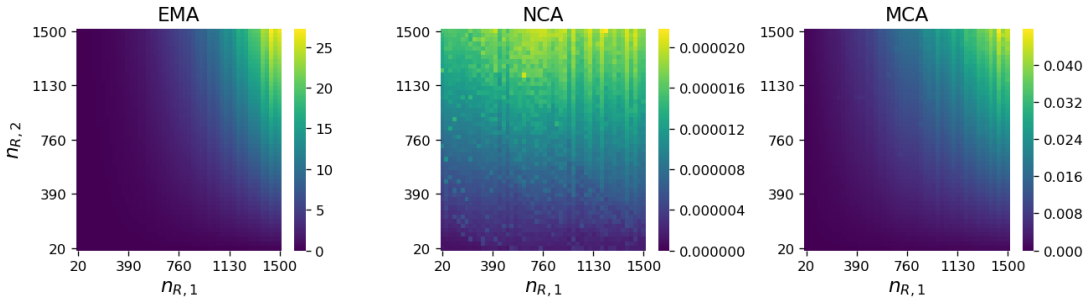

Figure 3.26: Comparison of the CPU time (in *minutes*) required to compute the mean time to reach $N = \frac{n_{R,2}}{2}$ complexes of type 2 for the EMA, NCA and MCA for different receptor numbers.

In this chapter, firstly a system was introduced in which there were two receptor types competing for a common ligand type, and then it was shown how the system could be extended to more receptor types. Both the analytic and approximate methods of computation of the stochastic descriptors analysed in this

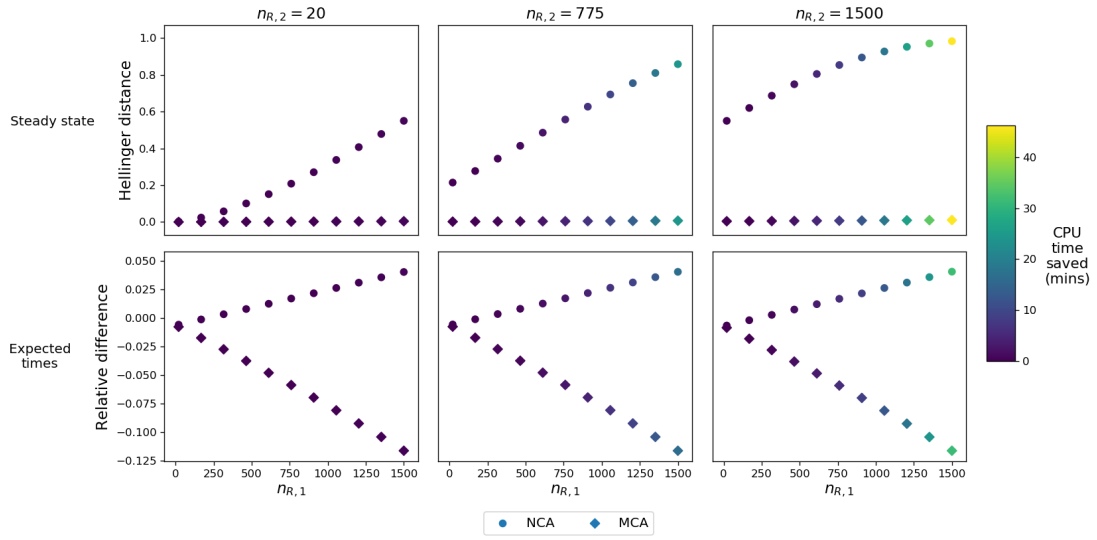

Figure 3.27: Top row: Trade-off between the HD, the value of $n_{R,1}$ and the CPU time saved by considering either the NCA or MCA instead of the EMA when computing the steady state distribution. **Bottom row:** Trade-off between the relative difference, the value of $n_{R,1}$ and the CPU time saved by considering either the NCA or MCA instead of the EMA when computing the expected time to reach N complexes of type 2. In all plots, $n_L = 10^4$ and $K_{d,1} = K_{d,2} = 10^3$.

chapter can be extended to these larger systems, however the approximations are much more efficient and computationally feasible for larger numbers of receptor types. To illustrate the need for a larger system, Roepstorff *et al.* (2009), describe how there are six different ligands which bind to EGFR (the epidermal growth factor receptor, discussed in Chapter 1). Hence one could imagine modelling this as a system of six variables (different ligand varieties), all competing for a common receptor (note that the system has been reversed here so that it is now ligands competing for a common receptor). To analyse a stochastic descriptor, such as the steady state distribution, for this system, it would be necessary to use the approximations presented in this chapter, as obtaining analytic results via the EMA would be intractable. As well as utilising the approximations for higher dimensional systems, one could also seek to use the ideas presented in this chapter to analyse further stochastic descriptors, similarly to as is done in Section 3.5.5 when considering productive complex formation. Moreover, since the methods presented here have been framed within processes originally devised

in Mathematical Ecology, it is to be expected that they could be extended to more general competition processes in this area (Gómez-Corral & López-García, 2012a,b, 2015), and are not restricted to molecular dynamics.

Finally, a clear limitation of the approaches in this chapter is that the model has focused on receptor-ligand binding dynamics without taking into account additional reactions such as receptor synthesis, degradation or internalisation. The analysis however, can be generalised to include those scenarios; for example, when modelling the surface *and* intracellular dynamics of two receptor types competing (on the surface) for a common ligand, one could use the techniques presented here to disentangle this competition, by the consideration of an effective number of ligands, so that the two intracellular processes can be separately studied.

Chapter 4

Mathematical modelling of cytokine receptor signalling

Cytokines are a class of small proteins, released by cells, which diffuse in the extracellular medium. A subclass of the cytokines are the interleukins, which are secreted by, and act upon, leukocytes, white blood cells of the immune system (Zhang & An, 2007). There are numerous (at least 40) interleukins which, in general, act as messenger molecules between cells and play a significant role in both the innate and the adaptive immune responses. They interact with cells by associating with membrane bound receptor molecules, and this interaction leads to cell signalling, for example, promoting differentiation, division or maturation of cells (Yuzhalin & Kutikhin, 2015). In this chapter, two specific interleukins are considered, interleukin 6 (IL-6) and interleukin 27 (IL-27). IL-6 acts primarily on B lymphocytes and hepatocytes and one of its main functions is to initiate, through binding with a receptor, the differentiation of B cells. IL-27 is a pro-inflammatory molecule and is produced by T cells (Petes *et al.*, 2018). A feature shared by these two interleukins is their ability to activate STAT proteins, specifically STAT1 and STAT3, which are part of the JAK/STAT pathway (Gordziel *et al.*, 2013; Hibbert *et al.*, 2003). Certain receptor molecules, such as glycoprotein 130 (GP130) and IL-27 receptor α (IL-27Rα), have phosphotyrosine residues which act as docking sites for STAT1 and STAT3 (Heinrich *et al.*, 1998). Such tyrosine residues become phosphorylated when the receptor forms a ligand-induced dimer with another receptor. Note that since cytokines can bind their associated

receptors, they behave as ligand molecules. Both IL-6 and IL-27 are capable of inducing such a dimer, where IL-6 induces the formation of a GP130 homodimer and IL-27 induces the formation of a GP130 - IL-27Rα heterodimer (Akdis *et al.*, 2011). Both such dimers are seen in Figure 4.1. As described in Chapter 1, dysregulation of the JAK/STAT pathway, for example through up-regulation of the cytokines IL-6 and IL-27, can have an impact on cell signalling, and the fate of a cell (for example, division, migration or death, see Chapter 1). In particular, this dysregulation can lead to autoimmune disorders such as SLE and Crohn's disease, and hence it is important to study the effects of up-regulation of these cytokines on different proteins in the JAK/STAT signalling pathway.

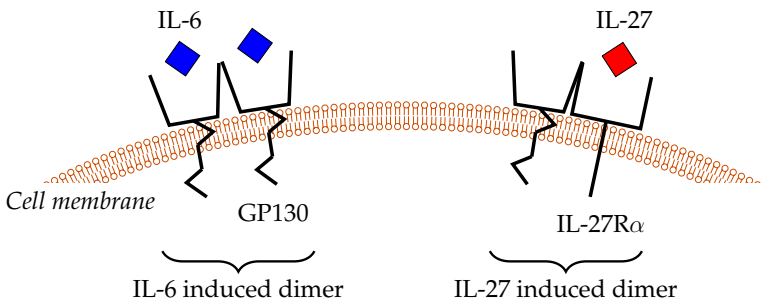

Figure 4.1: Diagram of the cytokine induced dimers formed under IL-6 and IL-27.

In this chapter, experimental data provided by Dr. Ignacio Moraga Gonzalez and Dr. Stephan Wilmes from the School of Life Sciences at the University of Dundee is used, which explores STAT1 and STAT3 activation by IL-6 and IL-27. They experimentally observed that, although these two cytokines can stimulate the same intracellular signalling pathway (JAK/STAT), they lead to differential signalling by STAT1. In particular, when the system is stimulated with IL-27, a greater and more sustained STAT1 signal is observed than when the system is stimulated with IL-6, whereas the STAT3 response is the same under both cytokines. Here, the aim is to understand the molecular mechanisms behind this differential STAT signalling induced by cytokine stimulation. Deterministic mathematical modelling and Bayesian model selection and inference are here used to learn about these potential molecular mechanisms. In Section 4.1, molecular

reactions which will inform mathematical models for both IL-6 and IL-27 stimulation are defined. Ordinary differential equation mathematical models are then developed based on these reactions. The experimental data used to compare with the mathematical model outputs is introduced in Section 4.2 and in Section 4.3, Bayesian model selection and Bayesian inference are used to inform hypotheses relating to the model reactions and to infer posterior distributions for the rate constants and initial concentrations in the models. In Section 4.4, the mathematical models are validated using additional experimental data and in Section 4.5, the models are used to predict changes in STAT signalling upon varying the concentrations of STATs and receptor molecules. Section 4.6 provides justification for choices made when developing the structure of the mathematical models with relation to other modelling work from the literature, specifically relating to negative feedback mechanisms. Finally, Section 4.7 is a discussion.

4.1 The IL-6 and IL-27 signalling mechanisms

The cytokines IL-6 and IL-27 both stimulate the JAK/STAT pathway, an important intracellular signal induction pathway. A summary of the JAK/STAT pathway is as follows (Dodington *et al.*, 2018). At the head of the pathway, spanning the cell surface membrane, is a ligand-induced receptor dimer, which is activated through autophosphorylation upon dimerisation. JAK proteins are then recruited to the intracellular tails of the receptors in the dimer and these molecules can, in turn, phosphorylate STAT proteins which also bind to the receptor tails. Phosphorylated STAT molecules then dissociate the receptor dimers and form dimers themselves before migrating to the nucleus to regulate gene transcription. The IL-6 and IL-27 JAK/STAT pathways differ due to the reactions involved in the formation of the signalling dimer under each cytokine. The signalling receptor complex formed under stimulation with IL-6 is a hexameric complex consisting of two molecules of IL-6, two molecules of IL-6 receptor α (IL-6Rα) and two molecules of GP130. In the experiments which produced the data used in this chapter however, a protein called hyper IL-6 (HypIL-6) was used, which is a molecular complex formed of IL-6 and IL-6Rα. This fusion protein was used in order to diminish signalling variability due to changes in IL-6Rα

expression. HypIL-6 then acts as the cytokine in the system and hence the signalling dimer is formed through two molecules of HypIL-6 and two of GP130. Under stimulation with IL-27, a heterodimer is formed consisting of one molecule of IL-27, one of IL-27Rα and one of GP130. Given that IL-27 has very weak affinity for GP130, it is assumed in the mathematical model that IL-27 binds firstly to IL-27Rα and then this complex forms a dimer with a GP130 unit.

In the mathematical models, it is assumed that the receptors in both dimer types phosphorylate immediately upon formation of the dimer and hence phosphorylation reactions are not included in the model. The JAK molecules do not feature as species in the mathematical models and hence it is assumed that they are constitutively bound to the corresponding receptors and that they phosphorylate immediately upon receptor phosphorylation (dimer formation) (Morris *et al.*, 2018). After the formation of the dimer, which will be denoted by either D_6 or D_{27}, formed by HypIL-6 or IL-27 respectively, the reactions for each mathematical model are similar, and are summarised as follows. A free cytoplasmic unphosphorylated STAT1 or STAT3 molecule can associate with either receptor in the dimer, provided that the intracellular tyrosine residue of the receptor in the dimer is free. The STAT1 or STAT3 molecule can subsequently unbind the receptor in the dimer or can become phosphorylated whilst bound to the dimer. Phosphorylated STAT1 (pSTAT1) and STAT3 (pSTAT3) molecules can also dissociate from the dimer where, once free in the cytoplasm, they can then become dephosphorylated. It is assumed that the rate of pSTAT dephosphorylation depends only on the concentration of the respective STAT type. No allostery is considered in the model and hence, phosphorylated and unphosphorylated (p)STAT molecules dissociate the receptor at the same rate. Given that STAT phosphorylation is independent of the receptor which the STAT molecule is bound to, and of the STAT type, there is only one rate of STAT phosphorylation in the model. Finally, species containing receptor molecules are removed from the system, due to receptor internalisation or degradation, via one of two hypothesised mechanisms,

- hypothesis 1 (H_1): receptors (free or bound, phosphorylated or unphosphorylated) are internalised/degraded with rate proportional to the concentration of the species in which they are contained, and

- hypothesis 2 (H_2): receptors (free or bound, phosphorylated or unphosphorylated) are internalised/degraded with rate proportional to the product of the concentration of the species in which they are contained and the sum of the concentrations of free cytoplasmic phosphorylated STAT1 and STAT3.

Hypothesis 1 assumes that receptor molecules (free or bound, phosphorylated or unphosphorylated) are being internalised/degraded as part of the natural cellular trafficking cycle. Hypothesis 2 is consistent with a potential negative feedback mechanism, whereby the free cytoplasmic pSTAT molecules would migrate to the nucleus and increase the translation of negative feedback proteins such as SOCS3, which down-regulate cytokine signalling (Brender *et al.*, 2007; Croker *et al.*, 2003). Thus, the internalization/degradation rate of receptor molecules (free or bound, phosphorylated or unphosphorylated) under hypothesis 2 increases with the total amount of free cytoplasmic phosphorylated STAT1 and STAT3, to account for this negative feedback on surface receptor expression. Although in cytokine systems, such as those considered here, there may be many other feedback mechanisms in place, such as positive feedback mechanisms (Arbouzova & Zeidler, 2006; Shuai & Liu, 2003), given the duration of the kinetic experiments (three hours), they would not be relevant here and hence only the two aforementioned hypotheses relating to internalisation/degradation of receptors are considered. A diagram which describes the molecular reactions in each model (HypIL-6 and IL-27) is shown in Figure 4.2 and the complete model reaction scheme is given in Figure 4.3, where a), c), e) and g) comprise the HypIL-6 model and b), d), f) and g) comprise the IL-27 model. In the figure, $i \in \{1, 3\}$ so that the reactions shown can either involve STAT1 or STAT3. Each reaction arrow has been shown with its rate (above or below the arrow). The notation for the rate constants and initial concentrations in the model, along with their descriptions and units, are given in Table 4.1. The initial concentrations of all model variables not included in Table 4.1 are fixed at 0 nM when integrating the mathematical models. Further steps in the pathway such as pSTAT dimerisation and gene transcription are not included in the mathematical models for simplicity, and since it is the initial steps in the JAK/STAT pathway that are of interest here. In Sections 4.1.1 and 4.1.2, ODE models are formulated for the HypIL-6 and IL-27 pathways, respectively, based on the reactions described here.

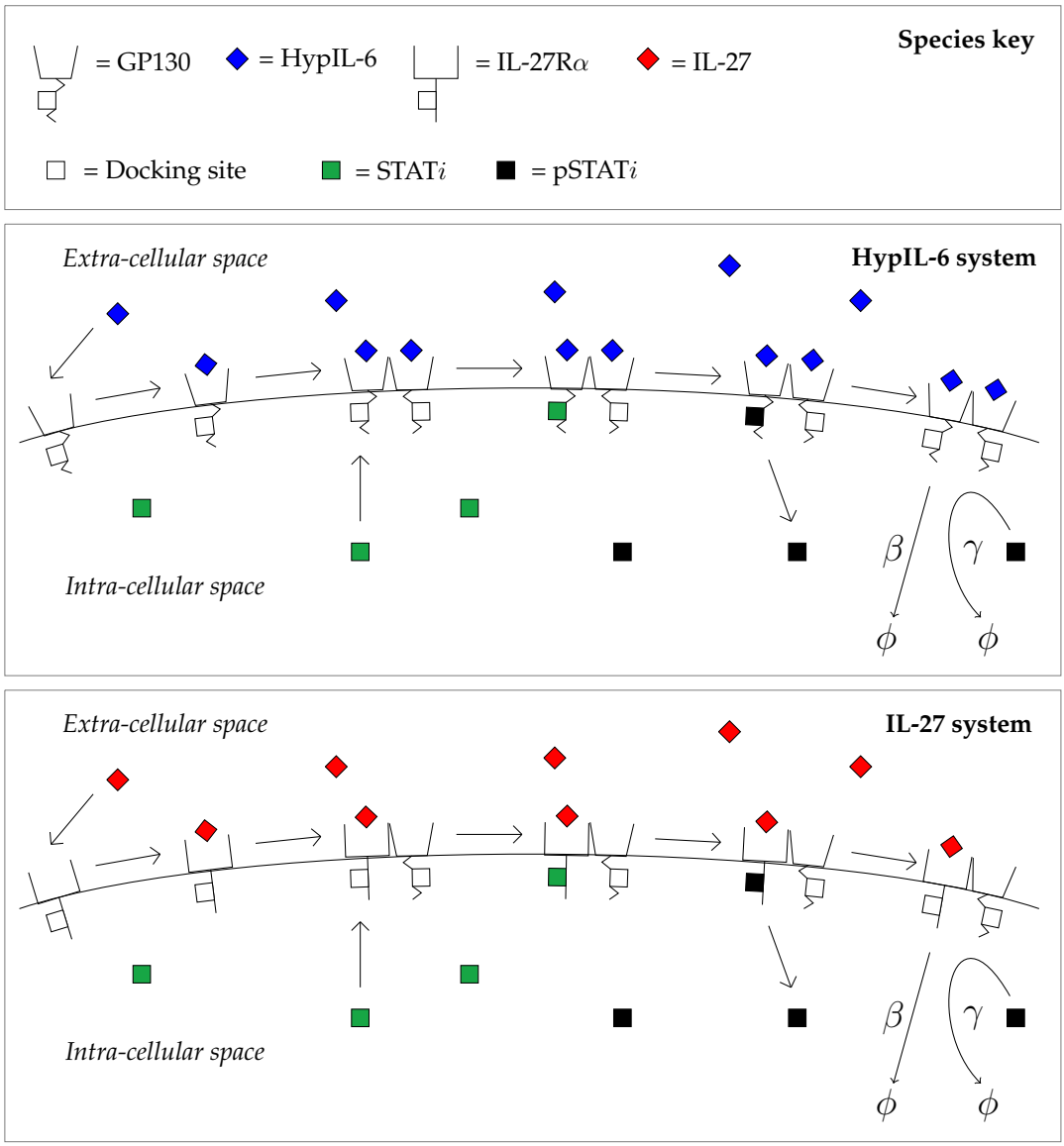

Figure 4.2: Diagram of the reactions for the HypIL-6 and IL-27 mathematical models. From left to right in a single model panel: cytokines can bind to unbound receptors, dimerisation of receptor complexes can occur and STAT molecules can bind to the dimers, where they can then phosphorylate and dissociate. Each panel is one such example of the model but in general STAT molecules can bind to either receptor in the dimer until two STATs are bound to a given receptor-ligand dimer. The reverse reactions are also included in the models, but have not been included in the diagram for simplicity. Finally, in each model (HypIL-6 or IL-27), any molecular species involving a receptor molecule of either type can be internalised/degraded.

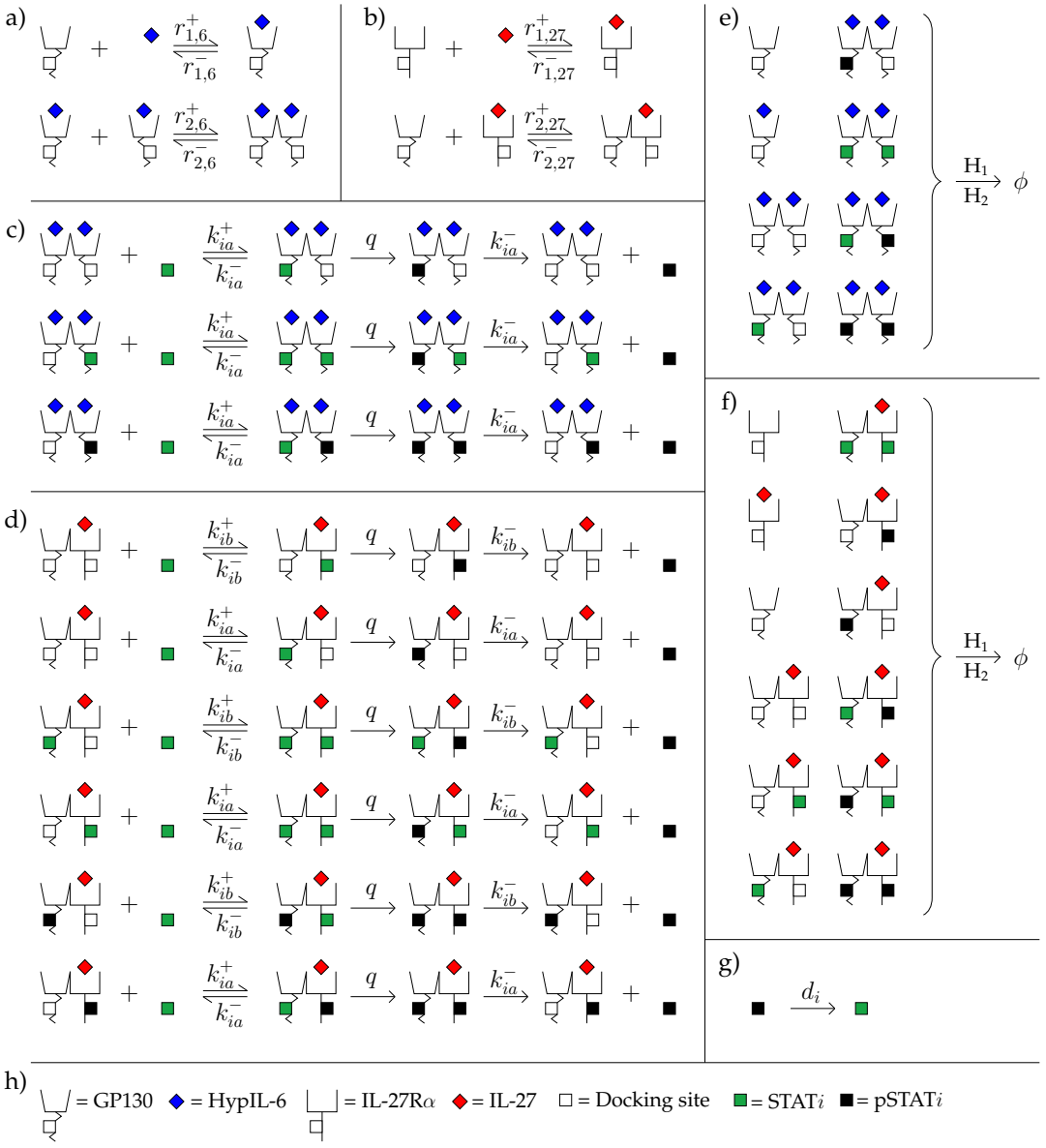

Figure 4.3: Depiction of the reactions defining the HypIL-6 and IL-27 mathematical models. a) Reactions involving ligand binding and dimerisation in the HypIL-6 model. b) Reactions involving ligand binding and dimerisation in the IL-27 model. c) Reactions involving STATi molecules, for $i \in \{1, 3\}$, in the HypIL-6 model. d) Reactions involving STATi molecules, for $i \in \{1, 3\}$, in the IL-27 model. e) Reactions involving receptor internalisation/degradation in the HypIL-6 model. Here $H_1 = \beta_6$ and $H_2 = \gamma_6([\text{pSTAT1}] + [\text{pSTAT3}])$ where square brackets around a species denote the concentration of the species. f) Reactions involving receptor internalisation/degradation in the IL-27 model. Here $H_1 = \beta_{27}$ and $H_2 = \gamma_{27}([\text{pSTAT1}] + [\text{pSTAT3}])$. g) Dephosphorylation of pS_i, for $i \in \{1, 3\}$, in the cytoplasm. This reaction occurs in both models. h) Key for the molecules in the reactions.

Parameter/IC	Description	Unit
$r^+_{1,6}$, $r^+_{1,27}$	Rate of receptor-ligand binding	$nM^{-1}s^{-1}$
$r^-_{1,6}$, $r^-_{1,27}$	Rate of receptor-ligand unbinding	s^{-1}
$r^+_{2,6}$, $r^+_{2,27}$	Rate of dimer association	$nM^{-1}s^{-1}$
$r^-_{2,6}$, $r^-_{2,27}$	Rate of dimer dissociation	s^{-1}
k^+_{ia}	Rate of STATi binding to GP130	$nM^{-1}s^{-1}$
k^+_{ib}	Rate of STATi binding to IL-27Rα	$nM^{-1}s^{-1}$
k^-_{ia}	Rate of STATi unbinding from GP130	s^{-1}
k^-_{ib}	Rate of STATi unbinding from IL-27Rα	s^{-1}
q	Rate of STAT phosphorylation on the dimer	s^{-1}
d_i	Rate of pSTATi dephosphorylation	s^{-1}
β_6, β_{27}	Rate of D internalisation/degradation	s^{-1}
γ_6, γ_{27}	Rate of D loss due to feedback	$nM^{-1}s^{-1}$
$[R_1](0)$	Initial concentration of GP130	nM
$[R_2](0)$	Initial concentration of IL-27Rα	nM
$[S_i(0)]$	Initial concentration of STATi	nM
$[L_6](0)$	Initial concentration of HypIL-6	nM
$[L_{27}](0)$	Initial concentration of IL-27	nM

Table 4.1: Definitions and units for the rate constants and initial concentrations in the mathematical models, where $i \in \{1, 3\}$ so that STATi corresponds to STAT1 or STAT3. A parameter with "6" in its notation is found only in the HypIL-6 model and likewise a parameter with "27" in its notation is found only in the IL-27 model.

4.1.1 HypIL-6 mathematical model

The HypIL-6 mathematical model was formulated based on biochemical reactions involving the following species:

- L_6 = HypIL-6,
- R_1 = GP130,

- C_1 = GP130 - HypIL-6 complex,
- D_6 = phosphorylated GP130 - HypIL-6 - HypIL-6 - GP130 homodimer,
- S_1 = unbound cytoplasmic unphosphorylated STAT1,
- S_3 = unbound cytoplasmic unphosphorylated STAT3,
- $D_6 \cdot S_1$ = dimer bound to STAT1,
- $D_6 \cdot S_3$ = dimer bound to STAT3,
- $D_6 \cdot pS_1$ = dimer bound to pSTAT1,
- $D_6 \cdot pS_3$ = dimer bound to pSTAT3,
- $S_1 \cdot D_6 \cdot S_1$ = dimer bound to two molecules of STAT1,
- $pS_1 \cdot D_6 \cdot S_1$ = dimer bound to two molecules of STAT1, one of which is phosphorylated,
- $pS_1 \cdot D_6 \cdot pS_1$ = dimer bound to two molecules of pSTAT1,
- $S_3 \cdot D_6 \cdot S_3$ = dimer bound to two molecules of STAT3,
- $pS_3 \cdot D_6 \cdot S_3$ = dimer bound to two molecules of STAT3, one of which is phosphorylated,
- $pS_3 \cdot D_6 \cdot pS_3$ = dimer bound to two molecules of pSTAT3,
- $S_1 \cdot D_6 \cdot S_3$ = dimer bound to one molecule of STAT1 and one of STAT3,
- $pS_1 \cdot D_6 \cdot S_3$ = dimer bound to one molecule of pSTAT1 and one of STAT3,
- $S_1 \cdot D_6 \cdot pS_3$ = dimer bound to one molecule of STAT1 and one of pSTAT3,
- $pS_1 \cdot D_6 \cdot pS_3$ = dimer bound to one molecule of pSTAT1 and one of pSTAT3,
- pS_1 = unbound cytoplasmic phosphorylated STAT1,
- pS_3 = unbound cytoplasmic phosphorylated STAT3.

The initial reactions in the HypIL-6 signalling pathway as described in Section 4.1 then inform the ODEs (4.1) - (4.22), under the law of mass action kinetics, where the terms involving the parameter β_6 apply only to the model under hypothesis 1 and the terms involving the parameter γ_6 apply only to the model

under hypothesis 2. Square brackets around a species denote the concentration of this species with units nM, and "·" implies a reaction bond between two molecules/species. The ODEs are valid for any time t, with $t \geq 0$, but time has been omitted in the species concentration notation for ease of notation, where for example $[R_1] = [R_1](t)$ for all $t \geq 0$.

$$\frac{d[R_1]}{dt} = -r^+_{1,6}[R_1][L_6] + r^-_{1,6}[C_1] - \beta_6[R_1] - \gamma_6([pS_1] + [pS_3])[R_1] \tag{4.1}$$

$$\frac{d[L_6]}{dt} = -r^+_{1,6}[R_1][L_6] + r^-_{1,6}[C_1] \tag{4.2}$$

$$\begin{aligned}\frac{d[C_1]}{dt} &= r^+_{1,6}[R_1][L_6] - r^-_{1,6}[C_1] - 2r^+_{2,6}[C_1]^2 + 2r^-_{2,6}[D_6] - \beta_6[C_1] \\ &\quad - \gamma_6([pS_1] + [pS_3])[C_1]\end{aligned} \tag{4.3}$$

$$\begin{aligned}\frac{d[D_6]}{dt} &= r^+_{2,6}[C_1]^2 - r^-_{2,6}[D_6] - 2k^+_{1a}[D_6][S_1] + k^-_{1a}([D_6 \cdot S_1] + [D_6 \cdot pS_1]) \\ &\quad - 2k^+_{3a}[D_6][S_3] + k^-_{3a}([D_6 \cdot S_3] + [D_6 \cdot pS_3]) - \beta_6[D_6] \\ &\quad - \gamma_6([pS_1] + [pS_3])[D_6]\end{aligned} \tag{4.4}$$

$$\begin{aligned}\frac{d[S_1]}{dt} &= -k^+_{1a}[S_1](2[D_6] + [D_6 \cdot S_1] + [D_6 \cdot S_3] + [D_6 \cdot pS_1] + [D_6 \cdot pS_3]) \\ &\quad + k^-_{1a}([D_6 \cdot S_1] + 2[S_1 \cdot D_6 \cdot S_1] + [S_1 \cdot D_6 \cdot S_3] + [S_1 \cdot D_6 \cdot pS_1] \\ &\quad + [S_1 \cdot D_6 \cdot pS_3]) + d_1[pS_1]\end{aligned} \tag{4.5}$$

$$\begin{aligned}\frac{d[S_3]}{dt} &= -k^+_{3a}[S_3](2[D_6] + [D_6 \cdot S_3] + [D_6 \cdot S_1] + [D_6 \cdot pS_3] + [D_6 \cdot pS_1]) \\ &\quad + k^-_{3a}([D_6 \cdot S_3] + 2[S_3 \cdot D_6 \cdot S_3] + [S_1 \cdot D_6 \cdot S_3] + [S_3 \cdot D_6 \cdot pS_3] \\ &\quad + [S_3 \cdot D_6 \cdot pS_1]) + d_3[pS_3]\end{aligned} \tag{4.6}$$

$$\begin{aligned}\frac{d[D_6 \cdot S_1]}{dt} &= 2k^+_{1a}[S_1][D_6] - k^-_{1a}[D_6 \cdot S_1] - k^+_{1a}[D_6 \cdot S_1][S_1] + 2k^-_{1a}[S_1 \cdot D_6 \cdot S_1] \\ &\quad - k^+_{3a}[D_6 \cdot S_1][S_3] + k^-_{3a}[S_3 \cdot D_6 \cdot S_1] - q[D_6 \cdot S_1] + k^-_{1a}[S_1 \cdot D_6 \cdot pS_1] \\ &\quad + k^-_{3a}[S_1 \cdot D_6 \cdot pS_3] - \beta_6[D_6 \cdot S_1] - \gamma_6([pS_1] + [pS_3])[D_6 \cdot S_1]\end{aligned} \tag{4.7}$$

$$\frac{d[D_6 \cdot S_3]}{dt} = 2k^+_{3a}[S_3][D_6] - k^-_{3a}[D_6 \cdot S_3] - k^+_{3a}[D_6 \cdot S_3][S_3] + 2k^-_{3a}[S_3 \cdot D_6 \cdot S_3]$$

$$- k_{1a}^{+}[D_6 \cdot S_3][S_1] + k_{1a}^{-}[S_1 \cdot D_6 \cdot S_3] - q[D_6 \cdot S_3] + k_{1a}^{-}[S_3 \cdot D_6 \cdot pS_1]$$
$$+ k_{3a}^{-}[S_3 \cdot D_6 \cdot pS_3] - \beta_6[D_6 \cdot S_3] - \gamma_6([pS_1] + [pS_3])[D_6 \cdot S_3] \tag{4.8}$$

$$\frac{d[D_6 \cdot pS_1]}{dt} = -k_{1a}^{+}[S_1][D_6 \cdot pS_1] + k_{1a}^{-}[S_1 \cdot D_6 \cdot pS_1] - k_{3a}^{+}[S_3][D_6 \cdot pS_1]$$
$$+ k_{3a}^{-}[S_3 \cdot D_6 \cdot pS_1] + q[D_6 \cdot S_1] - k_{1a}^{-}[D_6 \cdot pS_1] + 2k_{1a}^{-}[pS_1 \cdot D_6 \cdot pS_1]$$
$$+ k_{3a}^{-}[pS_1 \cdot D_6 \cdot pS_3] - \beta_6[D_6 \cdot pS_1]$$
$$- \gamma_6([pS_1] + [pS_3])[D_6 \cdot pS_1] \tag{4.9}$$

$$\frac{d[D_6 \cdot pS_3]}{dt} = -k_{3a}^{+}[S_3][D_6 \cdot pS_3] + k_{3a}^{-}[S_3 \cdot D_6 \cdot pS_3] - k_{1a}^{+}[S_1][D_6 \cdot pS_3]$$
$$+ k_{1a}^{-}[S_1 \cdot D_6 \cdot pS_3] + q[D_6 \cdot S_3] - k_{3a}^{-}[D_6 \cdot pS_3] + 2k_{3a}^{-}[pS_3 \cdot D_6 \cdot pS_3]$$
$$+ k_{1a}^{-}[pS_1 \cdot D_6 \cdot pS_3] - \beta_6[D_6 \cdot pS_3]$$
$$- \gamma_6([pS_1] + [pS_3])[D_6 \cdot pS_3] \tag{4.10}$$

$$\frac{d[S_1 \cdot D_6 \cdot S_1]}{dt} = k_{1a}^{+}[S_1][D_6 \cdot S_1] - 2k_{1a}^{-}[S_1 \cdot D_6 \cdot S_1] - 2q[S_1 \cdot D_6 \cdot S_1]$$
$$- \beta_6[S_1 \cdot D_6 \cdot S_1] - \gamma_6([pS_1] + [pS_3])[S_1 \cdot D_6 \cdot S_1] \tag{4.11}$$

$$\frac{d[S_3 \cdot D_6 \cdot S_3]}{dt} = k_{3a}^{+}[S_3][D_6 \cdot S_3] - 2k_{3a}^{-}[S_3 \cdot D_6 \cdot S_3] - 2q[S_3 \cdot D_6 \cdot S_3]$$
$$- \beta_6[S_3 \cdot D_6 \cdot S_3] - \gamma_6([pS_1] + [pS_3])[S_3 \cdot D_6 \cdot S_3] \tag{4.12}$$

$$\frac{d[pS_1 \cdot D_6 \cdot S_1]}{dt} = k_{1a}^{+}[pS_1 \cdot D_6][S_1] - 2k_{1a}^{-}[pS_1 \cdot D_6 \cdot S_1] + 2q[S_1 \cdot D_6 \cdot S_1]$$
$$- q[pS_1 \cdot D_6 \cdot S_1] - \beta_6[pS_1 \cdot D_6 \cdot S_1]$$
$$- \gamma_6([pS_1] + [pS_3])[pS_1 \cdot D_6 \cdot S_1] \tag{4.13}$$

$$\frac{d[pS_3 \cdot D_6 \cdot S_3]}{dt} = k_{3a}^{+}[pS_3 \cdot D_6][S_3] - 2k_{3a}^{-}[pS_3 \cdot D_6 \cdot S_3] + 2q[S_3 \cdot D_6 \cdot S_3]$$
$$- q[pS_3 \cdot D_6 \cdot S_3] - \beta_6[pS_3 \cdot D_6 \cdot S_3]$$
$$- \gamma_6([pS_1] + [pS_3])[pS_3 \cdot D_6 \cdot S_3] \tag{4.14}$$

$$\frac{d[pS_1 \cdot D_6 \cdot pS_1]}{dt} = q[pS_1 \cdot D_6 \cdot S_1] - 2k_{1a}^{-}[pS_1 \cdot D_6 \cdot pS_1] - \beta_6[pS_1 \cdot D_6 \cdot pS_1]$$
$$- \gamma_6([pS_1] + [pS_3])[pS_1 \cdot D_6 \cdot pS_1] \tag{4.15}$$

$$\frac{d[pS_3 \cdot D_6 \cdot pS_3]}{dt} = q[pS_3 \cdot D_6 \cdot S_3] - 2k_{3a}^{-}[pS_3 \cdot D_6 \cdot pS_3] - \beta_6[pS_3 \cdot D_6 \cdot pS_3] - \gamma_6([pS_1] + [pS_3])[pS_3 \cdot D_6 \cdot pS_3] \tag{4.16}$$

$$\frac{d[S_1 \cdot D_6 \cdot S_3]}{dt} = k_{1a}^{+}[S_1][D_6 \cdot S_3] - k_{1a}^{-}[S_1 \cdot D_6 \cdot S_3] + k_{3a}^{+}[S_1 \cdot D_6][S_3] - k_{3a}^{-}[S_1 \cdot D_6 \cdot S_3] - 2q[S_1 \cdot D_6 \cdot S_3] - \beta_6[S_1 \cdot D_6 \cdot S_3] - \gamma_6([pS_1] + [pS_3])[S_1 \cdot D_6 \cdot S_3] \tag{4.17}$$

$$\frac{d[pS_1 \cdot D_6 \cdot S_3]}{dt} = q[S_1 \cdot D_6 \cdot S_3] + k_{3a}^{+}[pS_1 \cdot D_6][S_3] - k_{3a}^{-}[pS_1 \cdot D_6 \cdot S_3] - q[pS_1 \cdot D_6 \cdot S_3] - k_{1a}^{-}[pS_1 \cdot D_6 \cdot S_3] - \beta_6[pS_1 \cdot D_6 \cdot S_3] - \gamma_6([pS_1] + [pS_3])[pS_1 \cdot D_6 \cdot S_3] \tag{4.18}$$

$$\frac{d[S_1 \cdot D_6 \cdot pS_3]}{dt} = q[S_1 \cdot D_6 \cdot S_3] + k_{1a}^{+}[S_1][D_6 \cdot pS_3] - k_{1a}^{-}[S_1 \cdot D_6 \cdot pS_3] - q[S_1 \cdot D_6 \cdot pS_3] - k_{3a}^{-}[S_1 \cdot D_6 \cdot pS_3] - \beta_6[S_1 \cdot D_6 \cdot pS_3] - \gamma_6([pS_1] + [pS_3])[S_1 \cdot D_6 \cdot pS_3] \tag{4.19}$$

$$\frac{d[pS_1 \cdot D_6 \cdot pS_3]}{dt} = q([S_1 \cdot D_6 \cdot pS_3] + [pS_1 \cdot D_6 \cdot S_3]) - [pS_1 \cdot D_6 \cdot pS_3](k_{1a}^{-} + k_{3a}^{-}) - \beta_6[pS_1 \cdot D_6 \cdot pS_3] - \gamma_6([pS_1] + [pS_3])[pS_1 \cdot D_6 \cdot pS_3] \tag{4.20}$$

$$\frac{d[pS_1]}{dt} = k_{1a}^{-}([D_6 \cdot pS_1] + [S_1 \cdot D_6 \cdot pS_1] + [S_3 \cdot D_6 \cdot pS_1] + [pS_3 \cdot D_6 \cdot pS_1] + 2[pS_1 \cdot D_6 \cdot pS_1]) - d_1[pS_1] \tag{4.21}$$

$$\frac{d[pS_3]}{dt} = k_{3a}^{-}([D_6 \cdot pS_3] + [S_3 \cdot D_6 \cdot pS_3] + [S_1 \cdot D_6 \cdot pS_3] + [pS_1 \cdot D_6 \cdot pS_3] + 2[pS_3 \cdot D_6 \cdot pS_3]) - d_3[pS_3] \tag{4.22}$$

4.1.2 IL-27 mathematical model

With some species in common with the HypIL-6 model, the IL-27 model has been formulated based on biochemical reactions involving the following species:

- L_{27} = IL-27,

- R_1 = GP130,
- R_2 = IL-27Rα,
- C_2 = IL-27Rα - IL-27 complex,
- D_{27} = phosphorylated IL-27Rα - IL-27 - GP130 heterodimer,
- S_1 = unbound cytoplasmic unphosphorylated STAT1,
- S_3 = unbound cytoplasmic unphosphorylated STAT3,
- $S_1 \cdot D_{27}$ = dimer bound to STAT1 via R_1,
- $S_3 \cdot D_{27}$ = dimer bound to STAT3 via R_1,
- $pS_1 \cdot D_{27}$ = dimer bound to pSTAT1 via R_1,
- $pS_3 \cdot D_{27}$ = dimer bound to pSTAT3 via R_1,
- $D_{27} \cdot S_1$ = dimer bound to STAT1 via R_2,
- $D_{27} \cdot S_3$ = dimer bound to STAT3 via R_2,
- $D_{27} \cdot pS_1$ = dimer bound to pSTAT1 via R_2,
- $D_{27} \cdot pS_3$ = dimer bound to pSTAT3 via R_2,
- $S_1 \cdot D_{27} \cdot S_1$ = dimer bound to two molecules of STAT1,
- $pS_1 \cdot D_{27} \cdot S_1$ = dimer bound to two molecules of STAT1, phosphorylated on R_1,
- $S_1 \cdot D_{27} \cdot pS_1$ = dimer bound to two molecules of STAT1, phosphorylated on R_2,
- $pS_1 \cdot D_{27} \cdot pS_1$ = dimer bound to two molecules of pSTAT1,
- $S_3 \cdot D_{27} \cdot S_3$ = dimer bound to two molecules of STAT3,
- $pS_3 \cdot D_{27} \cdot S_3$ = dimer bound to two molecules of STAT3, phosphorylated on R_1,
- $S_3 \cdot D_{27} \cdot pS_3$ = dimer bound to two molecules of STAT3, phosphorylated on R_2,
- $pS_3 \cdot D_{27} \cdot pS_3$ = dimer bound to two molecules of pSTAT3,
- $S_1 \cdot D_{27} \cdot S_3$ = dimer bound to STAT1 via R_1 and STAT3 via R_2,

- $S_3 \cdot D_{27} \cdot S_1$ = dimer bound to STAT1 via R_2 and STAT3 via R_1,
- $pS_1 \cdot D_{27} \cdot S_3$ = dimer bound to pSTAT1 via R_1 and STAT3 via R_2,
- $S_3 \cdot D_{27} \cdot pS_1$ = dimer bound to pSTAT1 via R_2 and STAT3 via R_1,
- $S_1 \cdot D_{27} \cdot pS_3$ = dimer bound to STAT1 via R_1 and pSTAT3 via R_2,
- $pS_3 \cdot D_{27} \cdot S_1$ = dimer bound to STAT1 via R_2 and pSTAT3 via R_1,
- $pS_1 \cdot D_{27} \cdot pS_3$ = dimer bound pSTAT1 via R_1 and pSTAT3 via R_2,
- $pS_3 \cdot D_{27} \cdot pS_1$ = dimer bound pSTAT3 via R_1 and pSTAT3 via R_1,
- pS_1 = unbound cytoplasmic phosphorylated STAT1,
- pS_3 = unbound cytoplasmic phosphorylated STAT3.

Again under the law of mass action kinetics, the initial reactions in the IL-27 signalling pathway can be described by the ODEs (4.23) - (4.55).

$$\begin{aligned}\frac{d[R_1]}{dt} &= -r^+_{2,27}[C_2][R_1] + r^-_{2,27}[D_{27}] - \beta_{27}[R_1] \\ &\quad - \gamma_{27}([pS_1] + [pS_3])[R_1] \end{aligned} \tag{4.23}$$

$$\begin{aligned}\frac{d[R_2]}{dt} &= -r^+_{1,27}[R_2][L_{27}] + r^-_{1,27}[C_2] - \beta_{27}[R_2] \\ &\quad - \gamma_{27}([pS_1] + [pS_3])[R_2] \end{aligned} \tag{4.24}$$

$$\frac{d[L_{27}]}{dt} = -r^+_{1,27}[R_2][L_{27}] + r^-_{1,27}[C_2] \tag{4.25}$$

$$\begin{aligned}\frac{d[C_2]}{dt} &= r^+_{1,27}[R_2][L_{27}] - r^-_{1,27}[C_2] - r^+_{2,27}[C_2][R_1] + r^-_{2,27}[D_{27}] - \beta_{27}[C_2] \\ &\quad - \gamma_{27}([pS_1] + [pS_3])[C_2] \end{aligned} \tag{4.26}$$

$$\begin{aligned}\frac{d[D_{27}]}{dt} &= r^+_{2,27}[C_2][R_1] - r^-_{2,27}[D_{27}] - (k^+_{1a} + k^+_{1b})[D_{27}][S_1] \\ &\quad + k^-_{1a}([S_1 \cdot D_{27}] + [pS_1 \cdot D_{27}]) + k^-_{1b}([D_{27} \cdot S_1] + [D_{27} \cdot pS_1]) \\ &\quad - (k^+_{3a} + k^+_{3b})[D_{27}][S_3] + k^-_{3a}([S_3 \cdot D_{27}] + [pS_3 \cdot D_{27}]) \\ &\quad + k^-_{3b}([D_{27} \cdot S_3] + [D_{27} \cdot pS_3]) - \beta_{27}[D_{27}] \\ &\quad - \gamma_{27}([pS_1] + [pS_3])[D_{27}] \end{aligned} \tag{4.27}$$

$$\begin{aligned}\frac{d[S_1]}{dt} &= -k_{1a}^{+}[S_1]([D_{27}] + [D_{27}\cdot S_1] + [D_{27}\cdot pS_1] + [D_{27}\cdot S_3] + [D_{27}\cdot pS_3]) \\ &+ k_{1a}^{-}([S_1\cdot D_{27}] + [S_1\cdot D_{27}\cdot S_1] + [S_1\cdot D_{27}\cdot pS_1] + [S_1\cdot D_{27}\cdot S_3] \\ &+ [S_1\cdot D_{27}\cdot pS_3]) - k_{1b}^{+}[S_1]([D_{27}] + [S_1\cdot D_{27}] + [pS_1\cdot D_{27}] + [S_3\cdot D_{27}] \\ &+ [pS_3\cdot D_{27}]) + k_{1b}^{-}([D_{27}\cdot S_1] + [S_1\cdot D_{27}\cdot S_1] + [pS_1\cdot D_{27}\cdot S_1] \\ &+ [S_3\cdot D_{27}\cdot S_1] + [pS_3\cdot D_{27}\cdot S_1]) + d_1[pS_1] \end{aligned} \tag{4.28}$$

$$\begin{aligned}\frac{d[S_3]}{dt} &= -k_{3a}^{+}[S_3]([D_{27}] + [D_{27}\cdot S_1] + [D_{27}\cdot pS_1] + [D_{27}\cdot S_3] + [D_{27}\cdot pS_3]) \\ &+ k_{3a}^{-}([S_3\cdot D_{27}] + [S_3\cdot D_{27}\cdot S_1] + [S_3\cdot D_{27}\cdot pS_1] + [S_3\cdot D_{27}\cdot S_3] \\ &+ [S_3\cdot D_{27}\cdot pS_3]) - k_{3b}^{+}[S_3]([D_{27}] + [S_1\cdot D_{27}] + [pS_1\cdot D_{27}] + [S_3\cdot D_{27}] \\ &+ [pS_3\cdot D_{27}]) + k_{3b}^{-}([D_{27}\cdot S_3] + [S_1\cdot D_{27}\cdot S_3] + [pS_1\cdot D_{27}\cdot S_3] \\ &+ [S_3\cdot D_{27}\cdot S_3] + [pS_3\cdot D_{27}\cdot S_3]) + d_3[pS_3] \end{aligned} \tag{4.29}$$

$$\begin{aligned}\frac{d[S_1\cdot D_{27}]}{dt} &= k_{1a}^{+}[S_1][D_{27}] - k_{1a}^{-}[S_1\cdot D_{27}] - q[S_1\cdot D_{27}] - k_{1b}^{+}[S_1][S_1\cdot D_{27}] \\ &+ k_{1b}^{-}[S_1\cdot D_{27}\cdot S_1] - k_{3b}^{+}[S_3][S_1\cdot D_{27}] + k_{3b}^{-}[S_1\cdot D_{27}\cdot S_3] \\ &+ k_{1b}^{-}[S_1\cdot D_{27}\cdot pS_1] + k_{3b}^{-}[S_1\cdot D_{27}\cdot pS_3] - \beta_{27}[S_1\cdot D_{27}] \\ &- \gamma_{27}([pS_1] + [pS_3])[S_1\cdot D_{27}] \end{aligned} \tag{4.30}$$

$$\begin{aligned}\frac{d[D_{27}\cdot S_1]}{dt} &= k_{1b}^{+}[S_1][D_{27}] - k_{1b}^{-}[D_{27}\cdot S_1] - q[D_{27}\cdot S_1] - k_{1a}^{+}[S_1][D_{27}\cdot S_1] \\ &+ k_{1a}^{-}[S_1\cdot D_{27}\cdot S_1] - k_{3a}^{+}[S_3][D_{27}\cdot S_1] + k_{3a}^{-}[S_3\cdot D_{27}\cdot S_1] \\ &+ k_{1a}^{-}[pS_1\cdot D_{27}\cdot S_1] + k_{3a}^{-}[pS_3\cdot D_{27}\cdot S_1] - \beta_{27}[D_{27}\cdot S_1] \\ &- \gamma_{27}([pS_1] + [pS_3])[D_{27}\cdot S_1] \end{aligned} \tag{4.31}$$

$$\begin{aligned}\frac{d[S_3\cdot D_{27}]}{dt} &= k_{3a}^{+}[S_3][D_{27}] - k_{3a}^{-}[S_3\cdot D_{27}] - q[S_3\cdot D_{27}] - k_{3b}^{+}[S_3][S_3\cdot D_{27}] \\ &+ k_{3b}^{-}[S_3\cdot D_{27}\cdot S_3] - k_{1b}^{+}[S_1][S_3\cdot D_{27}] + k_{1b}^{-}[S_3\cdot D_{27}\cdot S_1] \\ &+ k_{3b}^{-}[S_3\cdot D_{27}\cdot pS_3] + k_{1b}^{-}[S_3\cdot D_{27}\cdot pS_1] - \beta_{27}[S_3\cdot D_{27}] \\ &- \gamma_{27}([pS_1] + [pS_3])[S_3\cdot D_{27}] \end{aligned} \tag{4.32}$$

$$\begin{aligned}\frac{d[D_{27}\cdot S_3]}{dt} &= k_{3b}^{+}[S_3][D_{27}] - k_{3b}^{-}[D_{27}\cdot S_3] - q[D_{27}\cdot S_3] - k_{3a}^{+}[S_3][D_{27}\cdot S_3] \\ &+ k_{3a}^{-}[S_3\cdot D_{27}\cdot S_3] - k_{1a}^{+}[S_1][D_{27}\cdot S_3] + k_{1a}^{-}[S_1\cdot D_{27}\cdot S_3] \\ &+ k_{3a}^{-}[pS_3\cdot D_{27}\cdot S_3] + k_{1a}^{-}[pS_1\cdot D_{27}\cdot S_3] - \beta_{27}[D_{27}\cdot S_3] \end{aligned}$$

$$- \gamma_{27}([pS_1] + [pS_3])[D_{27} \cdot S_3] \tag{4.33}$$

$$\begin{aligned}\frac{d[pS_1 \cdot D_{27}]}{dt} &= -k_{1b}^{+}[pS_1 \cdot D_{27}][S_1] + k_{1b}^{-}[pS_1 \cdot D_{27} \cdot S_1] - k_{3b}^{+}[pS_1 \cdot D_{27}][S_3] \\ &+ k_{3b}^{-}[pS_1 \cdot D_{27} \cdot S_3] + q[S_1 \cdot D_{27}] - k_{1a}^{-}[pS_1 \cdot D_{27}] \\ &+ k_{1b}^{-}[pS_1 \cdot D_{27} \cdot pS_1] + k_{3b}^{-}[pS_1 \cdot D_{27} \cdot pS_3] - \beta_{27}[pS_1 \cdot D_{27}] \\ &- \gamma_{27}([pS_1] + [pS_3])[pS_1 \cdot D_{27}]\end{aligned} \tag{4.34}$$

$$\begin{aligned}\frac{d[D_{27} \cdot pS_1]}{dt} &= -k_{1a}^{+}[D_{27} \cdot pS_1][S_1] + k_{1a}^{-}[S_1 \cdot D_{27} \cdot pS_1] - k_{3a}^{+}[D_{27} \cdot pS_1][S_3] \\ &+ k_{3a}^{-}[S_3 \cdot D_{27} \cdot pS_1] + q[D_{27} \cdot S_1] - k_{1b}^{-}[D_{27} \cdot pS_1] \\ &+ k_{1a}^{-}[pS_1 \cdot D_{27} \cdot pS_1] + k_{3a}^{-}[pS_3 \cdot D_{27} \cdot pS_1] - \beta_{27}[D_{27} \cdot pS_1] \\ &- \gamma_{27}([pS_1] + [pS_3])[D_{27} \cdot pS_1]\end{aligned} \tag{4.35}$$

$$\begin{aligned}\frac{d[pS_3 \cdot D_{27}]}{dt} &= -k_{3b}^{+}[pS_3 \cdot D_{27}][S_3] + k_{3b}^{-}[pS_3 \cdot D_{27} \cdot S_3] - k_{1b}^{+}[pS_3 \cdot D_{27}][S_1] \\ &+ k_{1b}^{-}[pS_3 \cdot D_{27} \cdot S_1] + q[S_3 \cdot D_{27}] - k_{3a}^{-}[pS_3 \cdot D_{27}] \\ &+ k_{3b}^{-}[pS_3 \cdot D_{27} \cdot pS_3] + k_{1b}^{-}[pS_3 \cdot D_{27} \cdot pS_1] - \beta_{27}[pS_3 \cdot D_{27}] \\ &- \gamma_{27}([pS_1] + [pS_3])[pS_3 \cdot D_{27}]\end{aligned} \tag{4.36}$$

$$\begin{aligned}\frac{d[D_{27} \cdot pS_3]}{dt} &= -k_{3a}^{+}[D_{27} \cdot pS_3][S_3] + k_{3a}^{-}[S_3 \cdot D_{27} \cdot pS_3] - k_{1a}^{+}[D_{27} \cdot pS_3][S_1] \\ &+ k_{1a}^{-}[S_1 \cdot D_{27} \cdot pS_3] + q[D_{27} \cdot S_3] - k_{3b}^{-}[D_{27} \cdot pS_3] \\ &+ k_{3a}^{-}[pS_3 \cdot D_{27} \cdot pS_3] + k_{1a}^{-}[pS_1 \cdot D_{27} \cdot pS_3] - \beta_{27}[D_{27} \cdot pS_3] \\ &- \gamma_{27}([pS_1] + [pS_3])[D_{27} \cdot pS_3]\end{aligned} \tag{4.37}$$

$$\begin{aligned}\frac{d[S_1 \cdot D_{27} \cdot S_1]}{dt} &= k_{1a}^{+}[S_1][D_{27} \cdot S_1] - k_{1a}^{-}[S_1 \cdot D_{27} \cdot S_1] + k_{1b}^{+}[S_1 \cdot D_{27}][S_1] \\ &- k_{1b}^{-}[S_1 \cdot D_{27} \cdot S_1] - 2q[S_1 \cdot D_{27} \cdot S_1] - \beta_{27}[S_1 \cdot D_{27} \cdot S_1] \\ &- \gamma_{27}([pS_1] + [pS_3])[S_1 \cdot D_{27} \cdot S_1]\end{aligned} \tag{4.38}$$

$$\begin{aligned}\frac{d[pS_1 \cdot D_{27} \cdot S_1]}{dt} &= k_{1b}^{+}[pS_1 \cdot D_{27}][S_1] - k_{1b}^{-}[pS_1 \cdot D_{27} \cdot S_1] + q[S_1 \cdot D_{27} \cdot S_1] \\ &- q[pS_1 \cdot D_{27} \cdot S_1] - k_{1a}^{-}[pS_1 \cdot D_{27} \cdot S_1] - \beta_{27}[pS_1 \cdot D_{27} \cdot S_1] \\ &- \gamma_{27}([pS_1] + [pS_3])[pS_1 \cdot D_{27} \cdot S_1]\end{aligned} \tag{4.39}$$

$$\frac{d[S_1 \cdot D_{27} \cdot pS_1]}{dt} = k_{1a}^{+}[S_1][D_{27} \cdot pS_1] - k_{1a}^{-}[S_1 \cdot D_{27} \cdot pS_1] + q[S_1 \cdot D_{27} \cdot S_1]$$

$$- q[S_1 \cdot D_{27} \cdot pS_1] - k_{1b}^{-}[S_1 \cdot D_{27} \cdot pS_1] - \beta_{27}[S_1 \cdot D_{27} \cdot pS_1] - \gamma_{27}([pS_1] + [pS_3])[S_1 \cdot D_{27} \cdot pS_1] \tag{4.40}$$

$$\frac{d[pS_1 \cdot D_{27} \cdot pS_1]}{dt} = q([S_1 \cdot D_{27} \cdot pS_1] + [pS_1 \cdot D_{27} \cdot S_1]) - [pS_1 \cdot D_{27} \cdot pS_1](k_{1a}^{-} + k_{1b}^{-}) - \beta_{27}[pS_1 \cdot D_{27} \cdot pS_1] - \gamma_{27}([pS_1] + [pS_3])[pS_1 \cdot D_{27} \cdot pS_1] \tag{4.41}$$

$$\frac{d[S_3 \cdot D_{27} \cdot S_3]}{dt} = k_{3a}^{+}[S_3][D_{27} \cdot S_3] - k_{3a}^{-}[S_3 \cdot D_{27} \cdot S_3] + k_{3b}^{+}[S_3 \cdot D_{27}][S_3] - k_{3b}^{-}[S_3 \cdot D_{27} \cdot S_3] - 2q[S_3 \cdot D_{27} \cdot S_3] - \beta_{27}[S_3 \cdot D_{27} \cdot S_3] - \gamma_{27}([pS_1] + [pS_3])[S_3 \cdot D_{27} \cdot S_3] \tag{4.42}$$

$$\frac{d[pS_3 \cdot D_{27} \cdot S_3]}{dt} = k_{3b}^{+}[pS_3 \cdot D_{27}][S_3] - k_{3b}^{-}[pS_3 \cdot D_{27} \cdot S_3] + q[S_3 \cdot D_{27} \cdot S_3] - q[pS_3 \cdot D_{27} \cdot S_3] - k_{3a}^{-}[pS_3 \cdot D_{27} \cdot S_3] - \beta_{27}[pS_3 \cdot D_{27} \cdot S_3] - \gamma_{27}([pS_1] + [pS_3])[pS_3 \cdot D_{27} \cdot S_3] \tag{4.43}$$

$$\frac{d[S_3 \cdot D_{27} \cdot pS_3]}{dt} = k_{3a}^{+}[S_3][D_{27} \cdot pS_3] - k_{3a}^{-}[S_3 \cdot D_{27} \cdot pS_3] + q[S_3 \cdot D_{27} \cdot S_3] - q[S_3 \cdot D_{27} \cdot pS_3] - k_{3b}^{-}[S_3 \cdot D_{27} \cdot pS_3] - \beta_{27}[S_3 \cdot D_{27} \cdot pS_3] - \gamma_{27}([pS_1] + [pS_3])[S_3 \cdot D_{27} \cdot pS_3] \tag{4.44}$$

$$\frac{d[pS_3 \cdot D_{27} \cdot pS_3]}{dt} = q([S_3 \cdot D_{27} \cdot pS_3] + [pS_3 \cdot D_{27} \cdot S_3]) - [pS_3 \cdot D_{27} \cdot pS_3](k_{3a}^{-} + k_{3b}^{-}) - \beta_{27}[pS_3 \cdot D_{27} \cdot pS_3] - \gamma_{27}([pS_1] + [pS_3])[pS_3 \cdot D_{27} \cdot pS_3] \tag{4.45}$$

$$\frac{d[S_1 \cdot D_{27} \cdot S_3]}{dt} = k_{1a}^{+}[S_1][D_{27} \cdot S_3] - k_{1a}^{-}[S_1 \cdot D_{27} \cdot S_3] + k_{3b}^{+}[S_1 \cdot D_{27}][S_3] - k_{3b}^{-}[S_1 \cdot D_{27} \cdot S_3] - 2q[S_1 \cdot D_{27} \cdot S_3] - \beta_{27}[S_1 \cdot D_{27} \cdot S_3] - \gamma_{27}([pS_1] + [pS_3])[S_1 \cdot D_{27} \cdot S_3] \tag{4.46}$$

$$\frac{d[S_3 \cdot D_{27} \cdot S_1]}{dt} = k_{3a}^{+}[S_3][D_{27} \cdot S_1] - k_{3a}^{-}[S_3 \cdot D_{27} \cdot S_1] + k_{1b}^{+}[S_3 \cdot D_{27}][S_1] - k_{1b}^{-}[S_3 \cdot D_{27} \cdot S_1] - 2q[S_3 \cdot D_{27} \cdot S_1] - \beta_{27}[S_3 \cdot D_{27} \cdot S_1] - \gamma_{27}([pS_1] + [pS_3])[S_3 \cdot D_{27} \cdot S_1] \tag{4.47}$$

$$\frac{d[pS_1 \cdot D_{27} \cdot S_3]}{dt} = k_{3b}^{+}[pS_1 \cdot D_{27}][S_3] - k_{3b}^{-}[pS_1 \cdot D_{27} \cdot S_3] + q[S_1 \cdot D_{27} \cdot S_3]$$

$$
\begin{aligned}
&- q[pS_1 \cdot D_{27} \cdot S_3] - k_{1a}^{-}[pS_1 \cdot D_{27} \cdot S_3] - \beta_{27}[pS_1 \cdot D_{27} \cdot S_3] \\
&- \gamma_{27}([pS_1] + [pS_3])[pS_1 \cdot D_{27} \cdot S_3] \qquad (4.48)
\end{aligned}
$$

$$
\begin{aligned}
\frac{d[pS_3 \cdot D_{27} \cdot S_1]}{dt} &= k_{1b}^{+}[pS_3 \cdot D_{27}][S_1] - k_{1b}^{-}[pS_3 \cdot D_{27} \cdot S_1] + q[S_3 \cdot D_{27} \cdot S_1] \\
&- q[pS_3 \cdot D_{27} \cdot S_1] - k_{3a}^{-}[pS_3 \cdot D_{27} \cdot S_1] - \beta_{27}[pS_3 \cdot D_{27} \cdot S_1] \\
&- \gamma_{27}([pS_1] + [pS_3])[pS_3 \cdot D_{27} \cdot S_1] \qquad (4.49)
\end{aligned}
$$

$$
\begin{aligned}
\frac{d[S_1 \cdot D_{27} \cdot pS_3]}{dt} &= k_{1a}^{+}[S_1][D_{27} \cdot pS_3] - k_{1a}^{-}[S_1 \cdot D_{27} \cdot pS_3] + q[S_1 \cdot D_{27} \cdot S_3] \\
&- q[S_1 \cdot D_{27} \cdot pS_3] - k_{3b}^{-}[S_1 \cdot D_{27} \cdot pS_3] - \beta_{27}[S_1 \cdot D_{27} \cdot pS_3] \\
&- \gamma_{27}([pS_1] + [pS_3])[S_1 \cdot D_{27} \cdot pS_3] \qquad (4.50)
\end{aligned}
$$

$$
\begin{aligned}
\frac{d[S_3 \cdot D_{27} \cdot pS_1]}{dt} &= k_{3a}^{+}[S_3][D_{27} \cdot pS_1] - k_{3a}^{-}[S_3 \cdot D_{27} \cdot pS_1] + q[S_3 \cdot D_{27} \cdot S_1] \\
&- q[S_3 \cdot D_{27} \cdot pS_1] - k_{1b}^{-}[S_3 \cdot D_{27} \cdot pS_1] - \beta_{27}[S_3 \cdot D_{27} \cdot pS_1] \\
&- \gamma_{27}([pS_1] + [pS_3])[S_3 \cdot D_{27} \cdot pS_1] \qquad (4.51)
\end{aligned}
$$

$$
\begin{aligned}
\frac{d[pS_1 \cdot D_{27} \cdot pS_3]}{dt} &= q([S_1 \cdot D_{27} \cdot pS_3] + [pS_1 \cdot D_{27} \cdot S_3]) - [pS_1 \cdot D_{27} \cdot pS_3](k_{1a}^{-} + k_{3b}^{-}) \\
&- \beta_{27}[pS_1 \cdot D_{27} \cdot pS_3] - \gamma_{27}([pS_1] + [pS_3])[pS_1 \cdot D_{27} \cdot pS_3] \qquad (4.52)
\end{aligned}
$$

$$
\begin{aligned}
\frac{d[pS_3 \cdot D_{27} \cdot pS_1]}{dt} &= q([S_3 \cdot D_{27} \cdot pS_1] + [pS_3 \cdot D_{27} \cdot S_1]) - [pS_3 \cdot D_{27} \cdot pS_1](k_{3a}^{-} + k_{1b}^{-}) \\
&- \beta_{27}[pS_3 \cdot D_{27} \cdot pS_1] - \gamma_{27}([pS_1] + [pS_3])[pS_3 \cdot D_{27} \cdot pS_1] \qquad (4.53)
\end{aligned}
$$

$$
\begin{aligned}
\frac{d[pS_1]}{dt} &= k_{1a}^{-}([pS_1 \cdot D_{27}] + [pS_1 \cdot D_{27} \cdot S_1] + [pS_1 \cdot D_{27} \cdot pS_1] \\
&+ [pS_1 \cdot D_{27} \cdot S_3] + [pS_1 \cdot D_{27} \cdot pS_3]) + k_{1b}^{-}([D_{27} \cdot pS_1] \\
&+ [S_1 \cdot D_{27} \cdot pS_1] + [pS_1 \cdot D_{27} \cdot pS_1] + [S_3 \cdot D_{27} \cdot pS_1] \\
&+ [pS_3 \cdot D_{27} \cdot pS_1]) - d_1[pS_1] \qquad (4.54)
\end{aligned}
$$

$$
\begin{aligned}
\frac{d[pS_3]}{dt} &= k_{3a}^{-}([pS_3 \cdot D_{27}] + [pS_3 \cdot D_{27} \cdot S_3] + [pS_3 \cdot D_{27} \cdot pS_3] \\
&+ [pS_3 \cdot D_{27} \cdot S_1] + [pS_3 \cdot D_{27} \cdot pS_1]) + k_{3b}^{-}([D_{27} \cdot pS_3] \\
&+ [S_3 \cdot D_{27} \cdot pS_3] + [pS_3 \cdot D_{27} \cdot pS_3] + [S_1 \cdot D_{27} \cdot pS_3] \\
&+ [pS_1 \cdot D_{27} \cdot pS_3]) - d_3[pS_3] \qquad (4.55)
\end{aligned}
$$

Similarly to the HypIL-6 model, the terms in Equations (4.23) - (4.55) involving the parameter β_{27} apply only to the model under hypothesis 1 and the terms involving the parameter γ_{27} apply only to the model under hypothesis 2.

4.2 Experimental data

In this section, the experimental data which will be compared with the mathematical model outputs is presented and discussed. The experiments involved two types of cell, namely, retinal pigment epithelium 1 (RPE1) cells and T helper type 1 (Th-1) cells. For each cell type, the fluorescence intensity (FI) of antibodies for pSTAT1 and pSTAT3 were measured under three different experimental conditions: unstimulated, stimulated with 2 nM IL-27, and stimulated with 10 nM HypIL-6, and at 8 time points up to 180 minutes. Four replicates of each experiment were carried out for each cell type and stimulation condition and the raw data can be seen in Figure 4.4 for RPE1 cells and in Figure 4.5 for Th-1 cells.

4.2.1 Data normalisation

Given that the data have units of fluorescence intensity and the variables in the mathematical models have units of concentration (nM), both the data and model outputs must be normalised in order to be compared. Firstly it can be seen from the top row of Figures 4.4 and 4.5 that there is some FI detected for antibodies for pSTAT1 and pSTAT3 even when the cells are unstimulated. To account for this background fluorescence, a linear model was fitted to the data from unstimulated cells individually per cell type and STAT type. The value of this linear model at each time point was then subtracted from each of the four data points at each time point for the HypIL-6 and IL-27 data. Finally, the HypIL-6 and IL-27 data was normalised to the IL-27 data for each cell type and STAT type in the following way. Denoting by fl the experimental fluorescence intensity, $fl(r, i, tp, j, d)$ corresponds to the FI for the rth repeat, $r \in R = \{1, 2, 3, 4\}$ with

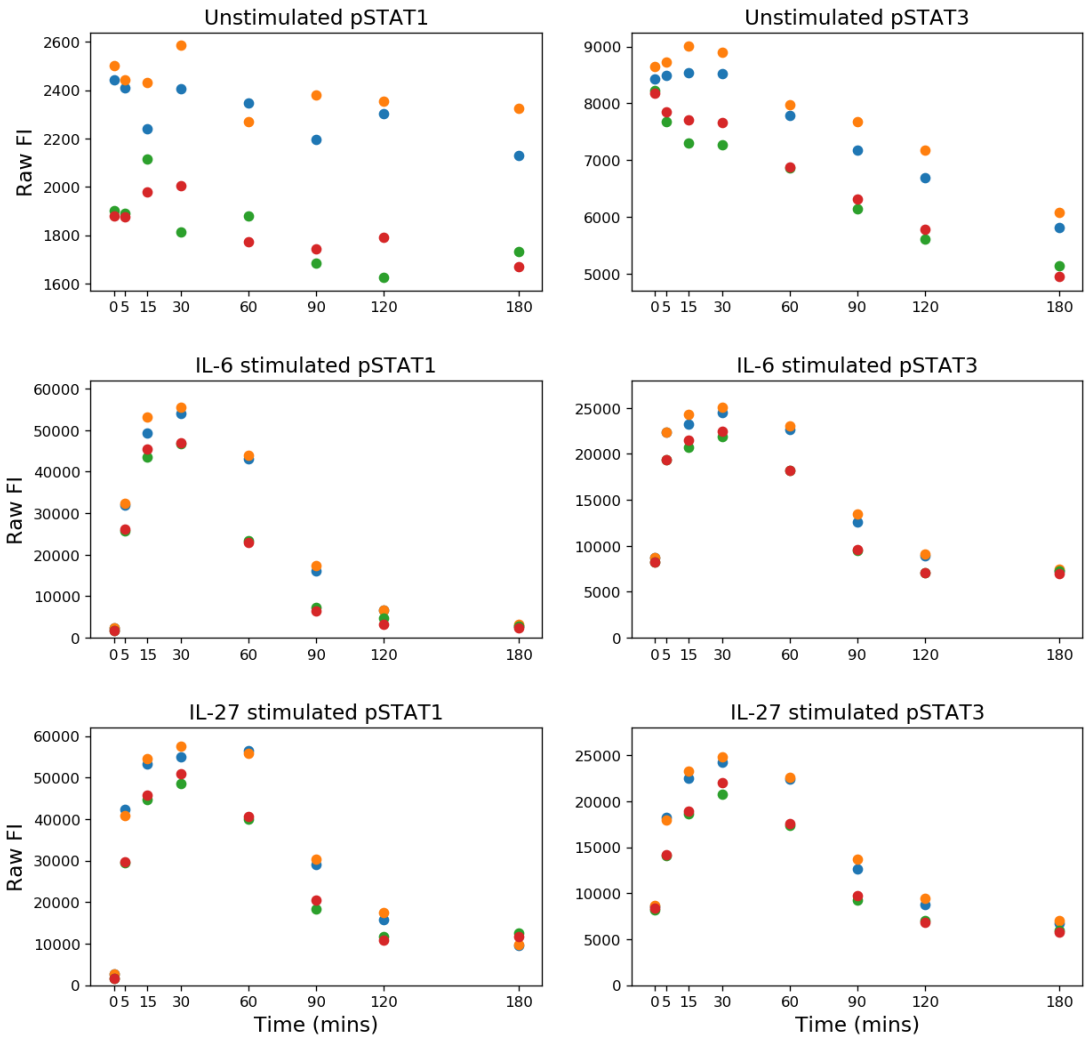

Figure 4.4: Raw FI data in unstimulated (**top row**), HypIL-6 stimulated (**middle row**), and IL-27 stimulated (**bottom row**) RPE1 cells. Each colour represents a different experimental replicate.

antibody for STATi, $i \in I = \{1, 3\}$ at time point

$$tp \in TP = \{0 \text{ min}, 5 \text{ min}, 15 \text{ min}, 30 \text{ min}, 60 \text{ min}, 90 \text{ min}, 120 \text{ min}, 180 \text{ min}\}$$

under stimulation by cytokine IL-j (HypIL-j when $j = 6$), with $j \in J = \{6, 27\}$ and in cell type $d \in D = \{\text{RPE1}, \text{Th-1}\}$. Each data point, $data(r, i, tp, j, d)$, to be used in the Bayesian inference and Bayesian model selection was then obtained

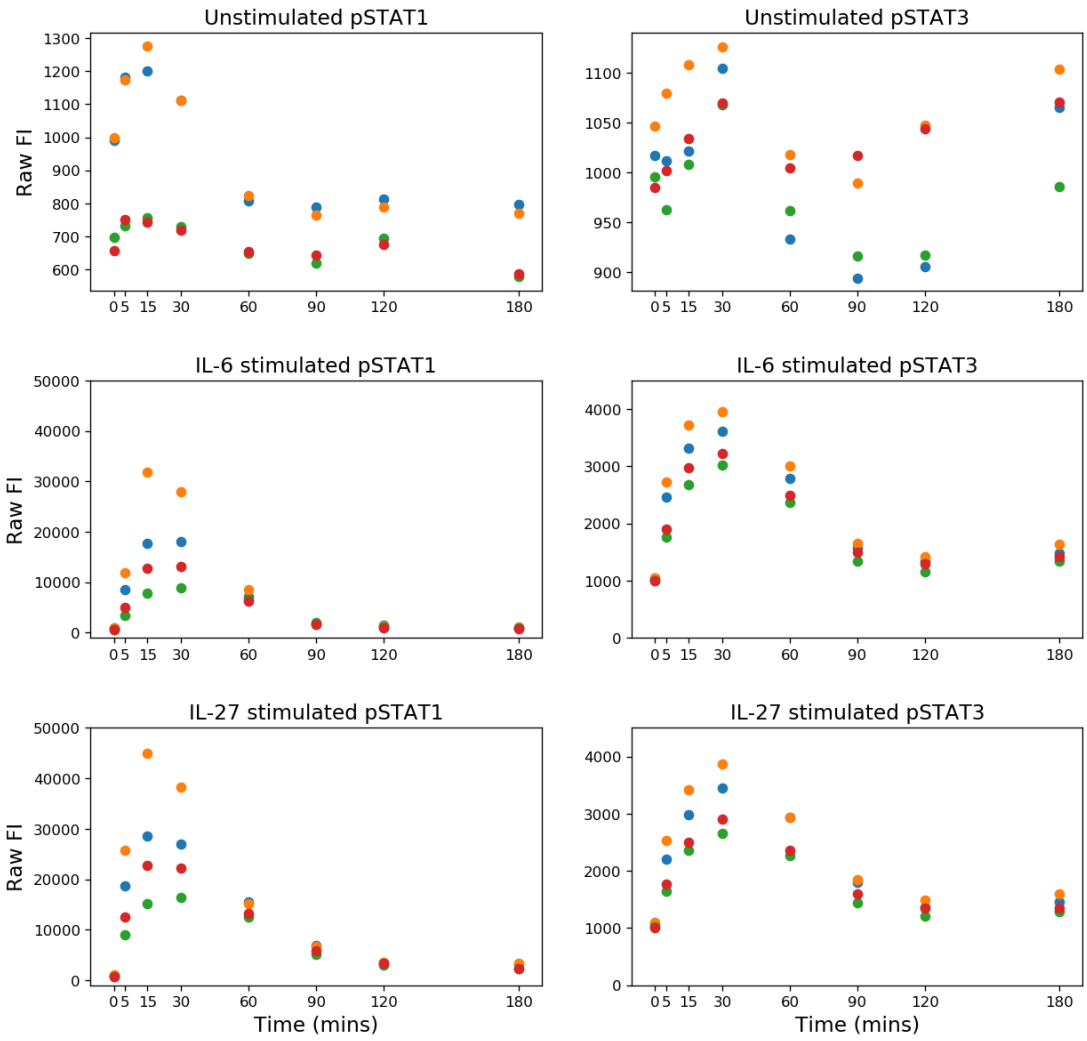

Figure 4.5: Raw FI data in unstimulated (**top row**), HypIL-6 stimulated (**middle row**), and IL-27 stimulated (**bottom row**) Th-1 cells. Each colour represents a different experimental replicate.

from $fl(r, i, tp, j, d)$ with the following normalisation,

$$data(r, i, tp, j, d) = \frac{fl(r, i, tp, j, d)}{fl(r, i, tp = 30 \text{ min}, j = 27, d)}. \tag{4.56}$$

That is, the normalisation has been chosen to be the time point 30 minutes with IL-27 stimulation. The data points for IL-27 stimulation at time 30 minutes were chosen as normalisation in Equation (4.56), since they correspond to the

maximal experimental value, and thus the data, when normalised in this way, were transformed to a scale in the interval $[0, 1]$. The mean of the four repeats of the normalised data is then given by

$$\text{MFI} = \mu_{data}(i, tp, j, d) = \frac{1}{4}\sum_{r=1}^{4} data(r, i, tp, j, d), \tag{4.57}$$

and the standard deviation (SD) by,

$$\sigma_{data}(i, tp, j, d) = \sqrt{\frac{1}{4}\sum_{r=1}^{4}(data(r, i, tp, j, d) - \mu_{data}(i, tp, j, d))^2}.$$

The mean and SD of the FI data can be seen in Figure 4.6 for the RPE1 cells (top row) and Th-1 cells (bottom row). From this figure, it can clearly be observed that there is a significant difference between HypIL-6 stimulated pSTAT1 signalling and IL-27 stimulated pSTAT1 signalling, but there is no significant difference in pSTAT3 signalling between the two cytokines.

4.2.2 Model output and normalisation

Since the experimental outputs are levels of pSTAT1 and pSTAT3 as a function of time under HypIL-6 and IL-27 stimulation (Figures 4.4 and 4.5), two model outputs of interest are considered for the HypIL-6 and IL-27 mathematical models, which are proportional to the experimental data in Figures 4.4 and 4.5. These outputs are, the sum of all molecular species (variables) containing phosphorylated STAT1 (free or bound) ($[pS_1]^{T,j}$, for $j \in \{6, 27\}$) and the sum of all species (variables) containing phosphorylated STAT3 (free or bound) ($[pS_3]^{T,j}$, for $j \in \{6, 27\}$). The total concentrations of the two model outputs of interest at any time t are defined by the following equations, where T denotes the total concentration of the given molecular species:

$$\begin{aligned}
[pS_1]^{T,6}(t) &= [D_6 \cdot pS_1](t) + [pS_1 \cdot D_6 \cdot S_1](t) + 2[pS_1 \cdot D_6 \cdot pS_1](t) \\
&\quad + [pS_1 \cdot D_6 \cdot S_3](t) + [pS_1 \cdot D_6 \cdot pS_3](t) + [pS_1](t), \\
[pS_3]^{T,6}(t) &= [D_6 \cdot pS_3](t) + [pS_3 \cdot D_6 \cdot S_3](t) + 2[pS_3 \cdot D_6 \cdot pS_3](t)
\end{aligned} \tag{4.58}$$

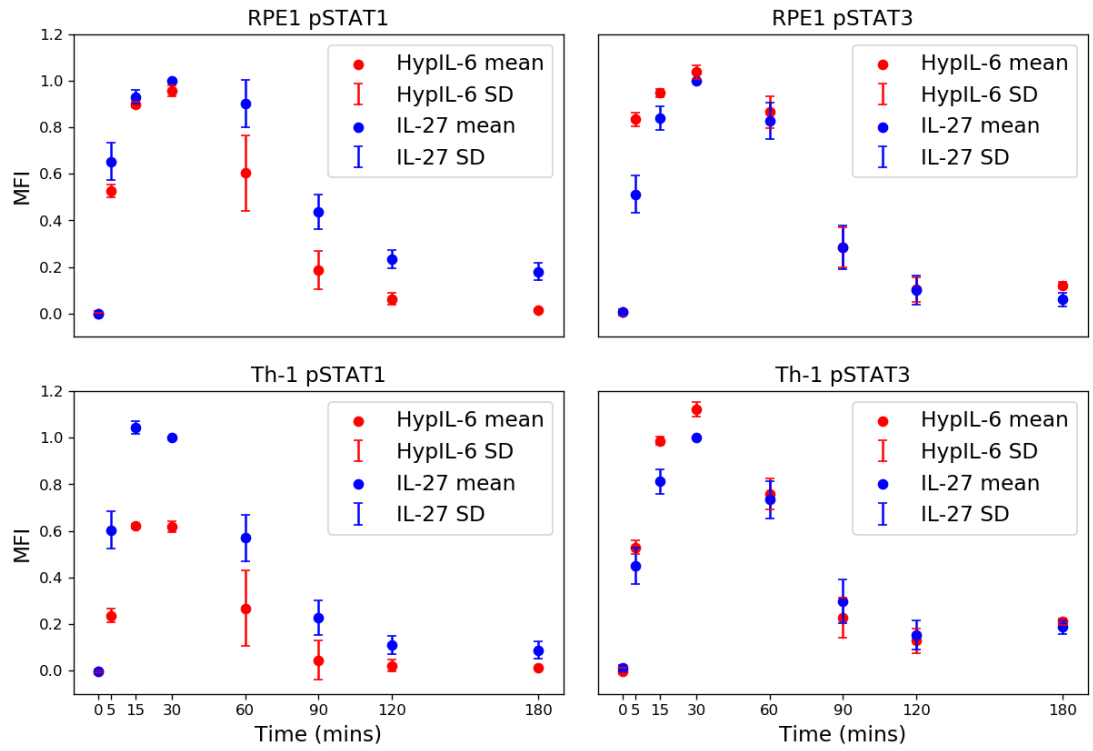

Figure 4.6: Top row: Mean and SD of the normalised FI data from RPE1 cells. **Bottom row:** Mean and SD of the normalised FI data from Th-1 cells. For both rows, the MFI is computed using Equation 4.57.

$$+ [pS_3 \cdot D_6 \cdot S_1](t) + [pS_3 \cdot D_6 \cdot pS_1](t) + [pS_3](t), \tag{4.59}$$

for the HypIL-6 model, and

$$\begin{aligned}[pS_1]^{T,27}(t) = {} & [pS_1 \cdot D_{27}](t) + [D_{27} \cdot pS_1](t) + [pS_1 \cdot D_{27} \cdot S_1](t) + [S_1 \cdot D_{27} \cdot pS_1](t) \\ & + 2[pS_1 \cdot D_{27} \cdot pS_1](t) + [pS_1 \cdot D_{27} \cdot S_3](t) + [S_3 \cdot D_{27} \cdot pS_1](t) \\ & + [pS_1 \cdot D_6 \cdot pS_3](t) + [pS_3 \cdot D_6 \cdot pS_1](t) + [pS_1](t), \end{aligned} \tag{4.60}$$

$$\begin{aligned}[pS_3]^{T,27}(t) = {} & [pS_3 \cdot D_{27}](t) + [D_{27} \cdot pS_3](t) + [pS_3 \cdot D_{27} \cdot S_3](t) + [S_3 \cdot D_{27} \cdot pS_3](t) \\ & + 2[pS_3 \cdot D_{27} \cdot pS_3](t) + [pS_3 \cdot D_{27} \cdot S_1](t) + [S_1 \cdot D_{27} \cdot pS_3](t) \\ & + [pS_1 \cdot D_6 \cdot pS_3](t) + [pS_3 \cdot D_6 \cdot pS_1](t) + [pS_3](t), \end{aligned} \tag{4.61}$$

for the IL-27 model. Denoting by sim the mathematical model output, $sim(i, tp, j, d)$ corresponds to the model output for STATi, $i \in I = \{1, 3\}$ at time point

$$tp \in TP = \{0 \text{ min}, 5 \text{ min}, 15 \text{ min}, 30 \text{ min}, 60 \text{ min}, 90 \text{ min}, 120 \text{ min}, 180 \text{ min}\}$$

in the IL-j (HypIL-j when $j = 6$) mathematical model, with $j \in J = \{6, 27\}$, when considering cell type $d \in D = \{\text{RPE1}, \text{Th-1}\}$. In order to compare the model output, $sim(i, tp, j, d)$, with the data, the output was normalised in the same way as the data, that is,

$$sim(i, tp, j, d) = \frac{[pS_i]^{T,j}(tp, d)}{[pS_i]^{T,27}(30 \text{ min}, d)},$$

where $[pS_i]^{T,j}(tp, d)$ denotes the total, T, concentration of phosphorylated STATi at time point tp (see Equations (4.58) - (4.61)) when considering cell type d and cytokine stimulation $j \in J = \{6, 27\}$. Once datasets and mathematical outputs have been normalised and can be compared, there is a need to quantify how close (or not) they are. To this end, one can make use of a quantitative measure, called a distance, and denoted by $\delta(sim, data)$. In this case, a generalisation of the Euclidean distance has been chosen, where

$$[\delta^d(sim, data)]^2 = \sum_{i \in I} \sum_{tp \in TP} \sum_{j \in J} [sim(i, tp, j, d) - \mu_{data}(i, tp, j, d)]^2, \qquad (4.62)$$

$d \in D = \{\text{RPE1}, \text{Th-1}\}$.

4.3 Modelling of the HypIL-6 and IL-27 pathways

With a choice of distance measure in hand, the primary aims of the modelling effort were as follows.

1. To determine, via a Bayesian model selection, which of the hypotheses relating to receptor internalisation/degradation, was most likely, given the data.

2. To use Bayesian parameter inference to obtain posterior distributions for the parameters in the mathematical models under the most likely hypothesis, in order to learn about which specific rate constants, and therefore reactions, were causing the differential signalling by pSTAT1 under the different cytokines.

3. To use the parametrised mathematical models to make predictions about pSTAT signalling under changes in receptor and STAT concentrations, relevant to different disease scenarios.

In this section, points 1 and 2 are addressed.

4.3.1 Prior distributions

Bayesian methods, such as Bayesian model selection and parameter inference, take into account prior beliefs about the parameter values in the models. In this section, a prior distribution is defined for each of the model parameters in Table 4.1, based either on information from the literature or independent experimental data.

Receptor-ligand kinetic parameters

From Biacore affinity measurements (Murphy *et al.*, 2006), the group at the University of Dundee obtained the following rates for the binding and unbinding of the cytokines to and from GP130 and IL-27Rα,

- HypIL-6 binding to GP130: $r_{1,6}^{+} = 9.86 \times 10^{-4}$ nM^{-1}s^{-1},
- HypIL-6 unbinding from GP130: $r_{1,6}^{-} = 1.26 \times 10^{-4}$ s^{-1},
- IL-27 binding to IL-27Rα: $r_{1,27}^{+} = 4.55 \times 10^{-3}$ nM^{-1}s^{-1}, and
- IL-27 unbinding from IL-27Rα: $r_{1,27}^{-} = 1.50 \times 10^{-3}$ s^{-1}.

Given that there is some uncertainty in the experimental data, the prior distributions used for these parameters were 10^{r} where r was sampled from a normal distribution with mean equal to the logarithm base 10 of the experimental values and standard deviation equal to 50% of the logarithm base 10 of the experimental

value. Using logarithms and sampling the exponent from a normal distribution, allows for equal prior sampling over several orders of magnitude. The prior distributions for these parameters were therefore,

- $r^+_{1,6}$ prior $= 10^r$, where $r \sim N(-3, 1.5)$,
- $r^-_{1,6}$ prior $= 10^r$, where $r \sim N(-3.9, 1.96)$,
- $r^+_{1,27}$ prior $= 10^r$, where $r \sim N(-2.34, 1.17)$, and
- $r^-_{1,27}$ prior $= 10^r$, where $r \sim N(-2.82, 1.41)$.

Dimerisation parameters

The rate of dimerisation ($r^+_{2,j}$) and dissociation of the dimer ($r^-_{2,j}$), for $j \in \{6, 27\}$, are difficult parameters to measure experimentally and hence the priors for these rates were based on values from the literature relating to a different receptor, EGFR, which has been extensively studied. Rate constants for the dimerisation and dissociation of two molecules of ligand-bound EGFR, are given by Kozer *et al.* (2013a). Since these values are not specific to either GP130 or IL-27Rα, there is considerable uncertainty in these parameter values, and thus uniform distributions were used for the priors, centred around the EGFR values. The prior distribution for $r^+_{2,j}$ was therefore 10^r, where $r \sim Unif(-2, 3)$ and the prior distribution for $r^-_{2,j}$ was 10^r, where $r \sim Unif(-3, 1)$ for $j \in \{6, 27\}$.

STAT-dimer interaction parameters

Here, the rates of STAT binding and unbinding to the receptors in the dimers, k^+_{ia}, k^-_{ia}, k^+_{ib} and k^-_{ib} for $i \in \{1, 3\}$, are assigned a prior distribution. These rates are difficult to determine experimentally and so again, the priors were informed using values from the literature. In particular, there are references in the literature of the dissociation constant, K_d, for the interaction between STAT1/3 and GP130 (as well as another receptor, namely the Interferon-gamma receptor 1) (Wiederkehr-Adam *et al.*, 2003), where the K_d value is defined as the ratio between the rate at which the STAT dissociates the receptor and the rate at which the STAT associates the receptor. One can then estimate that the K_d value should lie within the range $[10^1, 10^5]$ nM. A prior distribution was defined

for the dissociation rates k_{ia}^- and k_{ib}^- for $i \in \{1,3\}$ of 10^r, where $r \sim Unif(-2,1)$ based on discussion with experimentalists. A uniform distribution can then be found for the association constants k_{ia}^+ and k_{ib}^+ for $i \in \{1,3\}$, by taking the ratio of the dissociation rates range and the K_d value range. The prior distributions for the STAT-dimer interaction parameters were then defined as,

- k_{ia}^+, k_{ib}^+ for $i \in \{1,3\}$ prior $= 10^r$, where $r \sim Unif(-7,1)$, and
- k_{ia}^-, k_{ib}^- for $i \in \{1,3\}$ prior $= 10^r$, where $r \sim Unif(-2,1)$.

STAT phosphorylation parameter

The rate of STAT phosphorylation, q, is uncertain and cannot be estimated experimentally but is assumed to be relatively fast, based on experimental observations. Hence, a large range was defined for this rate of $[10^{-3}, 10^2]$ s^{-1}. The prior distribution for this parameter was therefore 10^r, where $r \sim Unif(-3,2)$.

pSTAT dephosphorylation parameters

Estimates of the pSTAT dephosphorylation parameters d_1 and d_3, can be made using experimental data from kinetic experiments, which again measured a FI corresponding to total pSTAT1 and total pSTAT3, but where after 15 minutes a JAK inhibitor, Tofacitinib, was added to the cells. Tofacitinib is a small molecule reversible inhibitor which competes with ATP for the binding to JAK molecules. Tofacitinib lacks a triphosphate group and hence, once bound to a JAK molecule, inhibits phosphorylation and activation of the the JAK (Hodge *et al.*, 2016). Upon JAK inhibition, the receptor dimers can no longer recruit and phosphorylate the STAT molecules and hence in the fluorescence readout, after 15 minutes, only dephosphorylation of the STAT molecules is measured. The raw data from these experiments can be seen in the top row of subplots in Figure 4.7.

Given that after 15 minutes in these experiments, any molecules comprised of a receptor dimer bound to a phosphorylated STAT molecule cannot be formed, it can be assumed that the concentration of all such molecules is 0 nM. The differential equations for $[pS_1]$ and $[pS_3]$ after 15 minutes in both the HypIL-6

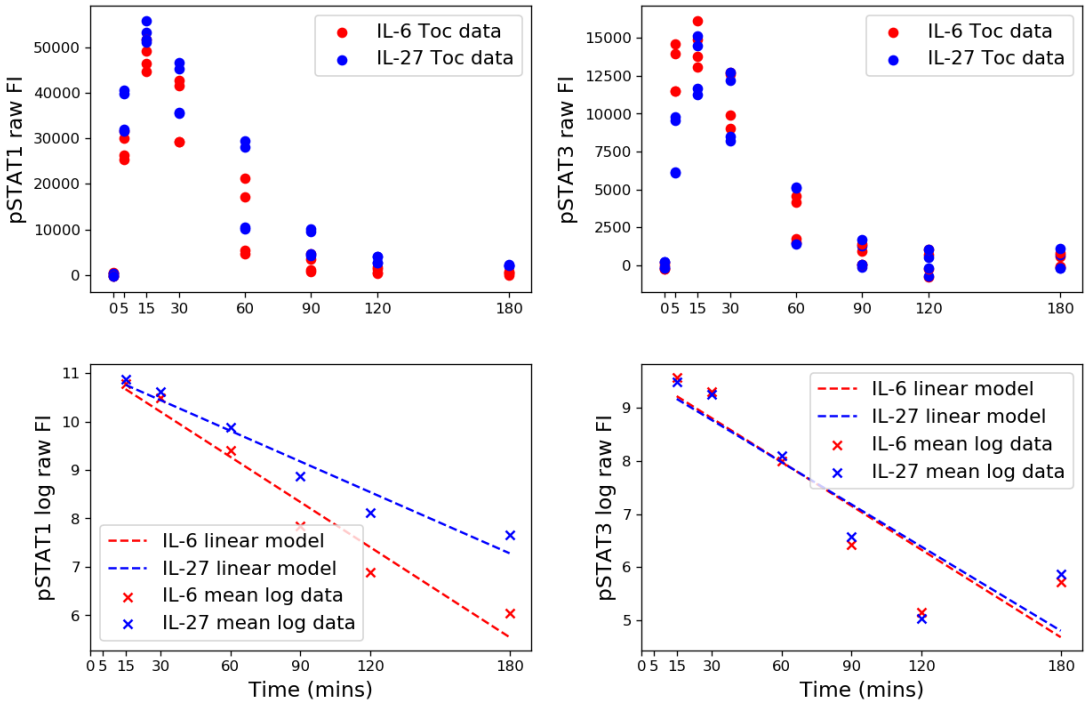

Figure 4.7: Top row: Raw FI of antibodies for pSTAT1 and pSTAT3 under stimulation with HypIL-6 and IL-27, where a JAK inhibitor, Tofacitinib, was added after 15 minutes. **Bottom row:** Linear model fitted to the logarithm of the mean raw data, from 15 to 180 minutes.

and IL-27 mathematical models then become

$$\frac{d[pS_1](t)}{dt} = -d_1[pS_1](t) \text{ and,}$$

$$\frac{d[pS_3](t)}{dt} = -d_3[pS_3](t),$$

for which the solutions are

$$[pS_1](t) = [pS_1](15)e^{-d_1 t} \text{ and,}$$

$$[pS_3](t) = [pS_3](15)e^{-d_3 t},$$

where $[pS_1](15)$ and $[pS_3](15)$ are the concentrations of pS_1 and pS_3 at time 15

minutes. Taking logarithms, one can write that

$$\log([pS_1](t)) = \log([pS_1](15)) - d_1 t \text{ and,}$$

$$\log([pS_3](t)) = \log([pS_3](15)) - d_3 t,$$

and hence one can estimate the parameters d_1 and d_3 as the slope of a linear model fitted to the logarithm of the mean data points after 15 minutes. Such linear models are shown, plotted with the logarithm of the mean data points, in the bottom row of subplots in Figure 4.7. From the linear models, the parameters for HypIL-6 stimulation can be estimated as

$$d_1 = 5.2 \times 10^{-4} \text{ s}^{-1} \text{ and,}$$
$$d_3 = 4.6 \times 10^{-4} \text{ s}^{-1},$$

and for IL-27 stimulation as

$$d_1 = 3.5 \times 10^{-4} \text{ s}^{-1} \text{ and,}$$
$$d_3 = 4.4 \times 10^{-4} \text{ s}^{-1}.$$

It can be seen that the estimates for these parameters are very similar for both cytokines, justifying the use of only one rate of pSTATi desphosphorylation, for $i \in \{1, 3\}$. Given the rates derived here and that there is some variation in the data around the linear model, the prior distribution for both d_1 and d_3 was defined as 10^r, where $r \sim Unif(-5, -2)$.

Receptor internalisation/degradation parameters

In the model, β_j or γ_j for $j \in \{6, 27\}$ (depending on the hypothesis), represents a rate of internalisation or degradation of any species involving a molecule of either receptor type. Some information about such rates can be gained using data from western blot experiments in which cells were treated with cycloheximide treatment. Cycloheximide blocks protein neosynthesis and gives an idea of basal protein turnover. In combination with HypIL-6 stimulation, it gives an estimate of the ligand-induced effect on GP130 degradation. It is assumed that in this

experimental set-up, protein degradation is the only process occurring and hence one can estimate the rate of degradation in the same way as the dephosphorylation rates were estimated. The logarithm of the mean turnover data from the cycloheximide experiments is taken, and a linear model fitted to this data. The rate constant for degradation of GP130, denoted k_{deg}, can then be read off as the slope of this linear model. The raw data can be seen in the left hand subplot of Figure 4.8 and the linear model fit to the log mean turnover data can be seen in the right hand subplot of the same figure. The linear fit allows one to conclude that $k_{deg} = 1.1 \times 10^{-4}$ s^{-1}.

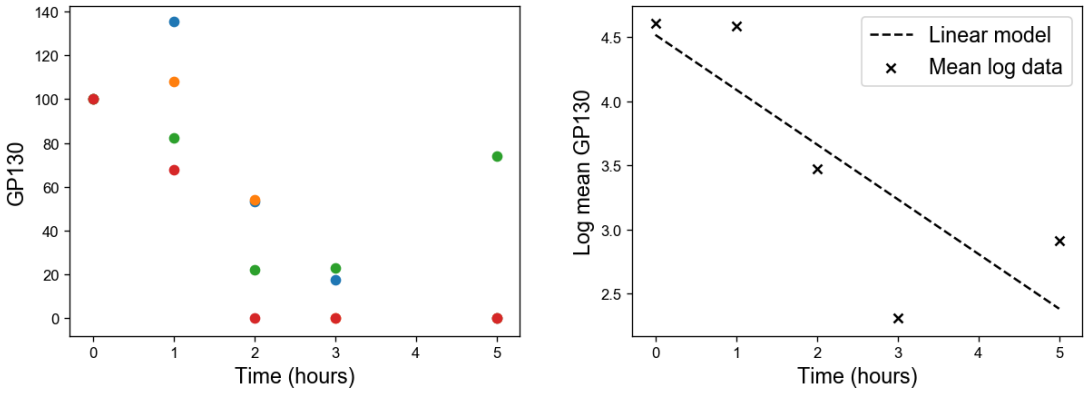

Figure 4.8: Left: Raw degradation data of GP130 under stimulation with HypIL-6 and treatment with cycloheximide. **Right:** A linear model fitted to the logarithm of the mean of the data.

Given that this parameter accounts for degradation of GP130 only (and not internalisation) and that similar data for IL-27Rα is not available, there is a lot of uncertainty in using this value as an estimate for β_j or γ_j ($j \in \{6, 27\}$) and hence a wide uniform distribution is used for the prior for both parameters. The prior distribution for β_j and γ_j was therefore 10^r, where $r \sim Unif(-5, -1)$ and $j \in \{6, 27\}$.

Receptor initial concentrations

Through total internal reflection fluorescence (TIRF) microscopy experiments, it was estimated that the numbers of each receptor type per square μm (micrometre) were,

- GP130 = 1.5 μm^{-2}, and
- IL-27Rα = 4 μm^{-2}.

Based on a cell surface area of 1600 μm^2 (Puck *et al.*, 1956), the per cell copy number for each receptor was given by

- GP130 = 2400 cell^{-1}, and
- IL-27Rα = 6400 cell^{-1}.

Given that the variables in the mathematical model have units of nM concentration, these values must be converted to a concentration by considering the volume in which the receptors are diffusing. It was assumed that the receptors were diffusing on the cell membrane and hence a depth 0.2 μm (approximately the length of a receptor molecule) into the cell from the surface was considered. Using the surface area of a cell, one can compute the radius of the cell, assuming that cells are spherical, as

$$r = \sqrt{\frac{1600}{4\pi}} = 11.28 \ \mu\text{m}.$$

The volume of the whole cell is therefore,

$$V_1 = \frac{4}{3}\pi 11.28^3 = 6012 \ \mu\text{m}^3.$$

One can then compute the volume of a smaller sphere, with radius $11.28 - 0.2 = 11.08$ μm as,

$$V_2 = \frac{4}{3}\pi 11.08^3 = 5698 \ \mu\text{m}^3.$$

Finally the volume of the cell in which the receptors can diffuse is the difference between the two computed volumes, *i.e.* $V_3 = V_1 - V_2 = 314$ μm^3. A depiction of a cell, showing the area in which receptor molecules are assumed to be diffusing can be seen in Figure 4.9.

In order to compute a concentration of receptors with units nM, firstly V_3 is converted to have units of L (litres), where $V_3 = 3.14 \times 10^{-13}$ L. Then, denoting

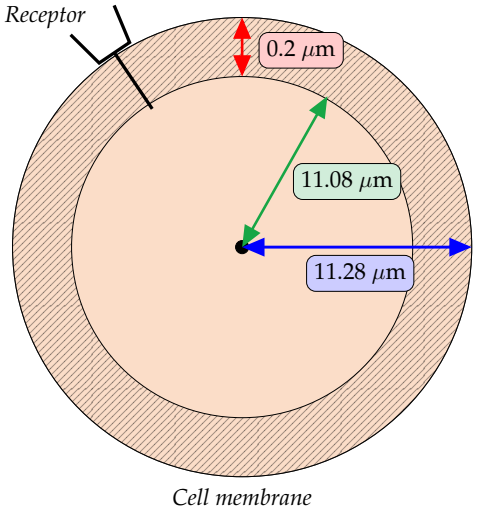

Figure 4.9: A depiction of a cell showing a receptor molecule whose tail crosses the cell membrane and protrudes into the cell to a depth of 0.2 μm. The dashed area is then the volume in which it is assumed that receptor molecules can diffuse. Figure not to scale.

by $[R_1](0)$ and $[R_2](0)$ the concentrations of GP130 and IL-27Rα, respectively, one can find that

$$[R_1](0) = \frac{2400}{3.14 \times 10^{-13} \times 6.022 \times 10^{23}} = 1.27 \times 10^{-8} \text{ M} = 12.7 \text{ nM}, \quad \text{and}$$
$$[R_2](0) = \frac{6400}{3.14 \times 10^{-13} \times 6.022 \times 10^{23}} = 3.38 \times 10^{-8} \text{ M} = 33.8 \text{ nM},$$

where 6.022×10^{23} is Avogadro's number with units mol^{-1}. Normal distributions were then used as the priors for these concentrations, where the mean for each distribution was the value computed above and the standard deviation was 50% of this value, and hence $[R_1](0) \sim N(12.7, 6.35)$ and $[R_2](0) \sim N(33.8, 16.9)$. These distributions were truncated so that the value of either receptor initial concentration must be positive.

STAT initial concentrations

The STAT initial concentrations were taken from the literature (Itzhak *et al.*, 2016) as $[S_1](0) \approx 300$ nM and $[S_3](0) \approx 400$ nM, and hence normal distributions were used as the priors for these parameters where, $[S_1](0) \sim N(300, 100)$ and $[S_3](0) \sim N(400, 100)$.

Summary of prior distributions

In summary, the prior distribution for each parameter is given in Table 4.2 and is plotted in Figure 4.10. In all of the modelling in this chapter, the cytokine initial concentrations are kept fixed at their experimental concentrations, $[L_6](0) = 10$ nM and $[L_{27}](0) = 2$ nM.

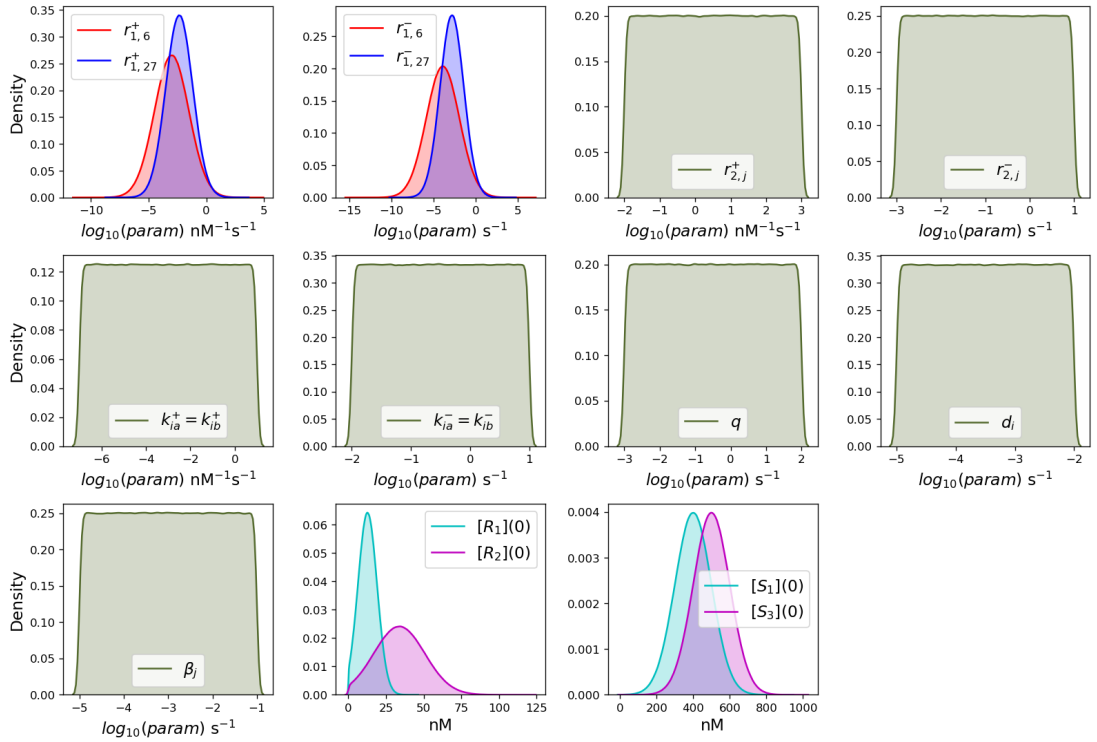

Figure 4.10: Prior distributions for the parameters in the mathematical model where $j \in \{6, 27\}$ and $i \in \{1, 3\}$.

Parameter/IC	Prior
$r^+_{1,6}$	10^r, where $r \sim N(-3, 1.5)$
$r^-_{1,6}$	10^r, where $r \sim N(-3.9, 1.96)$
$r^+_{1,27}$	10^r, where $r \sim N(-2.34, 1.17)$
$r^-_{1,27}$	10^r, where $r \sim N(-2.82, 1.41)$
$r^+_{2,j}$ for $j \in \{6, 27\}$	10^r, where $r \sim Unif(-2, 3)$
$r^-_{2,j}$ for $j \in \{6, 27\}$	10^r, where $r \sim Unif(-3, 1)$
k^+_{ia}, k^+_{ib} for $i \in \{1, 3\}$	10^r, where $r \sim Unif(-7, 1)$
k^-_{ia}, k^-_{ib} for $i \in \{1, 3\}$	10^r, where $r \sim Unif(-2, 1)$
q	10^r, where $r \sim Unif(-3, 2)$
d_i for $i \in \{1, 3\}$	10^r, where $r \sim Unif(-5, -2)$
β_j for $j \in \{6, 27\}$	10^r, where $r \sim Unif(-5, -1)$
γ_j for $j \in \{6, 27\}$	10^r, where $r \sim Unif(-5, -1)$
$[R_1](0)$	$N(12.7, 6.35)$
$[R_2](0)$	$N(33.8, 16.9)$
$[S_1](0)$	$N(300, 100)$
$[S_3](0)$	$N(400, 100)$

Table 4.2: Summary of the prior distributions for each of the parameters in the mathematical models.

4.3.2 Structural identifiability analysis

Before any parameter inference is carried out for the mathematical models, one should ensure that all parameters are structurally identifiable, meaning that they can be independently inferred (given the data available and any known initial conditions). To this end, a structural identifiability analysis was carried out using the method proposed by Castro & de Boer (2020) which is based on scale invariance of the equations. Structural identifiability is limited by the model as opposed to the quality of the data, and hence should be carried out in order to test the applicability of the model before parameter estimation. The method involves writing each ODE in the system as functionally independent terms, where for

example the terms ax_1x_2 and bx_1x_2 are not functionally independent (for variables x_1, x_2 and parameters a, b), whereas the terms ax_1x_2 and bx_1x_3 are independent. Consider a general model with λ_i parameters for $i = 1, \ldots, m$ and x_j variables, where the first r variables are observed and the remaining $j = r+1, \ldots, n$ variables are unobserved. The ODE model can then be written as

$$\frac{dx_j}{dt} = f_j(x_1, \ldots, x_r, x_{r+1}, \ldots, x_n; \lambda_1, \ldots, \lambda_m), \quad \text{for } j = 1, \ldots, n.$$

The first step in the method is to decompose each f_j into M functionally independent summands, f_{jk}, where

$$f_j(x_1, \ldots, x_r, x_{r+1}, \ldots, x_n; \lambda_1, \ldots, \lambda_m) = \sum_{k=1}^{M} f_{jk}(\tilde{x}_k, \tilde{\lambda}_k), \tag{4.63}$$

with f_{jk} functionally independent of f_{jl} for every $k \neq l$. In Equation (4.63), $\tilde{x}_k$ and $\tilde{\lambda}_k$ denote the subset of variables and parameters in the function f_{jk}. All parameters and all unobserved variables are then scaled by unknown scaling factors, u, so that

$$\begin{aligned} \lambda_i &\to u_{\lambda_i}\lambda_i \quad i = 1, \ldots, m \\ x_j &\to u_{x_j}x_j \quad j = r+1, \ldots, n. \end{aligned}$$

Having done this, each functionally independent term f_{jk} in each of the j ODEs can be equated to its scaled version via the formula

$$f_{jk}(\tilde{x}, \tilde{\lambda}) = \frac{1}{u_{x_j}} f_{jk}(u_{\tilde{x}}\tilde{x}, u_{\tilde{\lambda}}\tilde{\lambda}), \tag{4.64}$$

where $u_{x_j} = 1$ for $1 \leq j \leq r$. Finally, the identifiability equations given by (4.64) are solved simultaneously, and only parameters with $u_{\lambda_i} = 1$ are structurally identifiable. Likewise, the variables with $u_{x_j} = 1$ are observable. Full details of the method, including a proof are given by Castro & de Boer (2020). The structural identifiability method was applied to both the HypIL-6 and IL-27 mathematical models under each hypothesis and it was found that all parameters and initial conditions in all four models were structurally identifiable. Details of the method

applied to the HypIL-6 model under hypothesis 1 are given in Appendix A and details for the other three models are not presented in this thesis but follow very similar analysis to that of the HypIL-6 hypothesis 1 model.

4.3.3 Bayesian model selection

Having defined the mathematical models to describe the experimental systems under stimulation with HypIL-6 and IL-27, and determined that all parameters are structurally identifiable, in this section a Bayesian model selection is applied, as introduced in Section 2.5.3, Algorithm 5, to determine which of the model hypotheses relating to the internalisation/degradation of receptor molecules is most likely given the data. The RPE1 cell MFI data (top row of Figure 4.6 and defined by Equation 4.57) were used for the model selection, since the estimates of receptor and STAT numbers were more reliable in this cell type. The distance measure used to compare the model and the data was that given by Equation (4.62). At the first iteration of the model selection algorithm, the parameters were sampled from the prior distributions given in Table 4.2. The model (hypothesis) selection algorithm was run for a sample of size $N = 10^4$ and $Z = 15$ iterations. The following sequence of distance threshold values, ε_z for $z = 1, \ldots, Z$ was used,

$$\{100, 10, 5, 3, 2.5, 2.25, 2, 1.75, 1.5, 1.25, 1.1, 1, 0.9, 0.8, 0.7\},$$

where a particle (parameter set) at iteration z was accepted if it resulted in a value of the distance measure, $\delta^d(sim, data)$, less than the corresponding threshold value ε_z. A uniform perturbation kernel (Filippi *et al.*, 2013) was used to perturb the parameters sampled at each iteration z, for $z = 1, \ldots, 15$. The relative probability of each model hypothesis was computed for each iteration and the results are given in Figure 4.11. From the figure, it can be seen that, as ε decreases, *i.e.* with increasing z, the relative probability of hypothesis 1 tends to 1 and the relative probability of hypothesis 2 tends to 0. One can conclude that hypothesis 1 is the most likely hypothesis given the data, and hence the remaining modelling in this chapter is carried out considering the mathematical models under hypothesis 1 only.

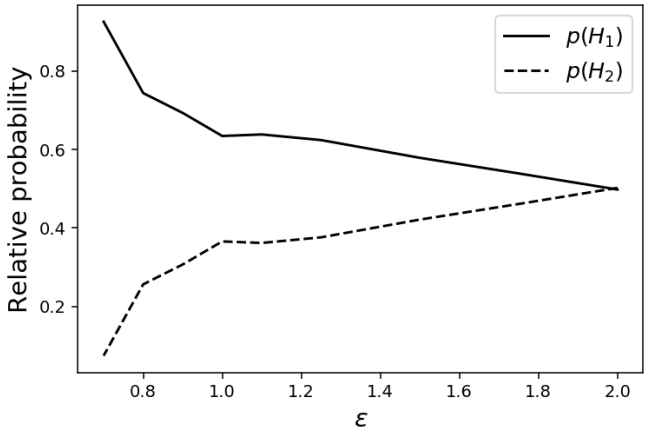

Figure 4.11: Result of the model selection to determine which hypothesis regarding internalisation/degradation of receptor molecules was most likely given the data.

The Bayesian model selection algorithm, Algorithm 5, incorporates the ABC-SMC parameter estimation algorithm, and hence at each iteration, posterior distributions for each model hypothesis can also be obtained. By taking the posterior distributions for hypothesis 1 at iteration $Z = 15$, one could learn about the parameters in the model, for the most likely hypothesis. However, although hypothesis 1 was predominantly selected at iteration 15, there were still some parameter sets accepted into hypothesis 2, and hence a full posterior of size $N = 10^4$ was not gained for a single hypothesis using the model selection algorithm. Therefore, the parameter estimation is carried out separately, using Algorithm 3, in Section 4.3.5. Firstly however, the sensitivity of each of the model parameters to the mathematical model output under hypothesis 1, is assessed in the following section.

4.3.4 Global sensitivity analysis

Before carrying out the Bayesian parameter inference to find posterior distributions for the model parameters, it is of interest to determine which of the model parameters in Table 4.1 are most influential to the output of the models. The Sobol method, described in Section 2.4, is used here to generate a time course of

the total order Sobol index for each parameter in the HypIL-6 and IL-27 mathematical models, where for each model there are two outputs of interest, $[pS_1]^{T,j}$ and $[pS_3]^{T,j}$ for $j \in \{6, 27\}$. A feasible range must be assigned to each parameter in the model, and these ranges are chosen here to reflect the prior distributions in Table 4.2 and are stated in Table 4.3. A time course of 0 to 180 minutes was used, similar to the time course of the experiments. The results of this analysis are plotted in Figure 4.12, where the top row corresponds to the HypIL-6 model outputs and the bottom row to the IL-27 model outputs, with $[pS_1]^{T,j}$ in the left hand column and $[pS_3]^{T,j}$ in the right hand column, for $j \in \{6, 27\}$. In each subplot, the time courses for the parameters with mean total order Sobol index greater than 0.15 for the output of interest, are plotted. One can assume that a parameter with mean total order Sobol index less than 0.15 is of little influence to the outputs of interest of the model.

From Figure 4.12, it can be seen that there are three rates which are particularly influential for both models and both outputs, β_j, $r^+_{1,j}$ for $j \in \{6, 27\}$ and q, *i.e.* the rate of receptor internalisation/degradation, the rate of receptor-ligand binding and the rate of STAT phosphorylation. There is a slight decrease in the total order Sobol index of $r^+_{1,j}$ for $j \in \{6, 27\}$ over time which can be explained since ligand stimulation occurs at time 0 and hence receptor-ligand binding is a reaction which one would expect to occur most towards the start of the time course of the experiments. Similarly, there is a slight decrease in the total order Sobol index for q over time. In contrast, the time course of the total order Sobol index for β_j ($j \in \{6, 27\}$) increases over time, indicating that receptor internalisation/degradation reactions happen more frequently in later times. For both the HypIL-6 and IL-27 models, the initial concentration of STATi is relatively important for the output $[pS_i]^{T,j}$, which is expected, since the concentration of STATi determines, ultimately, the maximal possible concentration of $[pS_i]^{T,j}$ with $i \in \{1, 3\}$ and $j \in \{6, 27\}$. Likewise, for both the HypIL-6 and IL-27 models, the rate of STATi dephosphorylation (d_i) becomes increasingly important for the output $[pS_i]^{T,j}$ over the time course with $i \in \{1, 3\}$ and $j \in \{6, 27\}$, indicating that STAT dephosphorylation happens most at later times, which is intuitive given that the pSTATi molecules are not present at the beginning of the time course and must first form before they can dephosphorylate. Finally, the remaining time

Parameter/IC	Range
$r_{1,6}^{+}$	10^r, where $r \in [-10, 5]$
$r_{1,6}^{-}$	10^r, where $r \in [-15, 5]$
$r_{1,27}^{+}$	10^r, where $r \in [-10, 5]$
$r_{1,27}^{-}$	10^r, where $r \in [-15, 5]$
$r_{2,j}^{+}$ for $j \in \{6, 27\}$	10^r, where $r \in [-2, 3]$
$r_{2,j}^{-}$ for $j \in \{6, 27\}$	10^r, where $r \in [-3, 1]$
k_{ia}^{+}, k_{ib}^{+} for $i \in \{1, 3\}$	10^r, where $r \in [-7, 1]$
k_{ia}^{-}, k_{ib}^{-} for $i \in \{1, 3\}$	10^r, where $r \in [-2, 1]$
q	10^r, where $r \in [-3, 2]$
d_i for $i \in \{1, 3\}$	10^r, where $r \in [-5, -2]$
β_j for $j \in \{6, 27\}$	10^r, where $r \in [-5, -1]$
$[R_1](0)$	$[0, 40]$
$[R_2](0)$	$[0, 100]$
$[S_1](0)$	$[0, 800]$
$[S_3](0)$	$[0, 1000]$

Table 4.3: Feasible ranges for each of the parameters in the mathematical models, used in the Sobol sensitivity analysis.

courses which can be seen in Figure 4.12 are those for the rates at which the STAT molecules bind to GP130 and IL-27Rα (k_{ia}^{+} and k_{ib}^{+}, respectively, for $i \in \{1, 3\}$), where only the time courses for the rates k_{ia}^{+} are seen in the top row of subplots corresponding to the HypIL-6 model outputs, since the HypIL-6 homodimer is formed of GP130 receptors only. The rates for STAT1 binding to a receptor (k_{1a}^{+} and k_{1b}^{+}) are influential to the model outputs for total phosphorylated STAT1 (left hand column), whereas the rates for STAT3 binding to a receptor (k_{3a}^{+} and k_{3b}^{+}) are influential to the model outputs for total phosphorylated STAT3 (right hand column). It can be seen from the HypIL-6 model output subplots, that the parameters k_{1a}^{+} and k_{3a}^{+} have an average total order Sobol index across the time course of approximately 0.5. Given that there are two receptors which the

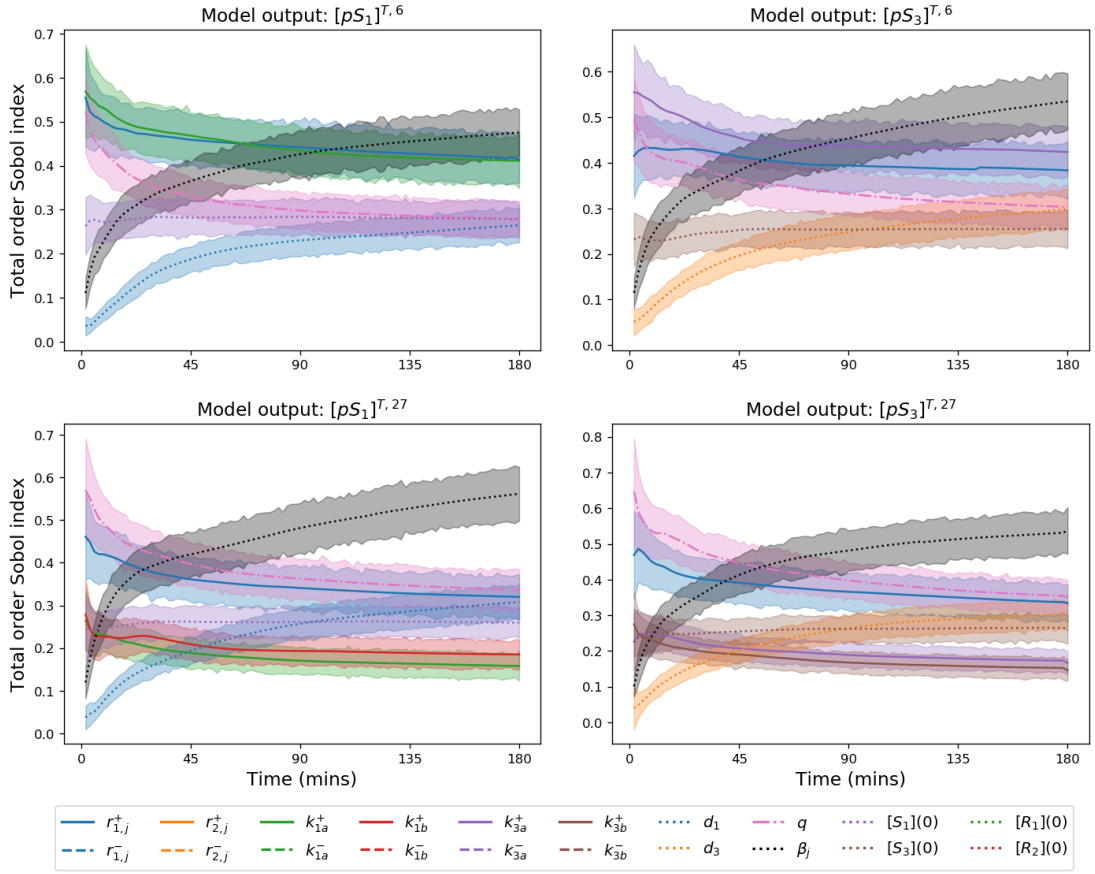

Figure 4.12: Top row: Means and 95% confidence intervals of the total order Sobol indices for the parameters of the HypIL-6 model. **Bottom row:** Means and 95% confidence intervals of the total order Sobol indices for the parameters of the IL-27 model. For each model and each output, the time courses of total order Sobol indices are plotted for parameters where the mean total order Sobol index across the whole time course is greater than 0.15.

STAT molecules can bind to in the IL-27 model, the average total order Sobol indices for the binding parameters k^+_{1a}, k^+_{1b}, k^+_{3a} and k^+_{3b} in the IL-27 model output subplots, are approximately half the value of the average for k^+_{1a} and k^+_{3a} in the HypIL-6 model output subplots, at around 0.25. Interestingly, for the IL-27 model outputs, it can be seen that the rate at which STAT1 binds to IL-27Rα is consistently more important than the rate at which STAT1 binds to GP130, whereas the opposite is true for STAT3.

There are a number of parameters which do not appear in any subplot of

Figure 4.12 and thus are deemed to be of the lowest influence to each model output. These parameters include the rate of receptor-ligand unbinding, the rate of dimerisation and dissociation of the dimer, the rate at which the STAT and pSTAT molecules dissociate from the dimer, and the initial concentrations of each receptor type.

4.3.5 Bayesian parameter inference

In this Section, predictions are made about the values of the parameters in terms of posterior distributions derived via ABC-SMC, as discussed in Section 2.5.2. The aim of this inference is both to narrow down the beliefs about the parameter values from the prior distributions, and to determine which reactions are influencing the differential signalling by pSTAT1 under different cytokine stimulation. The model output is compared with the MFI data in Figure 4.6 from both the RPE1 cells and Th-1 cells (defined by Equation 4.57). The ABC-SMC is carried out firstly using the RPE1 data, to estimate all of the model parameters and initial concentrations listed in Table 4.2. There are a number of parameters which one might expect to vary due to the cell type, in particular, q, d_1, d_3, β_6, β_{27}, $[R_1](0)$, $[R_2](0)$, $[S_1](0)$ and $[S_3](0)$, and hence these parameters are estimated separately for the Th-1 data. In the ABC-SMC using the Th-1 data, the remaining parameters which should not vary due to cell type, are drawn from the posterior distributions from the same analysis using the RPE1 data. The sequence of distance threshold values, ε, used for the ABC-SMC was

$$\{100, 10, 5, 3, 2.5, 2.25, 2, 1.75, 1.5, 1.25, 1.1, 1, 0.9, 0.8, 0.7, 0.6, 0.5\},$$

i.e. the same sequence as was used in the model selection (Section 4.3.3) but with two additional smaller values appended to the end. For each cell type, the ABC-SMC was run for 48 hours, and for the RPE1 cell type, the analysis reached iteration 16 ($\varepsilon_{16} = 0.6$), whereas for the Th-1 cell type, the analysis reached iteration 17 ($\varepsilon_{17} = 0.5$). In both cell types, a sample size of $N = 10^4$ was used. Kernel density estimates (KDEs) of the posterior distributions for each parameter in the mathematical models can be seen in Figure 4.13.

The top row of subplots in Figure 4.13 shows the posterior parameter estimates for the ligand binding/unbinding and dimerisation/dissociation rates, where the distributions in red are the posteriors under HypIL-6 stimulation and those in blue are under IL-27 stimulation. Although the Bayesian learning is only very slight for these parameters when compared to the prior distributions, it is worth noting that the parameter distributions for each cytokine are very similar. It is unlikely, therefore, that a difference in the rate of ligand binding/unbinding or dimerisation/dissociation is causing the difference in pSTAT1 signalling upon stimulation with the different cytokines. In the second row of subplots however, there are large differences in the posterior distributions for the different cytokines. In particular, the results indicate that STAT1 binds to IL-27Rα with a faster rate than it does to GP130 ($k_{1b}^{+} > k_{1a}^{+}$), and the opposite is true for STAT3, where it binds faster to GP130 than to IL-27Rα ($k_{3a}^{+} > k_{3b}^{+}$). The unbinding rates in this row (k_{ia}^{-} and k_{ib}^{-} for $i \in \{1, 3\}$) incorporate the rate of STATi and pSTATi (*i.e.* unphosphorylated and phosphoryalted STATi) dissociating the receptors, for $i \in \{1, 3\}$ and it can be observed that STAT1 is predicted to dissociate faster from IL-27Rα than GP130 and again the opposite is true for STAT3. From the fourth subplot in the third row of parameters, one can see that the rate of receptor internalisation/degradation (β_j for $j \in \{6, 27\}$) is likely to be very similar for both cytokines and both cell types. This rate is therefore not expected to account for the differential signalling by pSTAT1 under stimulation with HypIL-6 and IL-27.

The remaining subplots in Figure 4.13 show posterior distributions for parameters which do not depend on the cytokine, but may depend on the cell type. Firstly, the posterior distribution for q (the rate of STAT phosphorylation when bound to a receptor in either dimer type) is notably higher from the ABC-SMC using the RPE1 data, than the Th-1 data. The rate of phosphorylation of STATs is dependent on the number of JAK molecules present (and bound to the receptors) and the models presented in this chapter do not account for this. The observed difference in the parameter q, between the two cell types, may therefore be a result of differing copy numbers of JAK molecules in each cell type. The second and third subplots on the third row show the posterior distributions for the rates of pSTAT1 and pSTAT3 dephosphorylation in the cytoplasm, respectively. The modelling result here agrees with the experimental data used to inform the

prior distributions for these parameters in Section 4.3.1, whereby although uniform priors over three orders of magnitude were used for the rates d_1 and d_3, the Bayesian inference indicates that these rates should be very similar in RPE1 cells. The same is true for Th-1 cells, where the rate of pSTAT1 dephosphorylation is inferred to be slightly larger than the rate of pSTAT3 dephosphorylation, in line with the experimental result with Tofacitinib in Th-1 cells. The bottom row of Figure 4.13 shows the posterior distributions for the initial concentrations of both receptor types and both STAT types in RPE1 and Th-1 cells. The results indicate that there is a lower concentration of GP130 in Th-1 cells than in RPE1 cells, and that there is a slightly higher concentration of STAT1 than STAT3 in Th-1 cells.

As well as quantifying the individual parameter values in the model by inferring their posterior distribution, one can also evaluate the correlation between the posterior distributions of pairs of parameters in the models. For the RPE1 and Th-1 posteriors seen in Figure 4.13, the Pearson correlation coefficient (see Definition 12 of Chapter 2 and Peck *et al.* (2015)) between each pair of parameters was computed and for those pairs with an absolute value of the correlation coefficient greater than 0.5, indicative of a strong correlation, scatter plots of the posterior distributions are plotted in Figure 4.14. The top two rows show the posterior pairs with strong correlation from the RPE1 cells and the bottom two rows show the same pairs of parameters for the Th-1 cells. Also plotted is a linear model fitted to the accepted parameter value points for each subplot, and the Pearson correlation coefficient for each pair is stated in the legend of each subplot. There are four parameters which, when paired with each other, result in the highest values of the Pearson correlation coefficient. They are d_1, d_3, β_6 and β_{27}, *i.e.* the rates of STAT1/3 desphosphorylation and the rates of receptor internalisation/degradation under HypIL-6 and IL-27 stimulation.

From the first subplot in Figure 4.14 for each cell type, it can be seen that d_1 has a strong positive correlation with d_3. Likewise, the last subplot for each cell type shows a strong positive correlation between β_6 and β_{27}, *i.e.* when receptors are internalised/degraded rapidly under stimulation with HypIL-6, this is likely also to be the case under stimulation with IL-27. These two positive correlations further support the modelling result that it is the STAT binding

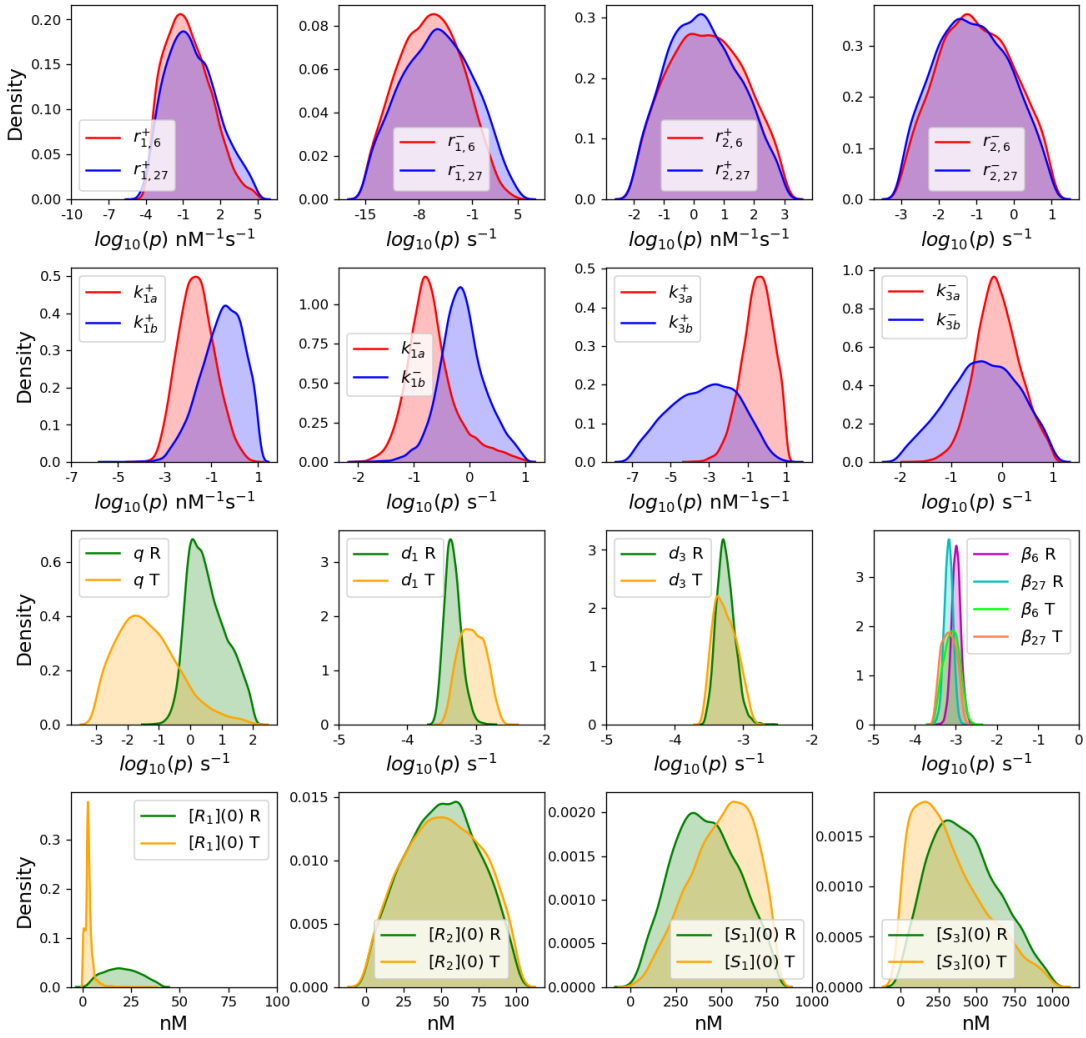

Figure 4.13: Kernel density estimates of the posterior distributions for each of the parameters in the mathematical models, as a result of the ABC-SMC, where p represents the parameter(s) stated in the legend of each subplot. In the figure legends, "R" stands for RPE1 and "T" stands for Th-1.

and unbinding to the receptor rates which are responsible for the differential signalling by pSTAT1 under different cytokines, and not any of the other rates, such as pSTAT dephosphorylation or receptor internalisation/degradation. The other four subplots for each cell type in Figure 4.13 show scatter plots of the posterior distributions for the parameter pair d_i versus β_j for $i \in \{1, 3\}$ and $j \in \{6, 27\}$, and for each pair there is a strong negative correlation. Both the d_i ($i \in \{1, 3\}$)

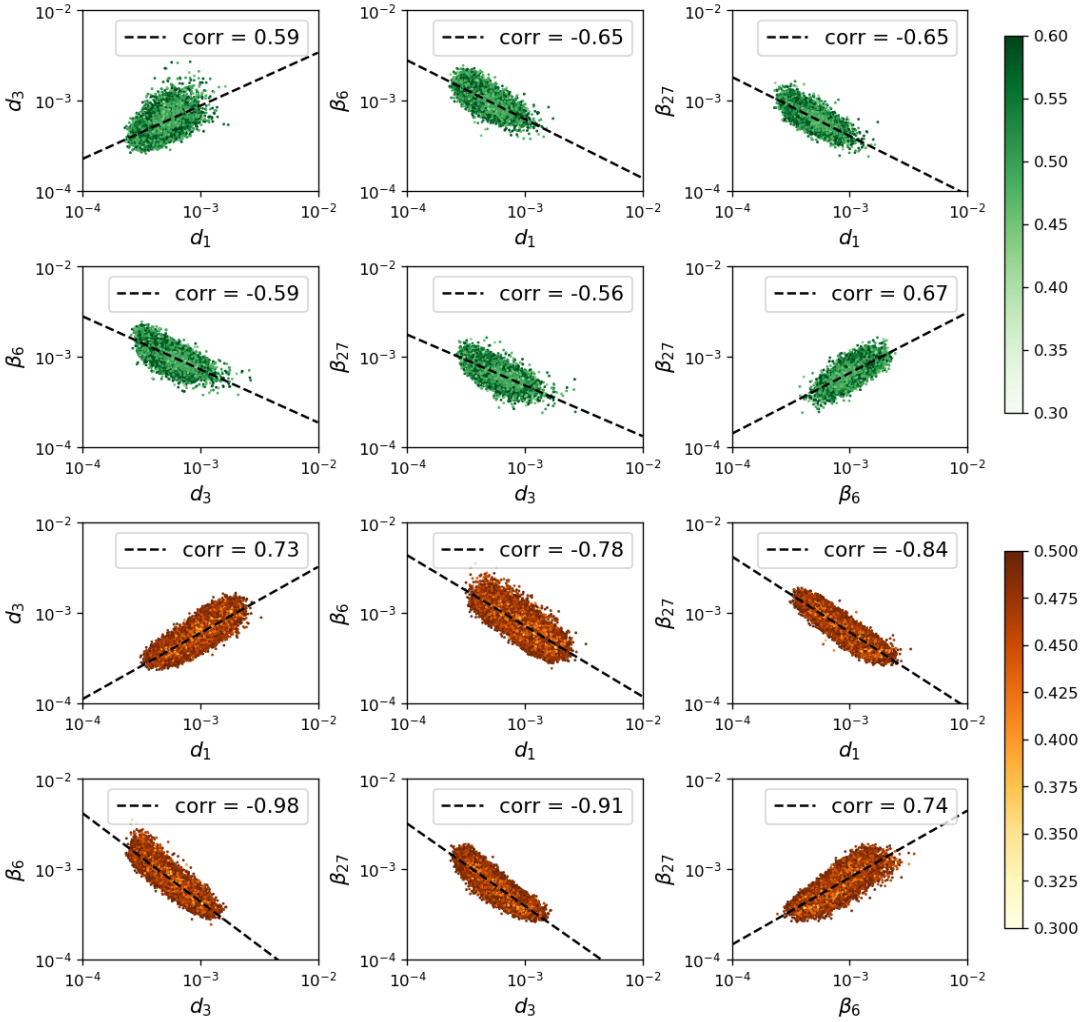

Figure 4.14: Scatter plots of the posterior distributions for pairs of parameters in the mathematical models whose Pearson correlation coefficient was greater than 0.5 or less than -0.5. The colour of each point represents the distance $\delta(sim, data)$ between the model simulation and the data points. The first two rows correspond to pairs of parameters inferred using the RPE1 cell data and the last two rows correspond to pairs of parameters inferred using the Th-1 cell data.

and β_j ($j \in \{6, 27\}$) parameters are rates for reactions which account for a decrease in $[pS_1]^{T,j}$ and/or $[pS_3]^{T,j}$, with $j \in \{6, 27\}$. To give the closest fit to the experimental data, the modelling result here indicates that if pSTAT dephosphorylation is relatively slow, then receptor internalisation/degradation should be relatively fast, and vice versa. Overall it is not surprising that these four param-

eters constituted pairs of parameters with the strongest correlations, given that they were highly influential to the model outputs in the Sobol sensitivity analysis. Also note that these parameters resulted in relatively narrow ABC-SMC posterior distributions.

Finally, In Figure 4.15, the pointwise median and 95% credible intervals of the model simulations using the parameter sets comprising the posterior distributions in Figure 4.13 are plotted, along with the normalised experimental data. It can be seen that the credible intervals capture the majority of the data points and the pointwise median of the simulations is a good representation of the data for each cell type and cytokine. The model therefore captures the prolonged pSTAT1 signalling under stimulation with IL-27 and together with the Bayesian learning one can conclude that this prolonged signalling is due to differences in the rates of STAT binding and unbinding to the two receptor types.

4.4 Model validation

The third objective as stated in Section 4.3 was to use the mathematical models to make predictions about pSTAT signalling in different concentration regimes of receptors and STATs. In order to trust the model predictions, in this section, further experimental datasets are used to validate the models. In particular, the mathematical models and posterior distributions inferred in Section 4.3.5 are used to simulate two experimental set-ups.

4.4.1 Chimera experiments

Firstly, to further corroborate the fact that it is interactions between the intracellular tail of the receptors and the STAT molecules that are responsible for the differential signalling by pSTAT1, experiments were carried out using an IL-27 GP130 *chimera* molecule. This is a receptor molecule with the extracellular head of IL-27Rα and the intracellular tail of GP130, and hence ligand binding and dimerisation occurs as in Figure 4.3 (b) but the STAT interaction reactions happen as in Figure 4.3 (c). The IL-27 model must therefore be modified to resemble the chimera system and this was done as follows. The IL-27 chimera

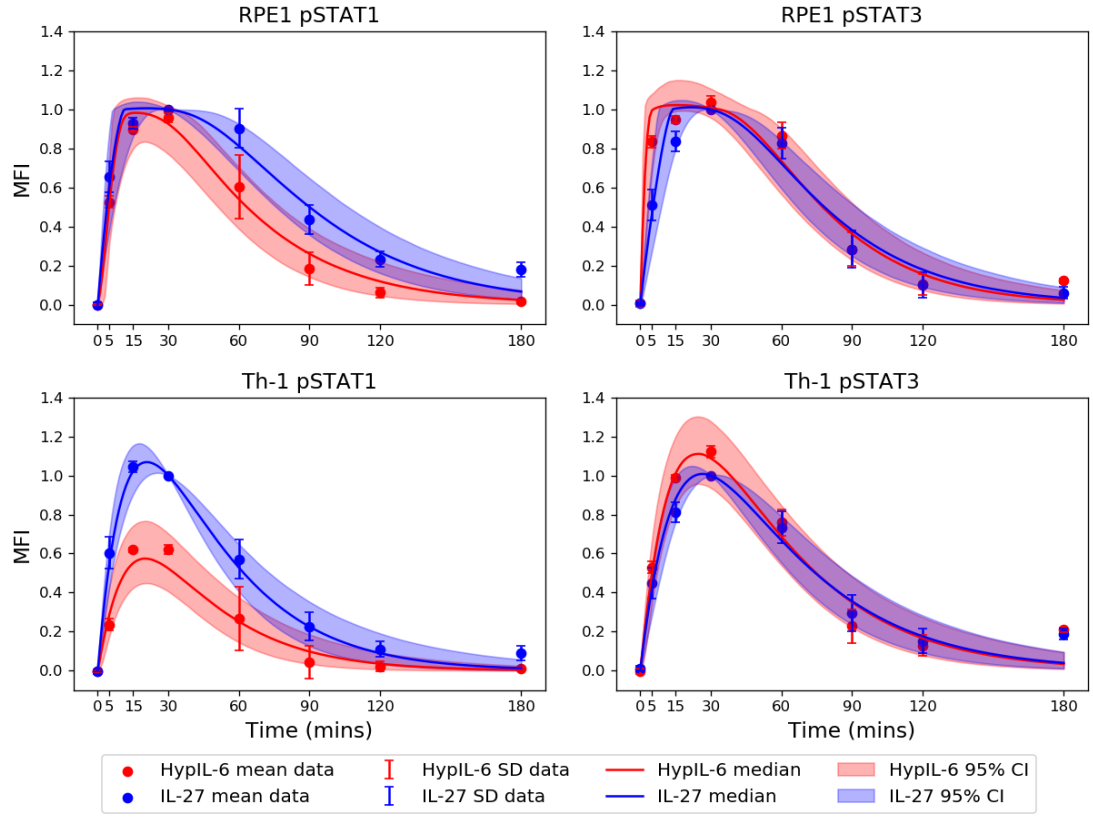

Figure 4.15: Pointwise median and 95% credible intervals of the model simulations using the parameters sets comprising the posterior distributions from the ABC-SMC.

model included the ODEs of the original IL-27 mathematical model involved in the formation of the dimer, Equations (4.23) - (4.26), and the ODEs of the HypIL-6 mathematical model post-dimer formation, Equations (4.5) – (4.22), in which D_6 was replaced by D_{27}. The ODE of the IL-27 induced dimer in the chimera model was modified as

$$\begin{aligned}\frac{d\left[D_{27}\right]}{dt} &= r^{+}_{2,27}\left[C_2\right]\left[R_1\right] - r^{-}_{2,27}\left[D_{27}\right] - 2k^{+}_{1a}\left[D_{27}\right]\left[S_1\right] + k^{-}_{1a}\left(\left[S_1 \cdot D_{27}\right] + \left[pS_1 \cdot D_{27}\right]\right) \\ &\quad - 2k^{+}_{3a}\left[D_{27}\right]\left[S_3\right] + k^{-}_{3a}\left(\left[S_3 \cdot D_{27}\right] + \left[pS_3 \cdot D_{27}\right]\right) - \beta_{27}\left[D_{27}\right]. \qquad (4.65)\end{aligned}$$

The HypIL-6 model remained the same as in Equations (4.1) - (4.22) for the chimera system, since the HypIL-6 system does not involve IL-27Rα. A depiction of the IL-27 chimera model reactions is given in Figure 4.16, and in line with the

model selection result in Section 4.3.3, the IL-27 chimera model was also based on hypothesis 1.

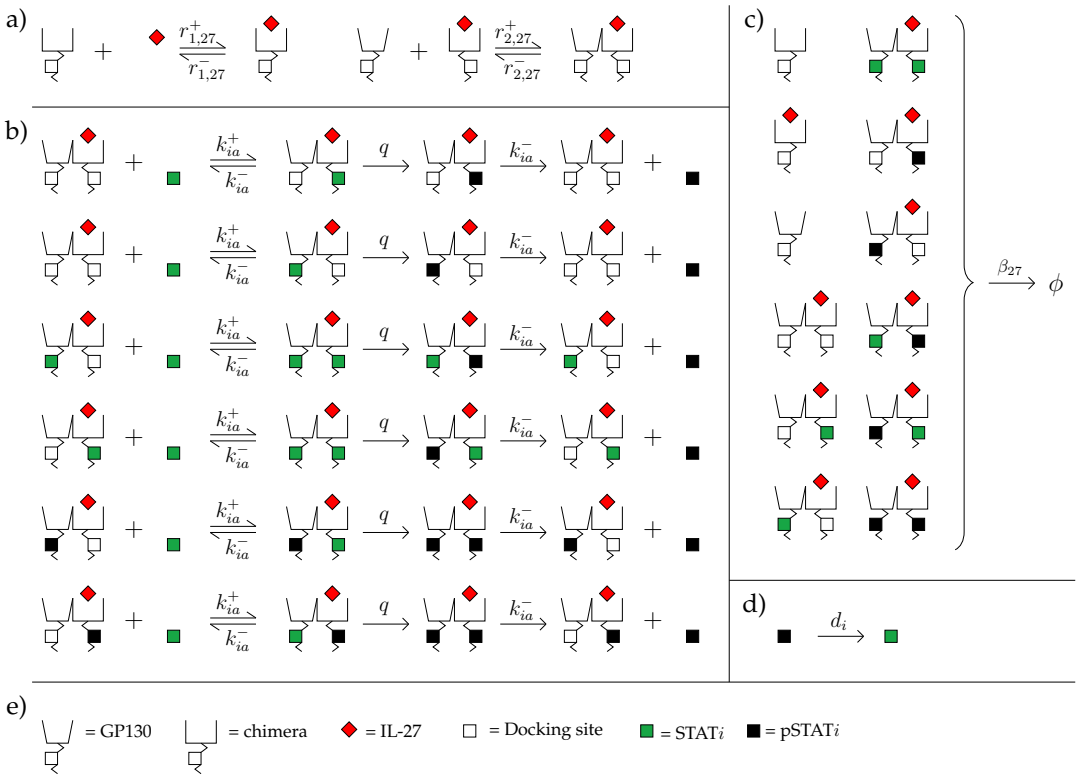

Figure 4.16: Depiction of the reactions comprising the IL-27 chimera mathematical model. a) Reactions involving ligand binding and dimerisation. b) Reactions involving STATi molecules, for $i \in \{1, 3\}$. c) Reactions involving receptor internalisation/degradation. d) Dephosphorylation of pS_i, for $i \in \{1, 3\}$, in the cytoplasm. h) Key for the molecules in the reactions.

The data from the chimera experiments resulted in a very similar time course for HypIL-6 and IL-27 induced pSTAT1 signalling, *i.e.* the prolonged signalling by pSTAT1 under IL-27 was no longer observed. This behaviour was expected, since removing the intracellular tail of IL-27Rα means that STAT1 no longer has a binding advantage over STAT3 when the system is stimulated with IL-27, and thus, pSTAT1 and pSTAT3 show similar behaviour. In agreement with the original experiments, the time course for pSTAT3 is also very similar between HypIL-6 and IL-27, and these data points can be seen in the top row of Figure

4.17. The data have been normalised in the same way as the original data (see Section 4.2.1).

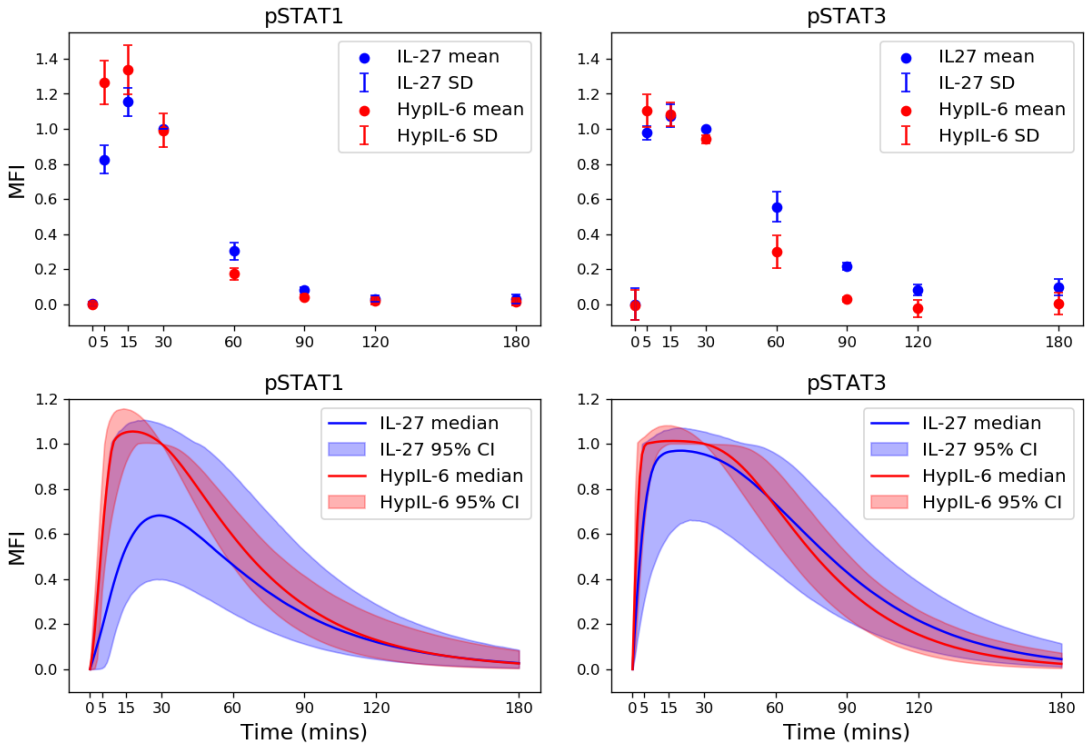

Figure 4.17: Top row: Normalised chimera data for pSTAT1 (**left**) and pSTAT3 (**right**) under stimulation with HypIL-6 and IL-27. **Bottom row:** Pointwise median and 95% credible intervals of the chimera mathematical model simulations.

To test the accuracy of the mathematical models in this situation, the HypIL-6 and IL-27 chimera models were simulated using the parameter sets comprising the RPE1 posterior distributions shown in Figure 4.13. The pointwise median and 95% credible intervals of the chimera model simulations are shown in the bottom row of Figure 4.17. The model simulations are not expected to overlay the data points perfectly in this case due to a difference in the experimental set-up, whereby "reverse order kinetics" were used in the chimera experiments as opposed to "correct order kinetics" which were used to generate the original data. The main difference between the methods is that in reverse order kinetics the experiments are started at different times and ended together in order to

gather data for each of the time points, whereas in correct order kinetics the experiments are started at the same time and ended at different time points. Correct order kinetics is a more accurate method but is also more time consuming and experimentally challenging and hence reverse order kinetics was used for some of the experiments. Under reverse order kinetics, one can expect the signalling to appear faster than in correct order kinetics, *i.e.* the data points are shifted to the left. Since the posteriors used for this analysis were obtained using data generated via correct order kinetics but the chimera data was generated using reverse order kinetics, it is expected that the data will appear to be shifted to earlier time points than the model simulations. As well as this difference in method, slightly differing concentrations of the cytokines were used in the chimera experiments and in the original experiments. The general trend from the data can however be compared with the trend of the model simulations, and firstly it can be seen that the pSTAT3 credible interval is a good representation of the pSTAT3 data points. The pointwise median of the IL-27 pSTAT1 model simulations appears to underestimate the peak of the IL-27 pSTAT1 data, however the 95% credible interval is relatively wide and, promisingly, the maximal value of the 95% credible interval almost reaches the maximal value of the HypIL-6 95% credible interval, representative of the data. It is clear from both the data and the mathematical model results that by removing the IL-27Rα tail and replacing this with the GP130 tail, the greater and more sustained signalling by pSTAT1 no longer occurs and hence one can be certain that this signalling difference is caused by STAT1 interactions with the intracellular tail of IL-27Rα.

4.4.2 Mutant experiments

As a second test of the accuracy of the mathematical models, another dataset can be used, with similar time course data as the original experiments but here using a mutant version of IL-27Rα. In particular, a specific tyrosine residue (Y613) on the intracellular tail of IL-27Rα was identified as being the high affinity tyrosine for STAT1 binding. A mutant variety of the receptor, known as the Y613F mutant, was generated with this specific tyrosine residue removed. Given that IL-27Rα only forms dimers and becomes activated when stimulated with IL-27,

only the IL-27 data and mathematical model are used in this section as one would not expect the mutation of IL-27Rα to have any effect on the HypIL-6 stimulated system. The wild type (WT) data, *i.e.* data from experiments in which IL-27Rα is not mutated, and the Y613F mutant data, upon stimulation with IL-27 are shown in the top row of Figure 4.18. From the data, it can be seen that there is a large decrease in pSTAT1 signalling from the mutant system compared with the WT. This observation is in line with the experimental hypothesis, that pSTAT1 signals mostly through the high affinity tyrosine residue (Y613) on IL-27Rα which has been removed in the mutant. STAT1 is therefore competing with STAT3 for binding to GP130 in order to signal, however from the posterior distributions in Figure 4.13 (comparing k_{1a}^{+} with k_{3a}^{+}) one can infer that STAT3 has a higher binding rate to GP130 than STAT1 and hence out-competes STAT1. From the right-hand subplot of the top row of Figure 4.18, the effect of mutating IL-27Rα on pSTAT3 can be seen, whereby the pSTAT3 signal is also slightly reduced in the mutant data, but only by 26% as opposed to 83% in the case of pSTAT1. There are two possible explanations for this slight decrease in pSTAT3 signalling. Firstly, similarly to STAT1, STAT3 could be doing a small amount of signalling through the tyrosine on IL-27Rα which is removed in the mutant version of the receptor, and secondly, in the mutant system there is more competition from STAT1 for binding to GP130.

Given the experimental hypothesis, that STAT1 binding to IL-27Rα is greatly reduced in the mutant system, to simulate the mathematical model under this scenario, the parameter k_{1b}^{+} representing this rate of binding, was fixed at a much lower value than is suggested by the posterior distribution of this parameter. The unbinding rate k_{1b}^{-} was also fixed in the simulations in order to keep a consistent binding affinity ($K_{d,1b}$) where $K_{d,1b} = k_{1b}^{-}/k_{1b}^{+}$. The values chosen for these parameters were $k_{1b}^{-} = 10^{0}$ s^{-1} (reflecting the median of the posterior distribution for this parameter) and $k_{1b}^{+} = 10^{-5}$ nM^{-1}s^{-1}, resulting in a binding affinity of $K_d = 100$ μM. These values reflect the assumptions from the experimentalists on this parameter value. In the simulations of the mutant mathematical model, all other parameters were sampled from their posterior distributions in Figure 4.13, as in the WT simulations. The pointwise medians and 95% credible intervals of the simulations for the WT and mutant mathematical models are seen in the

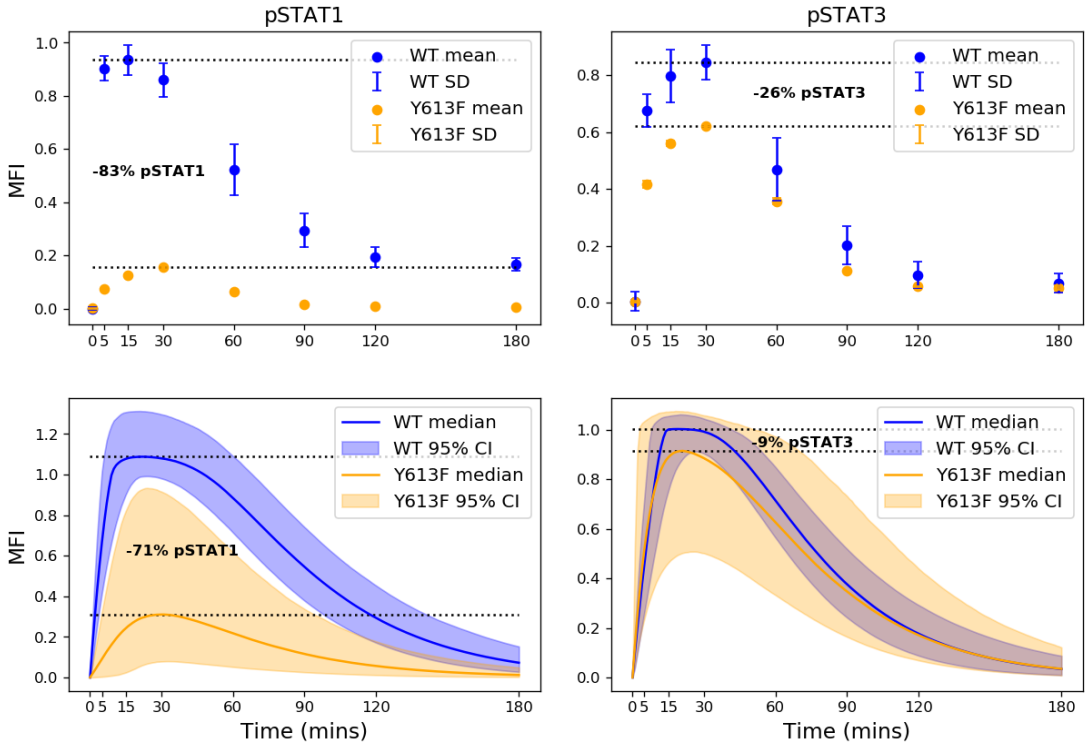

Figure 4.18: Top row: Normalised WT and mutant data for pSTAT1 (**left**) and pSTAT3 (**right**) under stimulation with IL-27. **Bottom row:** Pointwise median and 95% credible intervals of the WT and mutant mathematical model simulations.

bottom row of Figure 4.18, normalised to the WT model for each of pSTAT1 and pSTAT3. Similarly to the chimera experiments, the mutant experiments were carried out using reverse order kinetics and slightly different concentrations of cytokines compared with the original (WT) experiments and hence the model simulations will not overlay the data points exactly. The trend in the model simulations however, are in good agreement with the experimental data, where the pSTAT1 response is greatly reduced in the mutant simulations compared with the WT and the pSTAT3 response is also marginally reduced. This result, along with the chimera result gives confidence to the mathematical model and validates its use to make predictions about pSTAT signalling, as is discussed in the following section.

4.5 Model predictions

So far in this chapter, the focus has been to develop mathematical models of HypIL-6 and IL-27 induced STAT signalling and to use Bayesian methods to parameterise these models in order to infer specific reactions responsible for the prolonged activation of STAT1 under stimulation with IL-27. Having validated such models in the previous section, the focus here is on using the mathematical models to make predictions about pSTAT1/3 signalling in different biological regimes, specifically, looking at different concentration regimes of receptors and STATs. The interest in different concentration regimes stems from biological observations by Dr. Moraga Gonzalez and Dr. Wilmes, whereby patients with certain diseases, such as Crohn's disease and SLE, exhibited highly up-regulated levels of STAT1 and, in some cases, GP130. In the context of disease, it is interesting to see how these up-regulated protein levels affect the cytokine signalling response, given that IL-6 and IL-27 are known to be very important in the regulation of the adaptive immune system (O'Shea & Plenge, 2012; Yoshida & Hunter, 2015b). For a thorough analysis of the effect of concentrations on the HypIL-6 and IL-27 signalling responses, in this section model predictions are made with both up- and down-regulation of STAT1/3, GP130 and IL-27Rα.

The HypIL-6 and IL-27 mathematical models were simulated using the posterior distributions from the ABC-SMC with RPE1 cell data, but where the initial concentrations $[S_1](0)$, $[S_3](0)$, $[R_1](0)$ and $[R_2](0)$ were individually altered. A first observation was that the up-regulation of either receptor type had no effect on the signalling of pSTAT1 or pSTAT3 (data not shown here). This result can be explained by considering the initial concentrations of each cytokine used in the experiments and model simulations, $[L_6](0) = 10$ nM and $[L_{27}](0) = 2$ nM. From Figure 4.13, it can be seen that the posterior distributions for $[R_1](0)$ and $[R_2](0)$ are predicting that these concentrations should be, for the most part, greater than 10 nM, in particular in the case of $[R_2](0)$. It is therefore not surprising that increasing the receptor concentrations has no effect on the signalling of pSTAT1/3, since the receptor concentrations from the posterior distributions are already saturating, *i.e.* no more complexes or dimers could form even if the receptor concentrations were increased, since each ligand is already bound to a

receptor. Therefore, to examine the effect of altering receptor concentrations, 10 and 100 fold decreases from the median values of the posterior distributions for $[R_1](0)$ and $[R_2](0)$ were considered, where the median values are approximately 25 nM and 50 nM, respectively. These model simulations, showing the pointwise median and 95% credible intervals, can be seen in Figure 4.19. The top row shows the model predictions using the median values from the posteriors for $[R_1](0)$ and $[R_2](0)$, the second and third rows show predictions upon decreasing $[R_2](0)$ and the fourth and fifth rows show predictions upon decreasing $[R_1](0)$.

There is no difference between the first and second row of subplots in Figure 4.19, which can be explained since $[R_2](0) = 5$ nM is still a saturating concentration of receptor compared with the concentration of IL-27 at 2 nM. However, when the concentration of IL-27Rα is reduced further, to 0.5 nM, there is a reduction in both pSTAT1 and pSTAT3 signalling under stimulation with IL-27. The largest reduction is in pSTAT3 signalling, presumably because, since there is very little IL-27Rα in the system, there is much greater competition from STAT1 for GP130. The signalling from the HypIL-6 model is not affected since IL-27Rα does not feature as a variable in this model. When the concentration of GP130 is reduced 10 fold, from the fourth row of Figure 4.19 it can be observed that there is no effect on either pSTAT1 or pSTAT3 signalling in the IL-27 model. One can assume that this is because STAT1 signals mostly through IL-27Rα and 2.5 nM GP130 is still sufficient for STAT3 to reach maximal signalling. On the other hand, when the GP130 concentration is reduced 100 fold (fifth row of subplots), there is a dramatic effect on pSTAT1/3 signalling under both HypIL-6 and IL-27 stimulation. In the case of the HypIL-6 model, there is virtually no pSTAT1 signalling and a small amount of pSTAT3 signalling, roughly one tenth of the signal in the first row of subplots. In the IL-27 model there is slightly more signalling from both pSTAT1 and pSTAT3 compared with the HypIL-6 model, since GP130 forms only one half of the signalling dimer in this case.

Also of interest is to see the effect of varying STAT concentrations on pSTAT signalling, in particular increasing the STAT1 concentration since this scenario has been observed in Crohn's disease. To this end, in Figure 4.20 the initial concentrations of STAT1 and STAT3 are individually increased or decreased by 10 fold from their median posterior values, where the top row of subplots shows

the model simulations using these median values. In the second and third rows of subplots, the STAT3 concentration is varied, and in the fourth and fifth rows of subplots, the STAT1 concentration is varied.

From the second and fourth rows of Figure 4.20 it is noted that decreasing the STAT3 or STAT1 concentration by a factor of ten affects only the signalling output for the respective STAT type, *i.e.* only the pSTAT3 signal is reduced upon reduction of $[S_3](0)$ and only the pSTAT1 signal is reduced upon reduction of $[S_1](0)$. It is found that reducing STAT1/3 by a factor of ten, causes a ten fold reduction in the signalling output for pSTAT1/3. In the third row of subplots, the STAT3 initial concentration is increased to 5000 nM. In the HypIL-6 model, STAT3 therefore hugely out competes STAT1 for GP130 binding and there is a large increase in pSTAT3 signalling, correlating with a large decrease in pSTAT1 signalling. In the IL-27 model, the pSTAT1 signal decrease is only very slight, since STAT1 signals mostly through IL-27Rα, and although STAT3 is in great excess of STAT1, it has a very slow rate of binding to IL-27Rα. The pSTAT3 signal again increases, although not as greatly as in the HypIL-6 model, since there is only one GP130 receptor in each dimer in the IL-27 model. Finally, in the fifth row of subplots in Figure 4.20, the STAT1 concentration is increased while the STAT3 concentration remains at the median value of its posterior. As expected, this increase in STAT1 causes an increase in the pSTAT1 signal from both the HypIL-6 and IL-27 mathematical models, where the increase is greater in the IL-27 model since STAT1 has preferential binding to IL-27Rα. The pSTAT3 signal decreases in both models, interestingly more so in the IL-27 model than the HypIL-6. This can be explained by considering that STAT3 binds only to GP130 (since binding to IL-27Rα is seen to be very weak in the posterior distribution) and hence, due to the large amount of competition from STAT1 in the IL-27 system, there are very few receptors that STAT3 can bind with. In the HypIL-6 system however, since the dimer is formed of two GP130 molecules, there is double the binding opportunity for STAT3.

From simulations of the mathematical model, calibrated with the experimental data, one can predict changes in pSTAT signalling in different disease settings which can be useful in understanding how disease-induced dysregulation alters cellular signalling mechanisms.

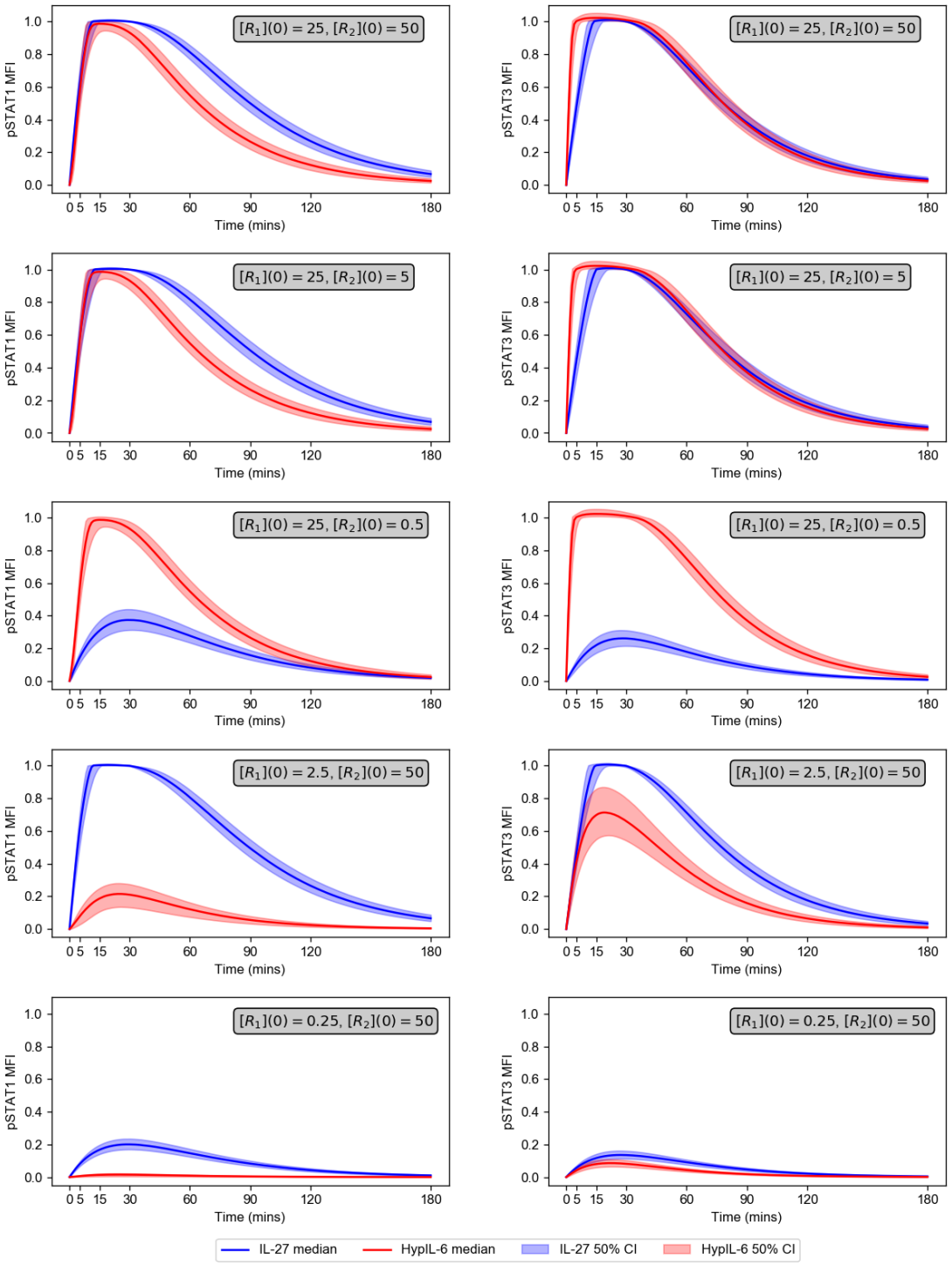

Figure 4.19: Model predictions for varying receptor initial concentrations. **Top row:** GP130 and IL-27Rα concentrations fixed at median values from posteriors. **Second and third rows:** Reduction in IL-27Rα. **Fourth and fifth rows:** Reduction in GP130.

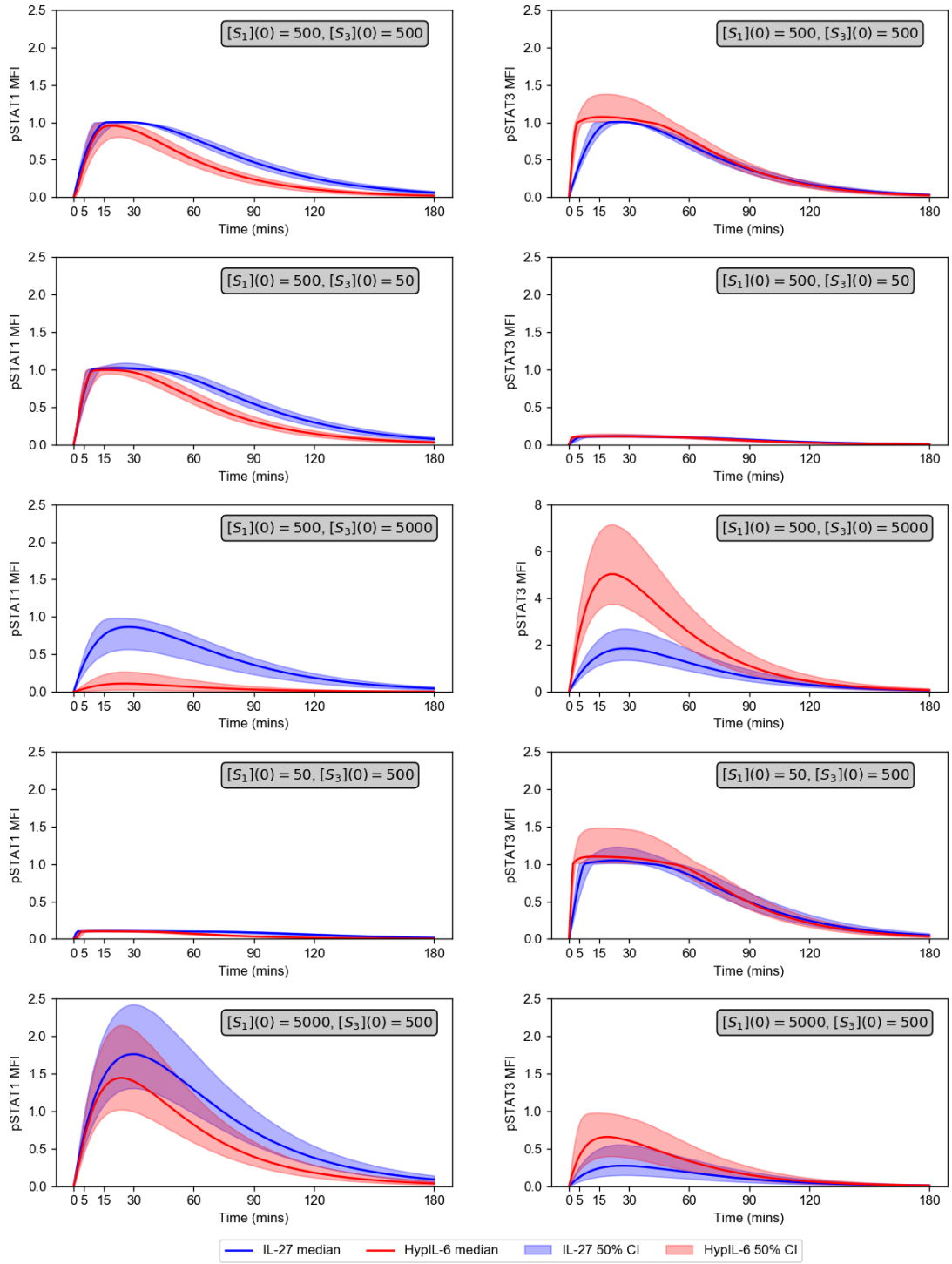

Figure 4.20: Model predictions for varying STAT1/3 initial concentrations. **Top row:** STAT1 and STAT3 concentrations fixed at median values from posteriors. **Second row:** Decrease in STAT3. **Third row:** Increase in STAT3. **Fourth row:** Decrease in STAT1. **Fifth row:** Increase in STAT1.

4.6 Model justification and limitations

In this section, the structure of the mathematical models presented in this chapter is justified with relation to other published mathematical models of the JAK/STAT pathway and further experimental data. The main point of discussion here will be the two model hypotheses relating to the internalisation and degradation of receptor molecules. In the HypIL-6 and IL-27 mathematical models, it is assumed (under either hypothesis) that receptor molecules (and therefore cytokine molecules) are internalised from the membrane and potentially then degraded. Indeed, upon stimulation with a cytokine, the internalisation and degradation pathway for the receptor molecules is greatly up-regulated. A good example of this is given by Tanaka *et al.* (2008) where the authors treat HeLa cells expressing GP130 with IL-6 and measure the GP130 level. They find that, after 120 minutes, GP130 has been completely depleted from the cell (see Figure 4.21). The cells were however, treated with cyclohexamide to prevent new protein synthesis and hence a *total* depletion of GP130 is unlikely to occur in reality. In a similar study (Marijanovic *et al.*, 2007), the authors look at the IFNα2 and IFNβ induced degradation of IFNAR1 and IFNAR2 (where IFN stands for interferon). They find that the ligands induce approximately a 60% reduction in IFNAR1 over three hours, again under cyclohexamide treatment (data not shown here, Figure 7 in Marijanovic *et al.* (2007)).

Although internalisation of receptor species is therefore certainly a process which occurs in cells, it is unlikely that a cell would completely deplete itself of surface receptor molecules through internalisation and degradation, since receptor recycling and synthesis can also occur. In fact, one study (Flynn *et al.*, 2021) found that internalisation of GP130 is a constitutive clathrin-mediated process independent of IL-6, and that upon stimulation with IL-6 the internalised receptors are directed predominantly to the recycling pathway as opposed to the degradation pathway. Recycling of receptors was therefore shown to be important in maintaining the signal from the JAK/STAT pathway. Given that recycling and synthesis reactions were not included in the mathematical models, it was of interest to see the effect of internalisation and degradation of receptors over time.

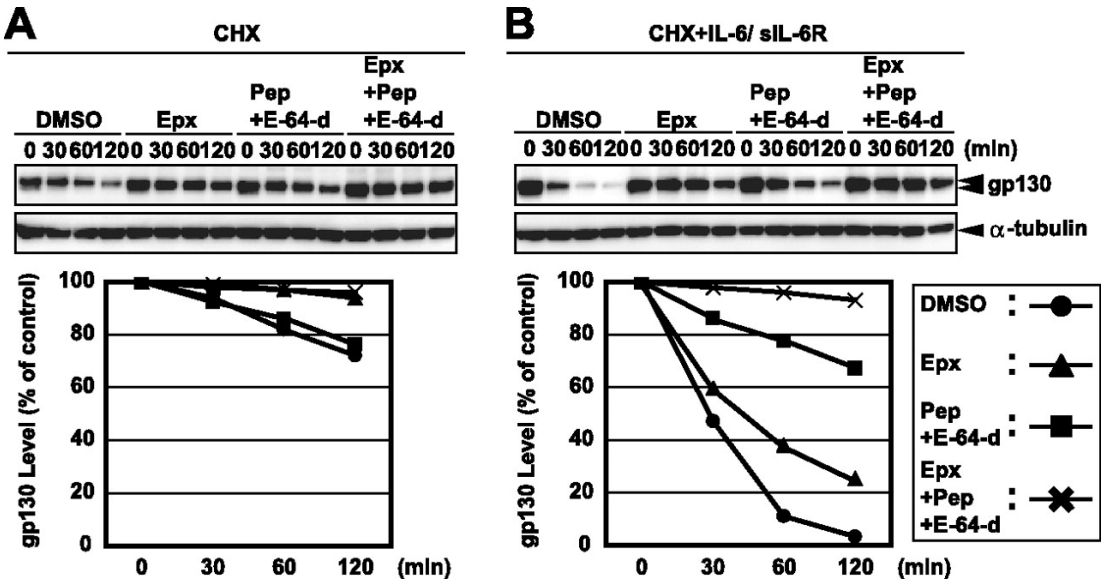

Figure 4.21: A figure showing the total cellular level of GP130 over time in the absence (**left**) or presence (**right**) of IL-6 stimulation, taken from Tanaka *et al.* (2008). When cells are pre-treated in DMSO alone (*i.e.* no lysosomal or proteosomal inhibitors) and stimulated with IL-6, GP130 is largely depleted.

To this end, in Figures 4.22 and 4.23, the total surface GP130, IL-27Rα, HypIL-6 and IL-27 concentrations (sum of all species containing these molecules) are plotted over time using the parameter values comprising the posterior distributions found via ABC-SMC in RPE1 cells (Figure 4.22) and Th-1 cells (Figure 4.23). Subplots are shown for both the IL-27 model parameters (top rows) and the HypIL-6 model parameters (bottom rows).

It can be seen that in both Figures 4.22 and 4.23, aside from the IL-6 concentration in the Th-1 cells, all other concentrations of surface receptors or cytokines approach 0 after approximately 1 hour for the majority of the simulations. This does not necessarily imply that the total receptor populations have been degraded, since the rates β_j and γ_j for $j \in \{6, 27\}$ are of internalisation/degradation and so it could be that the receptor molecules are residing in intracellular compartments. To address the fate of the internalised receptors, one could model an intracellular compartment along with recycling of receptors back to the membrane as is done in other receptor-ligand mathematical models (Lauffenburger & Linderman, 1996; Nazari *et al.*, 2018). This was not included in the models in this chapter

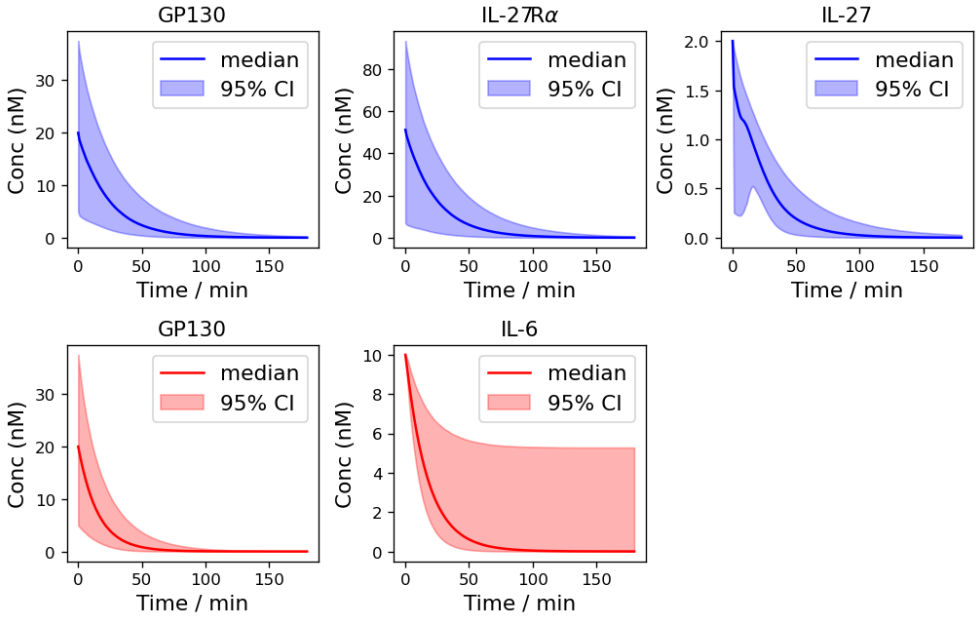

Figure 4.22: Pointwise median and 95% credible intervals of the model simulations for receptor and cytokine concentrations using the parameters sets of the posterior distributions from the ABC-SMC in RPE1 cells. The top row of subplots shows model outputs from the IL-27 mathematical model and the bottom row shows model outputs from the HypIL-6 mathematical model.

however, due to the relatively short time scale of the data, as well as the fact that the internal molecules could not be imaged, and hence any rates relating to these internal species could not be calibrated in the Bayesian inference. Additionally, adding more compartments and species to the mathematical models would increase the complexity of such models and therefore the simulation time. One way in which hypothesis 1 of the models presented here could be improved to give a more realistic representation of receptor internalisation and recycling processes could be to add a term for receptor synthesis to the differential equations for the unbound surface receptor species. Upon appropriate measuring or calibration of this synthesis rate, this may allow for the surface species to decrease to a concentration greater than zero that does not represent total surface receptor depletion. If it were necessary to simulate the model for a longer period of time than the experimental time course of the data set used in this chapter, one may need to consider also recycling of receptors.

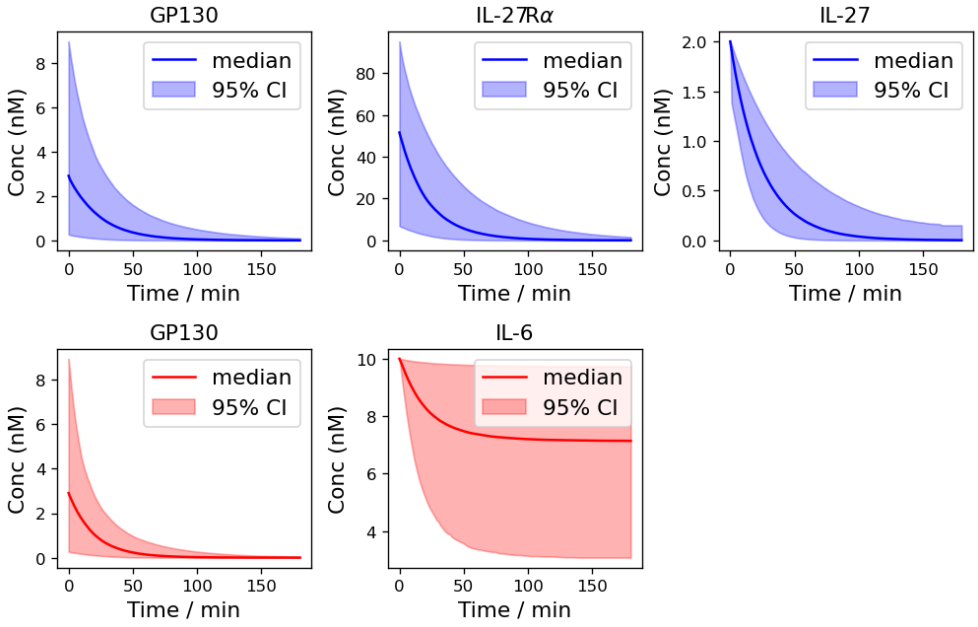

Figure 4.23: Pointwise median and 95% credible intervals of the model simulations for receptor and cytokine concentrations using the parameters sets of the posterior distributions from the ABC-SMC in Th-1 cells. The top row of subplots shows model outputs from the IL-27 mathematical model and the bottom row shows model outputs from the HypIL-6 mathematical model.

In many mathematical models of the JAK/STAT pathway from the literature (Blätke *et al.*, 2013; Reeh *et al.*, 2019; Sobotta *et al.*, 2017; Vera *et al.*, 2011), instead of modelling receptor internalisation/degradation as a means of pSTAT down-regulation, receptor *deactivation* by negative feedback molecules is modelled. This is what is implied by hypothesis 2 of the mathematical models presented in this chapter, whereby the species containing receptor molecules are down-regulated with a rate proportional to the concentration of pSTAT molecules. Although the pSTAT molecules themselves are not negative feedback molecules, once phosphorylated and dimerised, they can migrate to the nucleus and initiate the synthesis of negative feedback molecules such as SOCS3, via activation of the relevant transcription factors. SOCS3 can bind to either the receptors comprising the phosphorylated dimer (through different binding sites to the STAT molecules), or the JAK molecules bound to the dimer, inactivating the signalling complex (Babon *et al.*, 2012; Kershaw *et al.*, 2013; Rottenberg & Carow, 2014).

The dimer becomes unphosphorylated and therefore cannot recruit or phosphorylate STAT molecules, hence down-regulating STAT signalling. Given that the process of SOCS3 synthesis takes some time to occur, the way in which hypothesis 2 is implemented in the mathematical models in this chapter is not wholly realistic since the delay in production of the SOCS3 molecules is not accounted for. Indeed, one study found that after stimulation with 0.08 nM of IL-6 or HypIL-6, cells started to produce SOCS3 after approximately 30 minutes, around the same time at which pSTAT3 levels peaked (Reeh *et al.*, 2019).

In most of the JAK/STAT mathematical models from the literature which incorporate SOCS3 molecules, the dimer is modelled as a single, active, or inactive entity. STAT binding, phosphorylation and subsequent dissociation are then modelled as a single reaction, hence simplifying the model by neglecting intermediate complexes (Dittrich *et al.*, 2012; Reeh *et al.*, 2019). In these models, SOCS3 binding to the dimer causes it to become inactivated and hence no further STAT phosphorylation reactions can occur. In particular, in the model presented by Reeh *et al.* (2019), the authors consider two species for the dimeric receptor complex, one inactive and one active. The positive term in the ODE for the active dimer, which implies its production, is proportional to the concentration of inactive dimer divided by the concentration of SOCS3, so that with increasing SOCS3, the production rate of active dimer decreases. In the models developed in this chapter however, one of the main goals was to determine the binding rates of STAT molecules to the individual receptors in the dimer and hence a more complex mathematical description of the dimer and the STAT phosphorylation events was required. As well as this, it is assumed in the HypIL-6 and IL-27 models that the dimers become activated (phosphorylated) immediately upon formation and hence there is not a term in the ODEs for dimer activation. Incorporating SOCS3 molecules into the models would therefore add great complexity to the models, as is described in Section 4.6.1.

4.6.1 Mathematical modelling of SOCS3

Since there is no inactive dimer species in either of the mathematical models in this chapter, one cannot model SOCS3 negative feedback in the same way as

carried out by Reeh *et al.* (2019). Instead of this, the feedback could be modelled by adding SOCS3 as a species, which is then able to bind to and inactive the receptors in the dimer. This would mean adding reactions to the model whereby SOCS3, denoted here by X_3, can bind to either receptor in either dimer with rate α. Immediately upon binding, it is assumed that the receptor molecule which the SOCS3 is bound to becomes inactivated. For example in the HypIL-6 model, one would have the reaction

$$D_6 + X_3 \rightleftharpoons D_6 \cdot X_3,$$

where $D_6 \cdot X_3$ is a new species to be added to the mathematical model which is inactive and therefore cannot recruit or phosphorylate STATs. Similarly, X_3 could bind to either receptor in either dimer, where the receptor is unbound, bound to STATi, or bound to phosphorylated STATi for $i \in \{1, 3\}$, resulting in multiple additional species in the models. The following assumptions could be made:

- If a SOCS3 molecule binds to a STAT-unbound receptor, this receptor is then inactivated meaning that STAT molecules can no longer bind to it.
- If a SOCS3 molecule binds to a STAT-bound receptor, this receptor is then inactivated, meaning that the bound STAT molecule can no longer become phosphorylated and can only dissociate the receptor. No further STAT molecules can be recruited to the receptor.
- If a SOCS3 molecule binds to a pSTAT-bound receptor, this pSTAT molecule does not dephosphorylate and can dissociate from the receptor. No further STAT molecules can be recruited to the receptor.

In all three cases, the other receptor in the dimer is still free to recruit and phosphorylate STAT molecules unless also bound by SOCS3.

The delay in the production of SOCS3 molecules could be represented in a similar fashion to that used by Reeh *et al.* (2019). In the model presented in this work the authors account for the delay in SOCS3 production (starting after 30 minutes) by introducing "dummy" SOCS3 variables. For example SOCS3 dummy 1 has no effect on any other species in the model (it cannot inactivate the dimer)

and can only transform into SOCS3 dummy 2. One can introduce a chain of dummy molecules until eventually one of the dummy molecules transforms with some rate into active SOCS3. This SOCS3 variable is then capable of inactivating the receptor molecules. Here, three dummy SOCS3 molecules are introduced, namely X_3^a, X_3^b and X_3^c where the active form of SOCS3 is denoted by X_3. The following pathway of reactions can occur,

$$X_3^a \xrightarrow{\delta_1} X_3^b \xrightarrow{\delta_2} X_3^c \xrightarrow{\delta_3} X_3,$$

where then X_3, once formed, can bind to and inactivate the receptor molecules in an active dimer. It is assumed here that there is an initial concentration of X_3^a present in the cells at time 0 which then transforms into the other dummy SOCS3 molecules and eventually the active form, via the chain described above. In reality, the X_3^a would be produced by transcription and translation as a result of pSTAT dimers migrating to the nucleus, however a constant initial concentration is assumed here for simplicity in the models. Although introducing this pathway of dummy reactions allows for a delay in the production of SOCS3 in the model simulations, it is of course not biologically realistic and hence one issue with this method is that the rates cannot be measured and must be fixed to values which result in the required delay in SOCS3 production. Given that SOCS3 is not measured in the experiments which produced the data used in this chapter, the delay in production is here assumed to be 30 minutes, taken from Reeh *et al.* (2019). All other reactions in the HypIL-6 and IL-27 models remain the same, except for the removal of the receptor internalisation/degradation reactions (*i.e.* any reactions involving the parameter β_j or γ_j for $j \in \{6, 27\}$).

Owing to the number of additional variables and reactions that are included in the models after explicitly describing SOCS3 negative feedback, writing down and numerically solving the ODE system for the models would be highly complex and error-prone. To overcome this need to write down the ODEs for the systems, a rule-based modelling approach can be used (Danos *et al.*, 2007). Examples of rule-based modelling software include BioNetGen (Faeder *et al.*, 2009), Simmune (Meier-Schellersheim & Mack, 1999) and Copasi (Mendes *et al.*, 2009). Each software has its own advantages, where Copasi comes with a user friendly GUI

interface involving many inbuilt stochastic simulation algorithms, model analyses (such as sensitivity analysis and the linear noise approximation) and data visualisation methods. Simmune is also a model simulation software and comes with the additional advantages of being able to define molecule types graphically (for example specifying where there are binding sites for other molecules) and importantly, being able to define a 3D space for the model simulation to take place. Models can be simulated in both space and time and the model simulations displayed graphically. BioNetGen comes with a user interface known as RuleBender and several example codes for well-known signalling pathways. The RuleBender interface of BioNetGen is used here, whereby the user inputs a list of reaction rules, along with initial molecular concentrations and parameter values. One can specify for example, that X_3 can bind to any receptor with a free SOCS3 binding site with rate α, and that these receptors are then no longer capable to phosphorylating STATs. Each of the mathematical models in this chapter have been formulated in BioNetGen with the additional SOCS3 reactions. An example BioNetGen code for the SOCS3 HypIL-6 mathematical model is given in Appendix B. Upon writing of the models in the BioNetGen language, the program then generates the ODEs in the background of the simulation and solves them numerically. With the addition of negative feedback by SOCS3, the HypIL-6 model increases from 22 variables to 66 and the IL-27 model from 33 variables to 112. Given the increased dimensionality of the models as well as the increase in the number of parameters, carrying out parameter estimation or model selection using the Bayesian methods used in this chapter would be highly inefficient. The models can however be simulated using parameter values reflective of the posterior distributions generated from the ABC-SMC of the original models. To this end, the models are simulated using four randomly sampled parameter sets from each of the RPE1 posteriors and the Th-1 posteriors, and the outputs $[pS_i]^{T,j}$ for $i \in \{1, 3\}$ and $j \in \{6, 27\}$ (after normalisation) are plotted in Figure 4.24 for both cell types. The parameters δ_k for $k \in \{1, 2, 3\}$ are chosen as 5×10^{-4} such that SOCS3 production begins at around 30 minutes (representative of the peak in pSTAT levels seen in the data) as can be seen in Figure 4.25. The initial concentration of X_3^a, which represents the carrying capacity concentration of SOCS3 that can be produced, is set to 20 nM. The rate at which SOCS3 binds

either receptor in the dimer is set to $\alpha = 5 \times 10^{-3}$ nM^{-1}s^{-1}, a rate slower than the average rate of STAT binding to the receptors based on the binding affinities given by Dittrich *et al.* (2012).

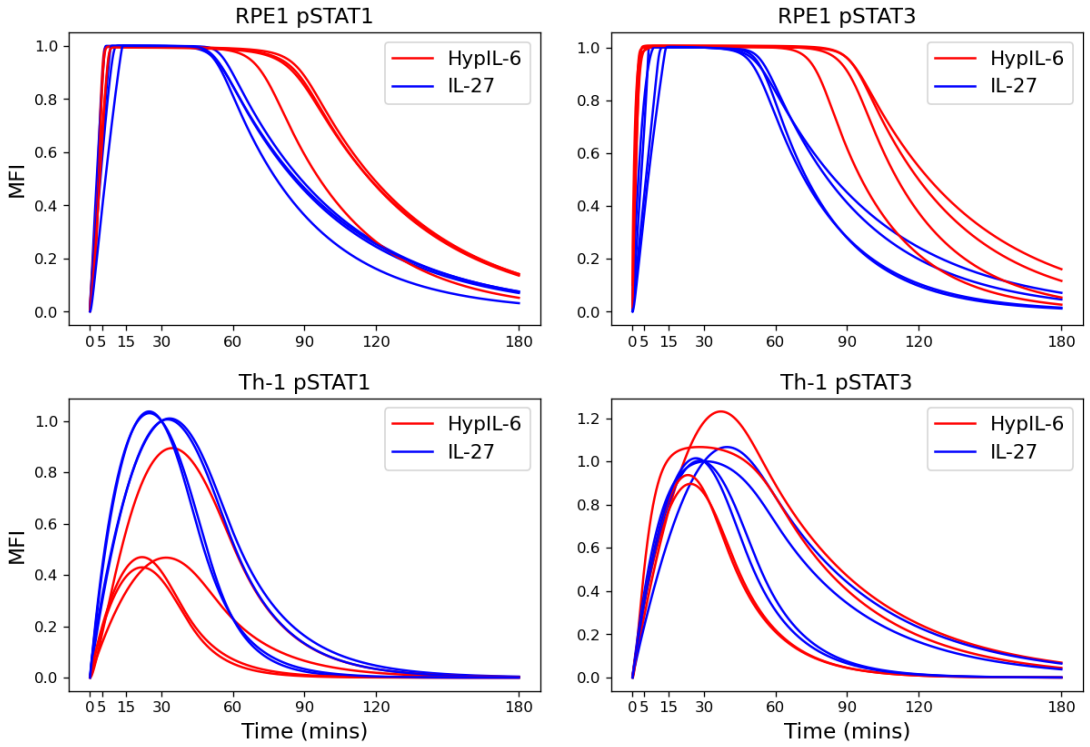

Figure 4.24: Outputs of total phosphorylated STATs from four simulations of the SOCS3 HypIL-6 and IL-27 mathematical models using BioNetGen in both RPE1 and Th-1 cells. The four parameter sets used were randomly sampled from the posterior distributions generated via ABC-SMC.

From Figure 4.24 it can be observed that the model outputs in the RPE1 cells (top row) are not at all reflective of the experimental data, but that the model outputs in the Th-1 cells (bottom row) do, in general, capture the trend of the data. It should be noted that the parameters α and $[X_3^a](0)$ have been fixed in the simulations and hence a better fit to the data might be found if these parameters were able to be calibrated. Calibration of these parameters however, would only be possible if additional data were collected for the MFI of SOCS3, which was not collected in the original experiments. The parameters δ_k for $k \in \{1, 2, 3\}$, are

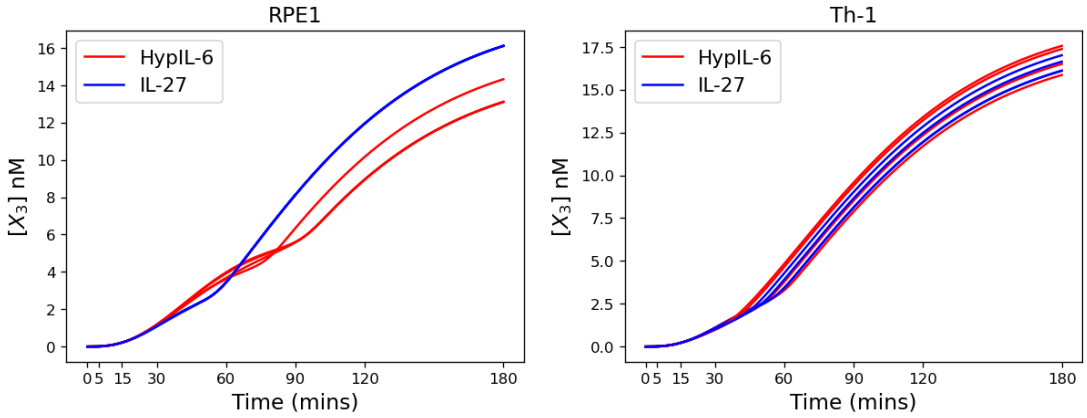

Figure 4.25: Output of SOCS3 ($[X_3]$) from four simulations of the SOCS3 HypIL-6 and IL-27 mathematical models using BioNetGen in both RPE1 and Th-1 cells. The four parameter sets used were randomly sampled from the posterior distributions generated via ABC-SMC.

also fixed at assumed values and hence further exploration of these parameter values may also result in a better model fit.

Given the analysis in the section, in particular the simulations for the Th-1 cells in Figure 4.24, it may be possible that negative feedback is indeed contributing to the down-regulation of receptor containing species in the experiments. However it is believed that for this particular experimental set-up, the down-regulation is mostly due to receptor internalisation. Indeed, an experiment was carried out whereby the cells were treated with cyclohexamide, meaning that SOCS3 molecules could no longer be produced. The STAT phosphorylation profiles were measured in these cells and compared with the profiles from cells in which SOCS3 molecules were able to be produced and this data can be found in Figure 3 supplement 2 of Wilmes *et al.* (2021). It was found that pSTAT1 is minimally affected by cyclohexamide treatment in both cell types under HypIL-6 stimulation, as is pSTAT3 in the Th-1 cells. This implies that SOCS3 molecules are not having a great effect on the deactivation of HypIL-6 induced dimers. Under stimulation with IL-27, when the cells are treated with cyclohexamide, the pSTAT profiles are slightly more sustained, although the experimentalists did not think this was very significant. Even upon treatment with cyclohexamide, there is still a general decrease in pSTAT over time, implying that the receptor

molecules are being internalised from the surface. If it were of particular interest to learn about whether a negative feedback mechanism is indeed in place in the cells which produced the data, one could use simpler models similar to those presented by Reeh *et al.* (2019) to test this hypothesis. In such models, there could be an active and inactive dimer form and therefore SOCS3 negative feedback could be added less explicitly than has been done in this section. This analysis however is out of the scope of this chapter, where the main aim was to identify the cause of the differential signalling by pSTAT1 under HypIL-6 and IL-27.

4.7 Discussion

In this chapter, deterministic mathematical models have been developed to describe the initial steps of intracellular signalling when cells are stimulated with one of two cytokine molecules, HypIL-6 or IL-27. Both cytokines are known to initiate the same intracellular signalling pathway, known as the JAK/STAT pathway, however it has been experimentally observed that IL-27 induces a stronger and more sustained pSTAT1 signal than HypIL-6, whereas the pSTAT3 response is similar between cytokines. ABC-SMC has been used here to calibrate the mathematical models by means of parameter inference in order to determine the cause of the differential pSTAT1 signalling, using experimental data from two cell types. Two hypotheses relating to the way in which receptor molecules are internalised/degraded were also considered and Bayesian model selection was used to infer the most likely mechanism. After further validation of the models, using different experimental datasets, the models were used to make predictions about pSTAT signalling in different receptor and STAT concentration regimes, some of which have been observed in human diseases.

The main result of this chapter is the identification of the specific reactions responsible for the stronger and more sustained pSTAT1 signal upon stimulation of the cells with IL-27. Through differences in the posterior distributions for some parameters obtained using ABC-SMC, one can conclude that pSTAT1 signalling is greater in IL-27 stimulated cells. This is due to the fast rate of binding of STAT1 to IL-27Rα, a receptor which forms one half of the IL-27 induced dimer, but is not part of the HypIL-6 dimer, which is instead formed of two molecules

of GP130. In particular, it was found that STAT1 binds faster to IL-27Rα than to GP130 ($k_{1b}^{+} > k_{1a}^{+}$), whereas STAT3 binds faster to GP130 than to IL-27Rα ($k_{3a}^{+} > k_{3b}^{+}$). A secondary result of the mathematical modelling is the Bayesian model selection. In particular, it was found that, for the data available in this chapter, the receptor molecules are more likely to be internalised/degraded with a constant rate proportional to their concentrations (hypothesis 1) rather than a rate proportional also to the sum of free cytoplasmic phosphorylated STAT1 and STAT3 (hypothesis 2). The aim of this model selection was to determine whether a negative feedback mechanism was in place, whereby the free pSTAT molecules were migrating to the nucleus and promoting the production of negative feedback proteins which would down-regulate the JAK/STAT pathway. Model selection together with the experimental data indicate that such a negative feedback mechanism is unlikely to have contributed to the loss of receptor molecules from the system, and that the receptors were more likely being internalised as part of natural trafficking mechanisms. A limitation of the method used here is that the experimental data used in this chapter only have a time course of up to 180 minutes and this time scale may not be long enough for a negative feedback mechanism to be induced. Therefore in order to test if negative feedback *ever* occurs for such cytokine induced JAK/STAT signalling pathways, one would require data with a significantly longer time course. In Section 4.6, the model hypotheses are further discussed and analysed with relation to other published modelling works which explicitly model negative feedback by including SOCS3 molecules in the model. Implicitly including the negative feedback mechanism in the model, as has been done in this chapter, is less realistic and therefore the result of the model selection is perhaps less trustworthy than it would have been had SOCS3 molecules been included in the model. It is discussed however, that including SOCS3 molecules in the models would hugely increase the complexity of the models and mean that Bayesian model selection would be very difficult and time consuming to conduct. As well as this, further experimental data shows that for this particular model set-up, knocking out SOCS3 molecules from the cells did not have a great effect on pSTAT signalling, and hence negative feedback is not expected to be very important.

Using the parameterised mathematical models, one can simulate pSTAT signalling dynamics under different concentrations of receptors and STATs. This is useful since it has been observed in some diseases such as Crohn's disease and SLE, that STAT1 is highly up-regulated and therefore pSTAT signalling which ultimately controls the fate of the cell, is dysregulated. In this particular situation, corresponding to the fifth row of Figure 4.20, the mathematical models predict that the pSTAT1 response will approximately double in the IL-27 stimulated system and will be up-regulated by a factor of 1.5 under stimulation with HypIL-6. On the other hand, the pSTAT3 response will decrease by approximately 75% in the IL-27 system and by approximately 30% in the HypIL-6 system. With observations from the mathematical model such as these, and knowledge about the specific cellular mechanisms induced by pSTAT1/3 signalling, one may be able to predict the fate of such cells in certain disease scenarios. As well as simulating pSTAT signalling outcomes for different concentration regimes, one could, as an extension of the work in this chapter, simulate the model varying some of the rate constants which may be biologically adjustable if the cells were treated with small molecule therapies. As an example of this, Nishimoto & Kishimoto (2004) discuss an anti IL-6Rα monoclonal antibody as a treatment for inflammatory diseases, which competes with IL-6 for the binding to IL-6Rα. An IL-6Rα bound to the IL-6Rα monoclonal antibody is not able to induce the formation of a GP130 homodimer and hence IL-6 signalling is suppressed. Such small molecule therapies could be included explicitly into the mathematical models developed in this chapter by adding them as extra variables, or implicitly, by reducing the rate at which HypIL-6 binds with GP130. The effect of these type of therapies on the JAK/STAT pathway could then be explored by simulating the mathematical models, similarly to Section 4.5.

The work in this chapter uses deterministic mathematical models and hence an assumption is being made that the concentrations of the reactant species are large enough that random fluctuations in species copy numbers are not important. For the concentrations used here, this assumption is valid, however if there were to be a large down-regulation in any of the species, for example in another disease scenario, then one may need to use a stochastic approach to characterise the system. Finally, a limitation of this work is that there are a large number of

variables and parameters in each of the mathematical models, and hence, in the Bayesian inference, for some of the parameters there is very little learning. For example, the KDEs of the posterior distributions for the ligand binding and dimerisation parameters (top row of Figure 4.13) span several orders of magnitude and hence are not informative. To overcome this problem, one could fix some of the model parameters at experimentally derived values. For example, if there was a lot of confidence in the receptor-ligand binding rates found from the Biacore affinity experiments (discussed in Section 4.3.1), then these rates could be fixed in the modelling. Likewise, if it were possible to experimentally derive any of the other rate constants in the mathematical models, these too could be fixed, resulting in a model with fewer parameters to estimate, yielding less uncertainty in the model predictions.

Chapter 5

Mathematical modelling of FGFR2 ternary complex formation

The fibroblast growth factor receptors are a class of receptor tyrosine kinases, of which there are four members, namely FGFR1-FGFR4, which bind to a family of 18 FGF ligands (Turner & Grose, 2010). Upon ligand binding to the extracellular region, the FGFRs form signalling dimers in which tyrosine sites on the intracellular tails of the receptors become phosphorylated and act as docking sites for various cytoplasmic proteins. This binding of cytoplasmic proteins to activated FGFR homo- or heterodimers induces one of four possible signalling cascades, known as the MAPK, JAK-STAT, PI3K-AKT and Plcγ pathways (Hallinan *et al.*, 2016). These diverse and complex signalling pathways ultimately lead to a cellular response which is regulated by FGFR trafficking, where receptors can be synthesised, internalised into the cell, recycled to the cell membrane and degraded. The cellular response is also controlled by the receptor-ligand binding pair, where it has been observed that different pairings can lead to different outcomes for the cell (Ornitz & Itoh, 2015). The FGFRs have been extensively studied over recent years due to their involvement in many different types of cancer. Aberrant FGFR signalling has been noted in some malignant cell types, which can be a result of over-expression of receptors on the cell surface, or mutations in the receptors leading to ligand-independent activation (Hallinan *et al.*,

2016).

Two cytoplasmic proteins which are involved in some of the aforementioned signalling pathways, and which can both bind to FGFRs are Src homology-2 domain-containing protein tyrosine phosphatase-2 (Shp2) and phospholipase C gamma 1 (Plcγ1). Shp2 is involved in many signalling pathways, including the MAPK, JAK-STAT and PI3K pathways, where it has multiple functionalities, with the ultimate effect of enhancing signal transduction (Qu, 2000). To date, however, Shp2 has not been reported to have an involvement in the Plcγ signalling pathway. In this pathway, Plcγ1 itself binds directly to the intracellular tails of the RTKs and becomes phosphorylated and activated. Activation of Plcγ1 leads to other intracellular events (such as the hydrolysis of a membrane bound protein known as PIP2) and the pathway induced can lead to a range of cellular outcomes, including cell division, migration, survival and death (Emmanouilidi *et al.*, 2017). In this chapter, a possible mechanism of increased signal transduction involving both Shp2 and Plcγ1 will be discussed.

In order to maintain a strong signal through any of the signalling pathways, cytoplasmic proteins must constantly be being recruited to the phosphorylated sites of the membrane bound receptor molecules to form complexes of two or more proteins. The interior of a cell is a crowded environment containing numerous organelles and proteins, and hence free diffusion of molecules is limited, meaning that the probability of an interaction between a receptor molecule and the cytoplasmic protein required for signalling, is low, especially when there are relatively low numbers of one or both molecule types (Cebecauer *et al.*, 2010). An interesting phenomenon that has been observed in the regulation of signalling pathways is the formation of liquid-liquid phase separated (LLPS) droplets containing signalling proteins (Cebecauer *et al.*, 2010). These regions of high concentrations of the proteins required for signalling, could be responsible for the maintenance of many cellular signals. One example of this phase separation involves the T cell receptor, where it has been found that the activation of this receptor leads to the formation of micrometer sized clusters of T cell receptors and other intracellular proteins (Su *et al.*, 2016). Interestingly, the clusters contained many kinases (enzymes which attach a phosphate group to another protein and therefore promote signalling), and few phosphatases (enzymes which remove a phosphate group from

another protein and therefore down-regulate signalling). There are many other examples of the formation of LLPS droplets enhancing cell signalling (Case *et al.*, 2019; Huang *et al.*, 2019; Li *et al.*, 2012; Zhang *et al.*, 2018), however, to date, it has not been investigated whether this phenomenon extends to RTK signalling.

Dr. Chi-Chuan Lin and other members of the group of molecular and cellular biology at the University of Leeds have been interested in exploring whether or not LLPS droplets occur involving RTKs. Through imaging experiments, they observed some level of phase separation involving many types of phosphorylated RTKs such as pEGFR, pFGFR1, pFGFR2 and pVEGFR1 (where an RTK prefixed with the letter "p" implies that it is phosphorylated), with Shp2. As a specific example, they decided to further investigate pFGFR2-Shp2 phase separation where they noticed that Plcγ1 was also able to phase separate with pFGFR2 and Shp2 in the droplets. They theorised, therefore, that these three proteins are capable of forming a ternary complex, in which pFGFR2 binds to Shp2, which subsequently binds with pPlcγ1 (phosphorylated Plcγ1), and that these ternary complexes are weakly bound together, to maintain the LLPS droplets. It is known however, that *both* Shp2 and Plcγ1 are capable of binding to pFGFR2, and in different states of phosphorylation, and hence it was unclear if this ternary complex was certainly the only one in the LLPS state. In this chapter, the formation of the pFGFR2 induced ternary complex is explored using a deterministic mathematical model. Through analysis of the steady state of the model, it is shown that the only possible ternary complex that prevails in the long time dynamics, is the one which was theorised and observed experimentally. The model is solved under different cellular conditions to determine the impact of the concentration of each of the molecular species in the ternary complex, on the level of formation of the ternary complex as a whole.

5.1 Experimental results

In this section, an explanation of the experimental work is given, and some of the experimental results are presented. Firstly, florescently tagged intracellular domains of several RTKs were imaged upon the addition of Shp2, which is known

to bind with each of the RTKs in order to initiate cell signalling pathways. A mutant variety of Shp2, known as the $Shp2_{C459S}$ phosphatase dead trapping mutant, was used in the experiments. As the name suggests, this mutant lacks phosphatase activity and was used so that Shp2 did not dephosphorylate the RTKs in the experiment, where the RTKs are required to be phosphorylated in order to bind with Shp2 or Plcγ1. For brevity, $Shp2_{C459S}$ will be referred to as $Shp2_C$ throughout this chapter. The experiment was set up by purifying the RTKs (separating them from other proteins) and subsequently phosphorylating them with ATP. With the phosphorylated RTKs then in solution, $Shp2_C$ was added to the solution and the mixture incubated at room temperature for 1 minute before imaging. Confocal imaging was used in order to produce a two-dimensional image of the cross section of protein droplets. This experiment was carried out in order to determine whether any of the RTKs were able to phase separate with this protein. The RTKs used in the experiment were three members of the ErbB family of receptors (pEGFR, pHer2 and pHer4), two members of the FGFR family (pFGFR1 and pFGFR2) and two members of the VEGFR family (pVEGFR1 and pVEGFR2). The whole protein was not used, and instead either the cytoplasmic domain (cyto) or the kinase-tail of the RTK was used (hence the proteins are "recombinant", *i.e.* modified or manipulated), and the results are seen in Figure 5.1. There are clearly visible droplets in the presence of pErbB family receptors (top row) and pFGFRs, however the pVEGF receptor family either do not form droplets with $Shp2_C$ or they are very small in relation to those formed with the other receptor types. Given the abundance of droplets formed upon the addition of $Shp2_C$ to pFGFR2, this pairing was chosen to study further in order to explore the possibility of other proteins being involved in the droplets, and the purpose of the formation of such droplets.

Plcγ1 is another cytoplasmic protein, important in the regulation of cell signalling, which binds with RTKs in order to initiate its own signalling pathway. It has not previously been reported to interact with Shp2, but given the importance of Shp2 in many other signalling pathways, this interaction was therefore explored. It was found, again through imaging experiments, that $Shp2_C$ and Plcγ1, although they are able to bind each other, do not phase separate in the absence of the receptor. It can be concluded therefore, that a pFGFR2-$Shp2_C$

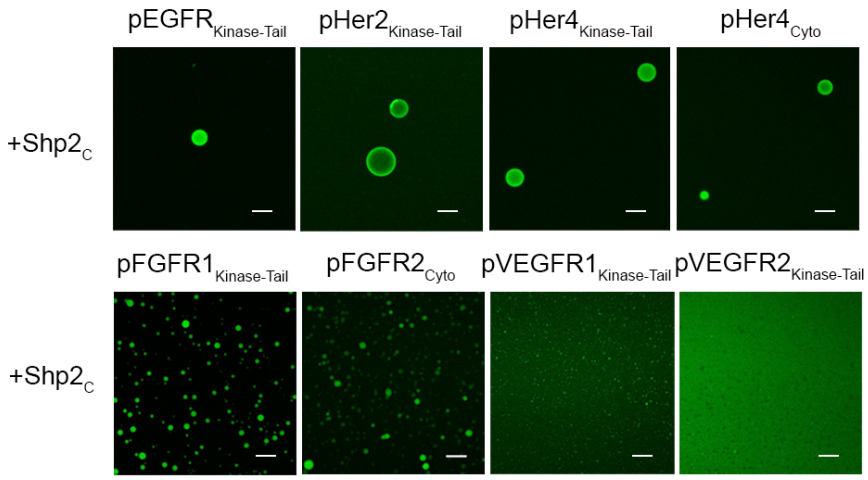

Figure 5.1: Droplet formation of recombinant pEGFR, pFGFR, and pVEGFR (6 μM each) upon adding 30 μM of $Shp2_C$ (Lin *et al.*, 2019). Scale bar = 10 μm.

interaction is essential for phase separation. In order to confirm that at least one other protein is required for phase separation, pFGFR2 was imaged in the absence of any other intracellular proteins and no phase separation was observed.

Given that both $Shp2_C$ and Plcγ1 are able to bind with pFGFR2, it was then theorised that they may be able to bind concurrently with the receptor to form a signalling ternary complex. When the two proteins were added to pFGFR2 containing medium (with the experimental set up similar to that used to produce Figure 5.1), phase separation was again observed, as can be seen in Figure 5.2. In this Figure, each of the first three panels shows the location, via confocal fluorescence imaging, of the proteins pFGFR2, $Shp2_C$ and Plcγ1, separately, and the fourth panel shows the overlay of the images, revealing that all of the proteins are located in the droplets.

Although Figure 5.2 confirms that the proteins co-locate, it does not explicitly confirm the formation of a ternary complex, where it may be possible that the proteins are binding only through binary interactions. To this end, pairwise interactions between the proteins were then characterised in order to test, for example, that two of the proteins do not compete for the same binding site on the third, making a ternary complex impossible. It was demonstrated that pFGFR2 and $Shp2_C$ engage with one another via interactions between the C-terminal tail

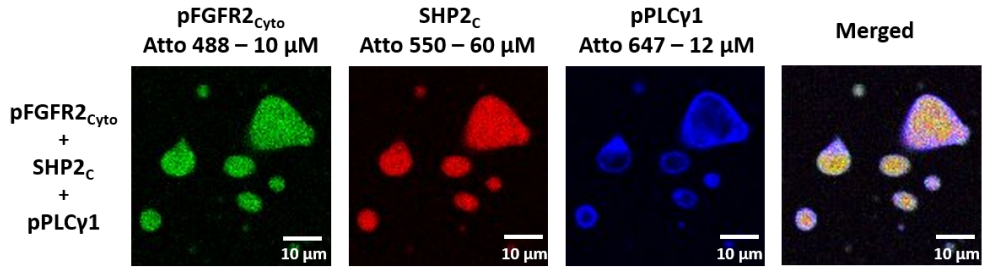

Figure 5.2: *In vitro* phase separation assay using Atto-labelled pFGFR2$_{\text{Cyto}}$ (280 μM), truncated Shp2$_{\text{2SH2}}$ (700 μM) and pPlcγ1$_{\text{2SH2}}$ (250 μM) (Lin *et al.*, 2019). Scale bar = 10 μm.

of the receptor, specifically tyrosine Y769, and the C-terminal Src homology 2 (CSH2) domain of Shp2$_{\text{C}}$ (see Figure 5.3, panel B). The binding between Shp2$_{\text{C}}$ and Plcγ1 is mediated between interactions within the tandem SH2 domains of the two proteins (see Figure 5.3, panel C). Plcγ1 can also be recruited to the pY769 site on the C-terminal tail of the receptor, through its NSH2 domain (see Figure 5.3, panel D). This interaction results in the phosphorylation of Plcγ1 on Y783. Clearly, since Plcγ1 and Shp2$_{\text{C}}$ bind the receptor tail at the same tyrosine residue, both proteins cannot be bound to the receptor at the same time. Therefore, a ternary complex must be comprised of the receptor bound to one of the proteins and the third protein bound to another site on the second protein. It was hypothesised that this ternary complex should be pFGFR2 bound to Shp2$_{\text{C}}$, which in turn is bound to Plcγ1, since this complex can form through non-exclusive surfaces (see Figure 5.3, panel E), although other formulations of the ternary complex are possible. Given that Plcγ1 can become phosphorylated by the receptor, the ternary complex may involve either unphosphorylated or phosphorylated Plcγ1. The binary complexes as well as the hypothesised ternary complex are depicted in Figure 5.3, where the domains are labelled as follows, 1: tyrosine kinase domain, 2: NSH2 domain, 3: CSH2 domain, 4: Phosphatase, and 5: SH3 domain. Dissociation constants (ratios of the backward are forward rate constants for binding) for some of the binary reactions were measured and are given in Table 5.1. It can be seen that, when the receptor is present with only one other species, the binding to the receptor is tighter from Plcγ1 than Shp2$_{\text{C}}$. Also Shp2$_{\text{C}}$ binds with higher affinity to the phosphorylated form of

Plcγ1, and therefore it was postulated that if a ternary complex exists it should involve phosphorylated Plcγ1, and thus the ternary complex of pFGFR2 bound to Shp2$_C$ which is in turn bound to pPlcγ1 is the suspected ternary complex being formed in the LLPS droplets.

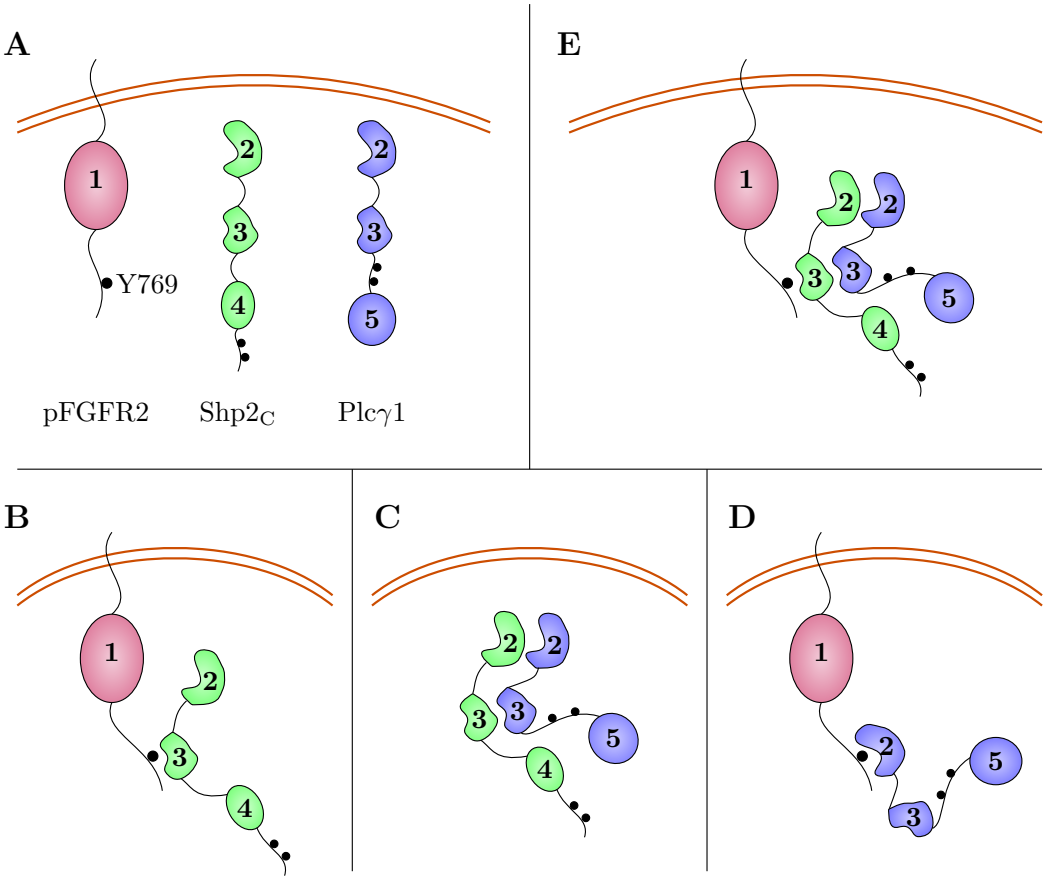

Figure 5.3: Diagrams of the interactions between pFGFR2, Shp2$_C$ and Plcγ1. **A**: The individual molecules in the system, where a black circle represents a phospho-tyrosine residue, 1: tyrosine kinase domain, 2: NSH2 domain, 3: CSH2 domain, 4: Phosphatase, and 5: SH3 domain. **B**: The binary reaction between pFGFR2 and Shp2$_C$. **C**: The binary reaction between Shp2$_C$ and Plcγ1. **D**: The binary reaction between pFGFR2 and Plcγ1. **E**: The hypothesised ternary complex formation between pFGFR2, Shp2$_C$ and Plcγ1.

In order to explore the possibility of ternary complex formation between the species pFGFR2, Shp2$_C$ and Plcγ1, in Section 5.2 a mathematical model is de-

Reaction	K_d value	Unit
Phosphorylated FGFR2 binding $Shp2_C$	25.1	μM
Phosphorylated FGFR2 binding Plcγ1	0.223	μM
$Shp2_C$ binding Plcγ1	1.16	μM
$Shp2_C$ binding phosphorylated Plcγ1	0.48	μM

Table 5.1: Experimentally derived dissociation constants (K_d values) for binary reactions involving the species pFGFR2, $Shp2_C$ and Plcγ1.

veloped to describe the reactions occurring in the experimental system. Then, in Section 5.3, the steady state of the mathematical model is analysed to determine which of the ternary complexes which *could* form, prevail in the long run.

5.2 Mathematical model

In this section, an ordinary differential equation mathematical model for the formation of ternary complexes involving pFGFR2, $Shp2_C$ and Plcγ1 is introduced. The model was formulated based on reactions involving the following species,

- F = unbound unphosphorylated FGFR2,
- pF = unbound phosphorylated FGFR2,
- S = unbound $Shp2_C$,
- $pF \cdot S$ = phosphorylated FGFR2 - $Shp2_C$ complex,
- P = unbound unphosphorylated Plcγ1,
- $pF \cdot P$ = phosphorylated FGFR2 - Plcγ1 complex,
- $pF \cdot pP$ = phosphorylated FGFR2 - phosphorylated Plcγ1 complex,
- pP = unbound phosphorylated Plcγ1,
- $S \cdot P$ = $Shp2_C$ - unphosphorylated Plcγ1 complex,
- $S \cdot pP$ = $Shp2_C$ - phosphorylated Plcγ1 complex,
- $pF \cdot S \cdot P$ = phosphorylated FGFR2 - $Shp2_C$ - unphosphorylated Plcγ1 complex,

- $pF \cdot P \cdot S$ = phosphorylated FGFR2 - unphosphorylated Plcγ1 - Shp2$_C$ complex,
- $pF \cdot S \cdot pP$ = phosphorylated FGFR2 - Shp2$_C$ - phosphorylated Plcγ1 complex,
- $pF \cdot pP \cdot S$ = phosphorylated FGFR2 - phosphorylated Plcγ1 - Shp2$_C$ complex.

Given that only one of either Shp2$_C$ and Plcγ1 can bind the receptor at any one time, and that Plcγ1 can be either phosphorylated or unphosphorylated, there are four possible ternary complexes which could form. The reactions of the model are depicted in Figure 5.4, where the ternary complex coloured in red, is the one which is experimentally hypothesised to exist in the LLPS droplets. The first reaction, at the top of the figure, illustrates the phosphorylation of FGFR2 with rate constant k_1. Although this is in reality a process involving ligand binding and dimerisation, here it is modelled as a spontaneous linear reaction in order to minimise complexity in the model. The second row of the figure shows the reversible binding of Shp2$_C$ to the phosphorylated tail of the receptor, where this reaction is known to occur with dissociation constant

$$K_{d,2} = \frac{k_{-2}}{k_{+2}} = 25.1\ \mu\text{M}.$$

The third row of Figure 5.4 indicates the reversible binding of Plcγ1 to pFGFR2, where

$$K_{d,3} = \frac{k_{-3}}{k_{+3}} = 0.223\ \mu\text{M}.$$

There are two further reactions on this row, whereby instead of dissociating the receptor unphosphorylated (with rate k_{-3}), Plcγ1 could become phosphorylated by the receptor with rate k_4 and then dissociate with rate k_5. These two reactions are irreversible, so that phosphorylated Plcγ1 cannot bind the receptor, and Plcγ1 cannot dephosphorylate whilst bound to the receptor. The phosphorylation and dissociation of Plcγ1 are thought to occur extremely quickly (Wahl *et al.*, 1992), so the rates k_4 and k_5 should be high. The fourth and fifth rows of Figure 5.4 depict the binding of Shp2$_C$ and Plcγ1, where in the fourth row Plcγ1 is

unphosphorylated and the reaction occurs with dissociation constant

$$K_{d,6} = \frac{k_{-6}}{k_{+6}} = 1.16\ \mu\text{M},$$

and in the fifth row, Plcγ1 is phosphorylated and the reaction occurs with dissociation constant

$$K_{d,7} = \frac{k_{-7}}{k_{+7}} = 0.48\ \mu\text{M}.$$

Finally, the bottom two rows of Figure 5.4 show a circuit of reactions in which binary complexes bind with single molecules to form ternary complexes. The phosphorylated receptor can reversibly bind with either of the binary complexes $S \cdot P$ or $S \cdot pP$ through the Shp2$_{\text{C}}$ molecule, where the rates of these reactions are k_{+2} and k_{-2}. Hence, there is no allostery considered in the model, so that the binding of a binary complex through Shp2$_{\text{C}}$ occurs with the same rate as the binding of a single molecule of Shp2$_{\text{C}}$. pFGFR2 can also bind the complex $S \cdot P$ through Plcγ1, again assuming no allostery so that these reversible reactions occur with rate constants k_{+3} and k_{-3}. Since phosphorylated Plcγ1 cannot bind to pFGFR2, the only reaction in the model involving the binding of the binary complex $S \cdot pP$, is through Shp2$_{\text{C}}$. Upon formation of the ternary complex $pF \cdot P \cdot S$, since Plcγ1 is directly bound to the receptor, it can become phosphorylated by the receptor and can subsequently dissociate, whilst still bound to Shp2$_{\text{C}}$. Again, since no allostery is considered in the model, these reactions are assumed to occur with rates k_4 and k_5, respectively. The last set of reversible reactions in the model is the binding of the binary complex $pF \cdot S$ to phosphorylated Plcγ1 (pP), with rates k_{+7} and k_{-7}. Given that the phosphorylation of Plcγ1 on the receptor and subsequent dissociation of pPlcγ1 are reactions supposed to occur very quickly, the complex $pF \cdot P$ will not exist for long enough periods of time such that the reactions $pF \cdot P + S \rightleftharpoons pF \cdot P \cdot S$ would become relevant, and hence they are not included in the model.

Based on the reactions in Figure 5.4, and under the assumption of mass action kinetics, the system can be described by the ODEs (5.1) - (5.14). Square brackets around a species notation denote the concentration of this species with units μM. The ODEs are valid for any time t, with $t \geq 0$, but time has been omitted in the

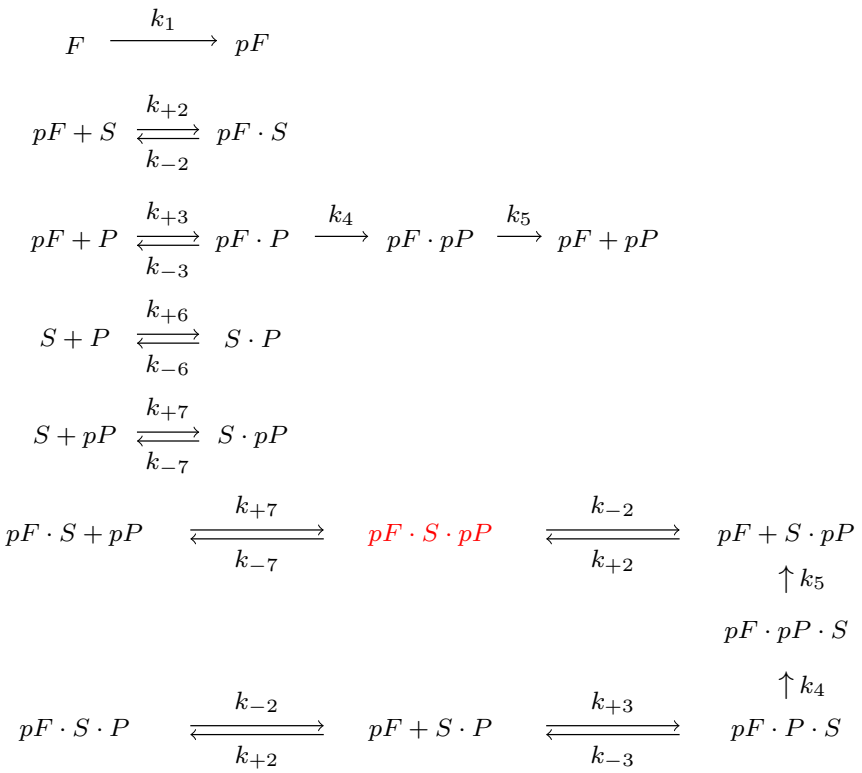

Figure 5.4: A depiction of the molecular reactions which define the mathematical model. In the figure a "·" indicates that the species are bound. The values k associated with the reaction arrows are the rates at which the reaction takes place. The species in red is the experimentally hypothesised ternary complex.

species concentration notation for ease of reading, where for example $[F] = [F](t)$ for all $t \geq 0$.

$$\frac{d[F]}{dt} = -k_1[F] \tag{5.1}$$

$$\begin{aligned}\frac{d[pF]}{dt} &= k_1[F] - k_{+2}[pF]([S] + [S \cdot P] + [S \cdot pP]) \\ &\quad + k_{-2}([pF \cdot S] + [pF \cdot S \cdot P] + [pF \cdot S \cdot pP]) - k_{+3}[pF]([P] + [S \cdot P]) \\ &\quad + k_{-3}([pF \cdot P] + [pF \cdot P \cdot S]) + k_5[pF \cdot pP] + k_5[pF \cdot pP \cdot S]\end{aligned} \tag{5.2}$$

$$\frac{d[S]}{dt} = -k_{+2}[pF][S] + k_{-2}[pF \cdot S] - k_{+6}[S][P] + k_{-6}[S \cdot P]$$

Parameter/IC	Description	Unit
$[F]^T$	Total FGFR2 concentration	μM
$[P]^T$	Total Plcγ1 concentration	μM
$[S]^T$	Total Shp2$_{\mathrm{C}}$ concentration	μM
k_1	Rate of FGFR2 phosphorylation	s^{-1}
k_{+2}	Rate of pFGFR2 binding to Shp2$_{\mathrm{C}}$	μM^{-1}s^{-1}
k_{-2}	Rate of pFGFR2 unbinding from Shp2$_{\mathrm{C}}$	s^{-1}
k_{+3}	Rate of pFGFR2 binding to Plcγ1	μM^{-1}s^{-1}
k_{-3}	Rate of pFGFR2 unbinding from Plcγ1	s^{-1}
k_4	Rate of Plcγ1 phosphorylation	s^{-1}
k_5	Rate of pFGFR2 unbinding from pPlcγ1	s^{-1}
k_{+6}	Rate of Shp2$_{\mathrm{C}}$ binding to Plcγ1	μM^{-1}s^{-1}
k_{-6}	Rate of Shp2$_{\mathrm{C}}$ unbinding from Plcγ1	s^{-1}
k_{+7}	Rate of Shp2$_{\mathrm{C}}$ binding to pPlcγ1	μM^{-1}s^{-1}
k_{-7}	Rate of Shp2$_{\mathrm{C}}$ unbinding from pPlcγ1	s^{-1}

Table 5.2: Definitions and units for the rate constants and initial concentrations in the FGFR2 mathematical model. A superscript T denotes the "total" (or initial) concentration of a molecule.

$$- k_{+7}[S][pP] + k_{-7}[S \cdot pP] \tag{5.3}$$

$$\frac{d[pF \cdot S]}{dt} = k_{+2}[pF][S] - k_{-2}[pF \cdot S] - k_{+7}[pF \cdot S][pP] + k_{-7}[pF \cdot S \cdot pP] \tag{5.4}$$

$$\frac{d[P]}{dt} = -k_{+3}[pF][P] + k_{-3}[pF \cdot P] - k_{+6}[S][P] + k_{-6}[S \cdot P] \tag{5.5}$$

$$\frac{d[pF \cdot P]}{dt} = k_{+3}[pF][P] - k_{-3}[pF \cdot P] - k_4[pF \cdot P] \tag{5.6}$$

$$\frac{d[pF \cdot pP]}{dt} = k_4[pF \cdot P] - k_5[pF \cdot pP] \tag{5.7}$$

$$\frac{d[pP]}{dt} = k_5[pF \cdot pP] - k_{+7}[pP]([S] + [pF \cdot S])$$

$$+ k_{-7}([S \cdot pP] + [pF \cdot S \cdot pP]) \tag{5.8}$$

$$\frac{d[S \cdot P]}{dt} = k_{+6}[S][P] - k_{-6}[S \cdot P] - k_{+2}[pF][S \cdot P] + k_{-2}[pF \cdot S \cdot P] - k_{+3}[pF][S \cdot P] + k_{-3}[pF \cdot P \cdot S] \tag{5.9}$$

$$\frac{d[S \cdot pP]}{dt} = k_{+7}[pP][S] - k_{-7}[S \cdot pP] - k_{+2}[pF][S \cdot pP] + k_{-2}[pF \cdot S \cdot pP] + k_5[pF \cdot pP \cdot S] \tag{5.10}$$

$$\frac{d[pF \cdot S \cdot P]}{dt} = k_{+2}[pF][S \cdot P] - k_{-2}[pF \cdot S \cdot P] \tag{5.11}$$

$$\frac{d[pF \cdot P \cdot S]}{dt} = k_{+3}[pF][S \cdot P] - k_{-3}[pF \cdot P \cdot S] - k_4[pF \cdot P \cdot S] \tag{5.12}$$

$$\frac{d[pF \cdot S \cdot pP]}{dt} = k_{+7}[pF \cdot S][pP] - k_{-7}[pF \cdot S \cdot pP] + k_{+2}[pF][S \cdot pP] - k_{-2}[pF \cdot S \cdot pP] \tag{5.13}$$

$$\frac{d[pF \cdot pP \cdot S]}{dt} = -k_5[pF \cdot pP \cdot S] + k_4[pF \cdot P \cdot S] \tag{5.14}$$

Conservation expressions can be written for the total concentration of F ($[F]^T$), S ($[S]^T$) and P ($[P]^T$), since it is assumed that the experimental volume of the system does not change with time. Therefore, one can write that

$$[F]^T = [F] + [pF] + [pF \cdot S] + [pF \cdot P] + [pF \cdot pP] + [pF \cdot S \cdot P] + [pF \cdot P \cdot S] + [pF \cdot S \cdot pP] + [pF \cdot pP \cdot S], \tag{5.15}$$

$$[S]^T = [S] + [pF \cdot S] + [S \cdot P] + [S \cdot pP] + [pF \cdot S \cdot P] + [pF \cdot P \cdot S] + [pF \cdot S \cdot pP] + [pF \cdot pP \cdot S], \text{ and,} \tag{5.16}$$

$$[P]^T = [P] + [pF \cdot P] + [pF \cdot pP] + [pP] + [S \cdot P] + [S \cdot pP] + [pF \cdot S \cdot P] + [pF \cdot P \cdot S] + [pF \cdot S \cdot pP] + [pF \cdot pP \cdot S]. \tag{5.17}$$

The above equations hold since the total concentration of a molecule at any time t is the sum of the concentrations of all species containing this molecule at that

time point. To explore the model dynamics, the model was numerically solved to reflect a further experimental setup of the same type as generated Figure 5.2, where, initially, Shp2_C and Plcγ1 were added to medium containing pFGFR2. Since in an *in vivo* scenario however, FGFR2 would exist in an unphosphorylated and a phosphorylated state, the model includes the phosphorylation of FGFR2 as a reaction and hence the model initial conditions for the receptor states are $[F](0) \neq 0$ and $[pF](0) = 0$. The initial concentrations of the species F, S and P were therefore non-zero and all other initial concentrations were zero. Specifically, in this experiment, the initial concentration of FGFR2 was, $[F](0) = 0.3$ μM, the initial concentration of Shp2_C was, $[S](0) = 147$ μM, and the initial concentration of Plcγ1 was $[P](0) = 140$ μM. These values are reported in Table 5.3.

Initial concentration	Value	Unit
$[F](0) = [F]^T$	0.3	μM
$[P](0) = [P]^T$	140	μM
$[S](0) = [S]^T$	147	μM

Table 5.3: Initial concentrations of FGFR2, Plcγ1 and Shp2_C, used in the experiment which the mathematical model is solved to reflect.

Figure 5.5 shows this model solution, where the dissociation rate constants were set as $k_{-2} = k_{-3} = k_{-6} = k_{-7} = k_5 = 10^{-1}$ s^{-1}, the phosphorylation rate constants were set as $k_1 = k_4 = 10^0$ s^{-1} and the association rate constants were fixed using the K_d values in Table 5.1. The figure is split into three panels where the left panel shows the concentration of monomeric species, the middle panel shows the dimeric species and the right panel shows the trimeric species (ternary complexes). The insets in the first two subplots are to show the dynamics of species at low concentrations.

From each panel of Figure 5.5 it can be seen that the phosphorylated form of Plcγ1 prevails in the long term dynamics. This behaviour can be expected when looking at the model diagram in Figure 5.4, since Plcγ1 can become phosphorylated on the receptor, but no dephosphorylation reaction is included in the model (for simplicity so as not to include further species in the model). From the inset in

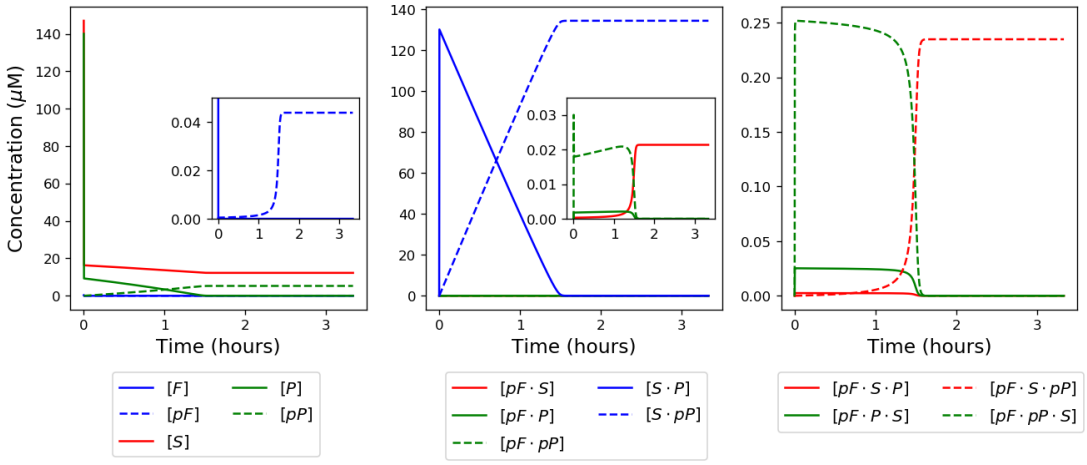

Figure 5.5: A numerical solution to the FGFR2 mathematical model using the experimental initial concentrations, $[F](0) = 0.3$ μM, $[S](0) = 147$ μM and $[P](0) = 140$ μM. The rate constants were set as $k_{-2} = k_{-3} = k_{-6} = k_{-7} = k_5 = 10^{-1}$ s^{-1}, $k_1 = k_4 = 10^0$ s^{-1} and the association rate constants were fixed using the K_d values in Table 5.1.

the first subplot, one can note that although $[F]$ goes rapidly to zero, due to the fast irreversible phosphorylation of this species, $[pF]$ appears to reach a steady state by the end of the time course considered. As indicated by the inset in the second subplot, $pF \cdot P$ and $pF \cdot pP$ are intermediary species in the formation of pP, where the concentration of pP also appears to reach steady state by the end of the time course (first subplot). $S \cdot P$ is a species which initially forms but is replaced by $S \cdot pP$ in later times, when more pP has been produced. Finally, from the last subplot, showing the ternary complexes, it can be seen that although $pF \cdot pP \cdot S$ and $pF \cdot P \cdot S$ form in relatively large concentrations at the beginning of the time course, by the end of the time course the only non-zero concentration of a ternary complex is that for the complex $pF \cdot S \cdot pP$. The concentration of this species also appears to have reached steady state, thus giving some verification for the hypothesis that this ternary complex may form experimentally. Figure 5.5 is an example of a model solution for specific rate constants and initial concentrations, and so in Section 5.3 the steady state of the system will be explored in more detail for general model parameters.

5.3 Steady states

For a system of differential equations

$$\frac{d\mathbf{X}}{dt} = \mathbf{F}(\mathbf{X}),$$

with $\mathbf{X} = (x_1, \ldots, x_n)^T$ and $\mathbf{F}(\mathbf{X}) = (f_1(x_1, \ldots, x_n), \ldots, f_n(x_1, \ldots, x_n))^T$, a steady state of such a system is a constant solution $\mathbf{X}^*$ such that $\mathbf{F}(\mathbf{X}^*) = \mathbf{0}$ (Allen (2007) and Section 2.3). For a biological system, such as the FGFR2 system modelled by Equations (5.1) - (5.14), if the system is in steady state, then the concentrations of each of the species are no longer changing with time. From Figure 5.5 it can be seen that after approximately 2 hours, the concentrations of each of the species remain constant. This indicates that the system of ODEs may have a steady state in which some species take positive values and others are zero. Steady states of a system of ODEs can be found by setting the right hand side of each of the ODEs to zero and solving the resulting equations simultaneously. For some ODE systems, usually with few variables and few underlying reactions, it is possible to solve for the steady state by hand, however for larger more complex systems this can become infeasible. In this case, symbolic mathematical software can be used to solve the system. Here, *Wolfram Mathematica*, was used to find steady state solutions for the system of Equations (5.1) - (5.14). In total, 8 sets of implicit solutions were found, whereby explicit steady state expressions were not given for every variable, but the solution for some variables could be written in terms of other variables and parameters. In each of the 8 solution sets, many of the variables go to zero in the steady state. For example, $[F]^* = 0$ in all sets of solutions, which is expected since the only reaction in the system involving F, is the phosphorylation of the species to form pF and so in the long term dynamics, this species is exhausted.

Of the 8 sets of steady state solutions, only 3 are biologically feasible, where all of the variables in these 3 sets are either equal to zero or are capable of taking positive values. Given that the variables represent concentrations of molecular species, a negative value is meaningless here. The first biologically relevant implicit steady state solution set is,

$$\begin{cases} [F]^* = [pF]^* = [pF \cdot S]^* = [pF \cdot P]^* = [pF \cdot pP]^* = [pF \cdot S \cdot P]^* \\ \qquad = [pF \cdot P \cdot S]^* = [pF \cdot S \cdot pP]^* = [pF \cdot pP \cdot S]^* = 0, \\ [P]^* = \dfrac{k_{-6}[S \cdot P]^*}{k_{+6}[S]^*}, \\ [S \cdot pP]^* = \dfrac{k_{+7}[S]^*[pP]^*}{k_{-7}}. \end{cases} \tag{5.18}$$

The second set is,

$$\begin{cases} [F]^* = [pF]^* = [S]^* = [pF \cdot S]^* = [pF \cdot P]^* = [pF \cdot pP]^* \\ \qquad = [S \cdot P]^* = [S \cdot pP]^* = [pF \cdot S \cdot P]^* = [pF \cdot P \cdot S]^* \\ \qquad = [pF \cdot S \cdot pP]^* = [pF \cdot pP \cdot S]^* = 0, \\ [P]^* \neq 0, \\ [pP]^* \neq 0. \end{cases} \tag{5.19}$$

Finally the third solution set for which all variables may take zero or positive values is,

$$\begin{cases} [F]^* = [P]^* = [pF \cdot P]^* = [pF \cdot pP]^* = [S \cdot P]^* = [pF \cdot S \cdot P]^* \\ \qquad = [pF \cdot P \cdot S]^* = [pF \cdot pP \cdot S]^* = 0, \\ [pF \cdot S]^* = \dfrac{k_{+2}[pF]^*[S]^*}{k_{-2}}, \\ [S \cdot pP]^* = \dfrac{k_{+7}[S]^*[pP]^*}{k_{-7}}, \\ [pF \cdot S \cdot pP]^* = \dfrac{k_{+2}k_{+7}[pF]^*[S]^*[pP]^*}{k_{-2}k_{-7}}. \end{cases} \tag{5.20}$$

In solution set (5.18), all of the variables for species involving FGFR2 are equal to zero. Clearly, since there is no synthesis or degradation of any species in the model, this steady state can only be reached when the concentration of all FGFR2 containing species is zero initially. In this situation, the only species which can exist are S, P, $S \cdot P$, pP and $S \cdot pP$. Since pP is formed through phosphorylation

of P on the receptor, the only way in which the species containing pP can be non-zero in this steady state is if either $[pP]$ or $[S \cdot pP]$ were non-zero initially. When considering the formation of a ternary complex, this steady state is not relevant as one can assume that each of S, P and F should be present initially.

The implicit steady state described by Equation (5.19) is similar to that described by Equation (5.18), in that again all species containing FGFR2 are equal to zero. Here however, all species containing Shp2_C are also equal to zero, and for the same reasoning as above, one can conclude that this steady state can only be reached if all species containing either F or S are zero initially. The only species which prevail in this steady state are P and pP and clearly, due to the lack of receptor in the system, the concentration of both of these species can only be positive if they take positive values initially. Given that there are no other species present, there would be no reactions occurring, and hence the steady state values for these species would be equal to their respective initial concentrations. This steady state therefore, can be thought of not only as a point but as a whole plane in the two dimensions, $[P]$ and $[pP]$. Again, this steady state is not relevant for the exploration of ternary complex formation.

Solution set (5.20) is the only implicit steady state which can result in zero or positive values for all species, where there can be positive values for species involving each of FGFR2, Shp2_C and Plcγ1. Crucially, there is a non-zero value for one of the four possible ternary complexes in the steady state, and moreover, it is the ternary complex which was hypothesised to form through the experimental work, namely $pF \cdot S \cdot pP$. Thus, through analysis of the steady state, even when only implicit solutions are available, it can be shown that there is only one steady state in which all species can take either the value zero or positive values, which can be reached when each of F, S and P are initially non-zero, as in the experimental set-up. The six species which prevail in this steady state are pF, S, $pF \cdot S$, pP, $S \cdot pP$ and $pF \cdot S \cdot pP$. Although the steady state (5.20) is implicit, and there are only 3 expressions in the 6 prevailing variables, the conservation equations (5.15) - (5.17) can be used as additional expressions. In particular, in the steady state, since many of the variables go to zero, the conservation equations become

$$[F]^T = [pF]^* + [pF \cdot S]^* + [pF \cdot S \cdot pP]^*, \tag{5.21}$$

$$[S]^T = [S]^* + [pF \cdot S]^* + [S \cdot pP]^* + [pF \cdot S \cdot pP]^*, \text{ and,} \tag{5.22}$$

$$[P]^T = [pP]^* + [S \cdot pP]^* + [pF \cdot S \cdot pP]^*. \tag{5.23}$$

Therefore, combining of Equations (5.21) - (5.23) with the steady state Equations (5.20) results in six equations in six variables, which can be solved to give analytic expressions for each variable in steady state. These equations can be solved using *Wolfram Mathematica*, however the solutions are complex and unwieldy and thus difficult to analyse. In the following section therefore, an efficient method of numerically solving polynomial systems is discussed and applied to the steady state equations.

5.3.1 Numerical homotopy continuation

Numerical algebraic geometry is an area of mathematics which has been in development since the late 1990s (Sommese *et al.*, 2005) with the aim to find solutions of systems of polynomial equations, usually for which symbolic solutions are difficult to compute. A numerical continuation method, known as *homotopy continuation* is applied which is described briefly here based on the works by Sommese *et al.* (2005) and Li (1997) as well as the user manuals and associated publications for the programs *Bertini* (Bates, 2012; Bates *et al.*, 2013) and *Paramotopy* (Bates *et al.*, 2018). These two programs will be used in this chapter to conduct homotopy continuation and subsequently parameter homotopy continuation in order to study the steady states of the FGFR2 system. The aim of homotopy continuation is to find the isolated solutions of a system of n polynomial equations $\mathbf{P}(\mathbf{x}) = \mathbf{0}$ in n unknowns, known as a square polynomial system. Denoting $\mathbf{P} = (p_1, \ldots, p_n)$ and $\mathbf{x} = (x_1, \ldots, x_n)$, the aim is to find the solutions of the system of equations

$$p_1(x_1, \ldots, x_n) = 0$$

$$\vdots \tag{5.24}$$
$$p_n(x_1, \ldots, x_n) = 0.$$

Homotopy continuation is a method proven (by Garcia & Zangwill (1979) and Drexler (1977), proof not discussed here) to be able to numerically find *all* solutions of (5.24). The general idea is to firstly define a separate system of polynomials, known as the *start system*, $\mathbf{Q}(\mathbf{x}) = (q_1(\mathbf{x}), \ldots, q_n(\mathbf{x})) = \mathbf{0}$ for which the solutions are trivially known. A homotopy $H(\mathbf{x}, t)$ is defined such that

$$H(\mathbf{x}, t) = \gamma t \mathbf{Q}(\mathbf{x}) + (1 - t)\mathbf{P}(\mathbf{x}), \tag{5.25}$$

where $\gamma \in \mathbb{C}$ is randomly chosen and $t \in [0, 1]$ is known as the *continuation parameter*, which deforms the system from the start system $\mathbf{Q}(\mathbf{x})$ to the target system $\mathbf{P}(\mathbf{x})$ as it varies from 1 to 0. If $\mathbf{Q}(\mathbf{x}) = \mathbf{0}$ is chosen to satisfy the following three conditions, then the solution paths can be traced from the start system to the target system, and the solutions of $\mathbf{P}(\mathbf{x}) = \mathbf{0}$ are hence found.

1. **Triviality:** The solutions of $\mathbf{Q}(\mathbf{x}) = \mathbf{0}$ are known.
2. **Smoothness:** There are a finite number of smooth solution paths defined by $H(\mathbf{x}, t) = \mathbf{0}$ and parametrised by $t \in (0, 1]$.
3. **Accessibility:** Every isolated solution of $H(\mathbf{x}, 0) = \mathbf{P}(\mathbf{x}) = \mathbf{0}$ can be reached by a path which starts at a solution of $H(\mathbf{x}, 1) = \mathbf{Q}(\mathbf{x}) = \mathbf{0}$.

It is possible for $\mathbf{Q}(\mathbf{x}) = \mathbf{0}$ to have more solutions than $\mathbf{P}(\mathbf{x}) = \mathbf{0}$ in which case some of the solution paths tend to infinity as $t \to 0$. A key point of the method is that with probability 1, the solution paths will not collide in the interval $t \in (0, 1]$. Strikingly, almost any value of $\gamma \in \mathbb{C}$ will allow for the continuation method to work, and hence in the software which will be used here, γ is chosen at random. There are several examples of homotopies which can be used, but the most simple example is known as the total-degree homotopy. Denoting by d_i the degree of the polynomial p_i for $i = 1, \ldots, n$, a system $\mathbf{Q}(\mathbf{x}) = \mathbf{0}$ which is easy to solve and has $D = d_1 \cdots d_n$ solutions is the system

$$q_i(\mathbf{x}) = x_i^{d_i} - 1 \quad \text{for } i = 1, \ldots, n.$$

This choice of start system used with the homotopy (5.25) is the default method used by the software *Bertini*.

Given a start system, such as the total-degree homotopy, predictor-corrector methods are then used to track the solutions from $\mathbf{Q}(\mathbf{x}) = \mathbf{0}$ to $\mathbf{P}(\mathbf{x}) = \mathbf{0}$. Although there are many different predictor-corrector methods, one of the most simple methods, used by *Bertini*, is the Euler-Newton method. The method begins with $t_0 = 1$ and $\mathbf{r}_0$ a known initial value solving $H(\mathbf{r}_0, t_0) = \mathbf{Q}(\mathbf{x}) = \mathbf{0}$ and then successive values $\mathbf{r}_1, \mathbf{r}_2, \ldots$ are computed at $t_1 > t_2 > \cdots > 0$. The predictor part of the continuation is the Euler method, commonly used for numerically solving ordinary differential equations (Butcher & Goodwin, 2008). Here it is formulated as follows

$$\mathbf{r}_{i+1} = \mathbf{r}_i - JH(\mathbf{r}_i, t_i)^{-1} \frac{\partial H(\mathbf{r}_i, t_i)}{\partial t} \Delta t_i,$$

where $JH(\mathbf{x}, t)$ is the Jacobian matrix with respect to the variables $\mathbf{x}$, $\frac{\partial H}{\partial \mathbf{x}}$ and $\Delta t_i = t_{i+1} - t_i$. Once a new value $\mathbf{r}_{i+1}$ has been predicted, it is then corrected using Newton's method for $H(\mathbf{x}, t_i)$ starting with $\mathbf{x}_0 = \mathbf{r}_{i+1}$, where

$$\mathbf{x}_{i+1} = \mathbf{x}_i - JH(\mathbf{x}_i, t_{i+1})^{-1} H(\mathbf{x}_i, t_{i+1}).$$

The value $\mathbf{r}_{i+1}$ is then replaced with the corrected value after a small number of iterations of Newton's method and the whole procedure is continued for decreasing values of t. A geometric interpretation of the Euler-Newton path tracking method is given in Figure 5.6. Each line represents a solution path tracked from the start system at $t = 1$ to the target system at $t = 0$, where green lines represent convergent paths and red dashed lines represent divergent paths, *i.e.* paths which start at a solution for $\mathbf{Q}(\mathbf{x}) = \mathbf{0}$ but for which there is no corresponding solution in $\mathbf{P}(\mathbf{x}) = \mathbf{0}$. The Euler prediction step corresponds to a move along the line tangent to the current point on the solution curve and the Newton correction brings the new point closer to the true solution. More details on the homotopy continuation method, specifically with relation to the program *Bertini* are given by Bates *et al.* (2013).

Bertini can be used to find all solutions to the set of polynomials defined

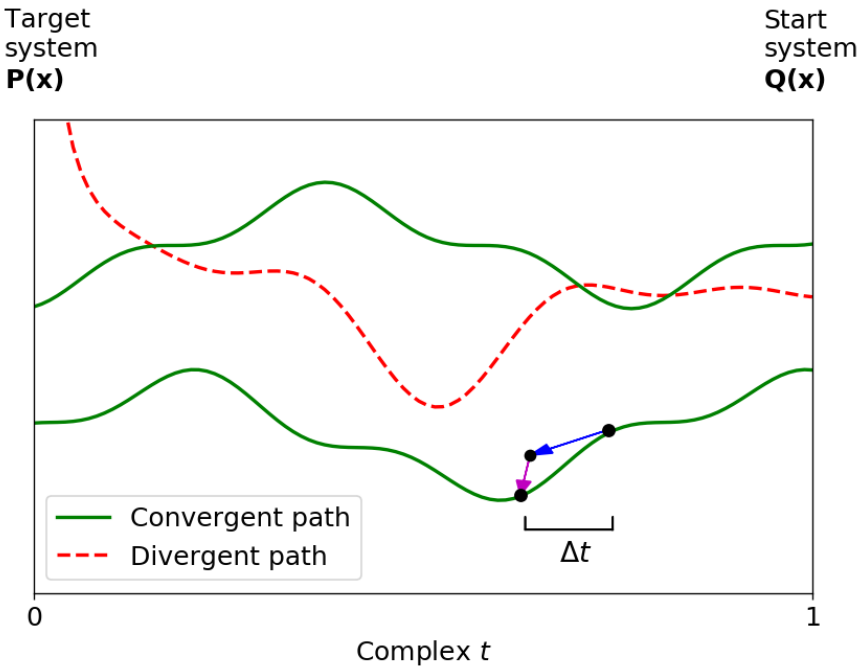

Figure 5.6: A figure illustrating the Euler-Newton homotopy continuation method used by *Bertini* where the lines represent solutions paths from the start system at $t = 1$ to the target system at $t = 0$. The blue arrow represents the Euler prediction step in the method and the magenta arrow represents the Newton correction. Figure inspired by similar figures by Bates *et al.* (2013) and Bates *et al.* (2018).

by Equations (5.20) and Equations (5.15) - (5.17). The homotopy continuation method can be applied by the program for this system of equations using the experimental initial concentrations (Table 5.3) and experimentally derived K_d values (Table 5.1) for $K_{d,2}$ and $K_{d,7}$. In total, four sets of real non-singular solutions were found, meaning that four of the paths from the start system did not diverge to infinity. The values of the variables for these four solution sets are given in Table 5.4 and it can be seen that there is only one solution set (set 2) for which all of the variables take a positive value. Therefore set 2 is the only biologically feasible numerical steady state of the FGFR2 model and, as expected, the values of the variables from the table match with those seen in Figure 5.5 towards the end of the time course.

The steady state defined by set 2 in Table 5.4 is the numerical steady state

Variable	Set 1	Set 2	Set 3	Set 4
$[pF]^*$	-1.72×10^{2}	4.28×10^{-2}	-1.72×10^{2}	4.28×10^{-2}
$[S]^*$	9.84×10^{-1}	1.22×10^{1}	-2.10×10^{0}	-5.74×10^{0}
$[pF \cdot S]^*$	-6.73×10^{0}	2.14×10^{-2}	1.44×10^{1}	-1.00×10^{-2}
$[pP]^*$	-1.28×10^{1}	5.27×10^{0}	5.27×10^{0}	-1.28×10^{1}
$[S \cdot pP]^*$	-2.61×10^{1}	1.34×10^{2}	-2.30×10^{1}	1.52×10^{2}
$[pF \cdot S \cdot pP]^*$	1.79×10^{2}	2.35×10^{-1}	1.58×10^{2}	2.66×10^{-1}

Table 5.4: Solution sets (with units μM) to Equations (5.20) and (5.15) - (5.17) found by numerical homotopy continuation in *Bertini* using the experimental initial concentrations (Table 5.3) and experimentally derived K_d values (Table 5.1) for $K_{d,2}$ and $K_{d,7}$.

reached by the system under the specific experimental conditions. In reality however, the concentrations of each of the species, FGFR2, Shp2 and Plcγ1 will vary based on the cell type and the rates of production and degradation of each species. Likewise, the experimentally derived K_d values may differ in different conditions, and therefore it is important to consider the steady state of the system not only for the experimental conditions but also for more general conditions. For this type of problem, where the same polynomial system is to be solved for differing values of the coefficients, an efficient method can be used, known as *parameter* homotopy continuation. For a system of polynomials, $\mathbf{P}(\mathbf{x}, \mathbf{y})$, parametrised by some parameters $\mathbf{y} \in \mathcal{Y}$, where $\mathcal{Y}$ may be a very large number of sets of parameter values, parameter homotopy continuation occurs in two steps. The first step is to draw random parameter values $\mathbf{y}_0 \in \mathbb{C}$ and with these parameters, carry out a general homotopy continuation as described above (this can be done using *Bertini* or using *Paramotopy*, with which the second step can also be carried out). Theory then guarantees that, for any randomly chosen starting parameter set, the number of finite, isolated solutions of $\mathbf{P}(\mathbf{x}, \mathbf{y})$ is the same for all $\mathbf{y} \in \mathcal{Y}$ (Bates *et al.*, 2018). The second step of the method therefore, is to follow the solutions from $\mathbf{y}_0$ to each $\mathbf{y}_i \in \mathcal{Y}$ using the homotopy

$$H(\mathbf{x}, t) = t\mathbf{P}(\mathbf{x}, \mathbf{y}_0) + (1 - t)\mathbf{P}(\mathbf{x}, \mathbf{y}_i).$$

The major benefit of this method is that one must only run step 1 (the general homotopy continuation) once, before running step 2 (the parameter homotopies) for each parameter set and it is guaranteed that all solutions will be found. Since all paths in step 2 start from the same parameter set $\mathbf{y}_0$, in theory, all of the parameter homotopies in step 2 can be carried out in parallel, greatly reducing the computational cost and time required to solve such a parametrised polynomial system. *Paramotopy* is a program which can carry out both steps, after being given an input system of polynomials and a grid of parameter values with which to solve the system.

To this end, the amplitudes of each variable in steady state were computed using *Paramotopy* where the parameters were varied within the ranges given in Table 5.5. Given that experimentally it is often only possible to measure K_d values of reactions, and that the expressions (5.20) can be written in terms of the K_d values, $K_{d,2}$ and $K_{d,7}$, these parameters are considered here instead of the individual association and dissociation rate constants. The parameters were varied in pairs around the experimental values (see Table 5.3 and Table 5.1), and the steady state computed, where the parameters not being varied were fixed at $[F]^T = 10^{-1}$, $[P]^T = 10^2$, $[S]^T = 10^2$, $K_{d,2} = 25.1$ and $K_{d,7} = 0.48$ (such that they are reflective of the experimental values), all with units μM. 50 points, r, were uniformly sampled within the range given for each parameter being varied and the parameter used to compute the steady state concentrations was then 10^r. The results are seen in Figures 5.7 and 5.8 for the variable $pF \cdot S \cdot pP$, since this is the complex of interest here (the ternary complex). Figure 5.7 shows the parameter pairings involving $[F]^T$ and Figure 5.8 shows all other parameter pairings, where the results are separated into two figures due to the different scales on the colour bar. In each individual heatmap, the colour represents the concentration of the ternary complex, as indicated by the colour bar, with units μM. The x and y-axes show the logarithm base 10 of the μM value of the parameter labelled.

Figure 5.7 shows the concentration of the ternary complex when the initial concentration of FGFR2 is varied with the other initial concentrations and the two K_d values. From all four subplots it can be seen that for most initial concentrations of FGFR2, there is a low production of the ternary complex, however towards the higher end of the range for $[F]^T$, the concentration of the ternary

Parameter/IC	Range	Unit
$[F]^T$	10^r, where $r \in [-2, 0]$	μM
$[P]^T$	10^r, where $r \in [1, 3]$	μM
$[S]^T$	10^r, where $r \in [1, 3]$	μM
$K_{d,2}$	10^r, where $r \in [0, 2]$	μM
$K_{d,7}$	10^r, where $r \in [-2, 0]$	μM

Table 5.5: Ranges for each of the parameters in the steady state of the FGFR2 mathematical model in which the ternary complex $pF \cdot S \cdot pP$ is present.

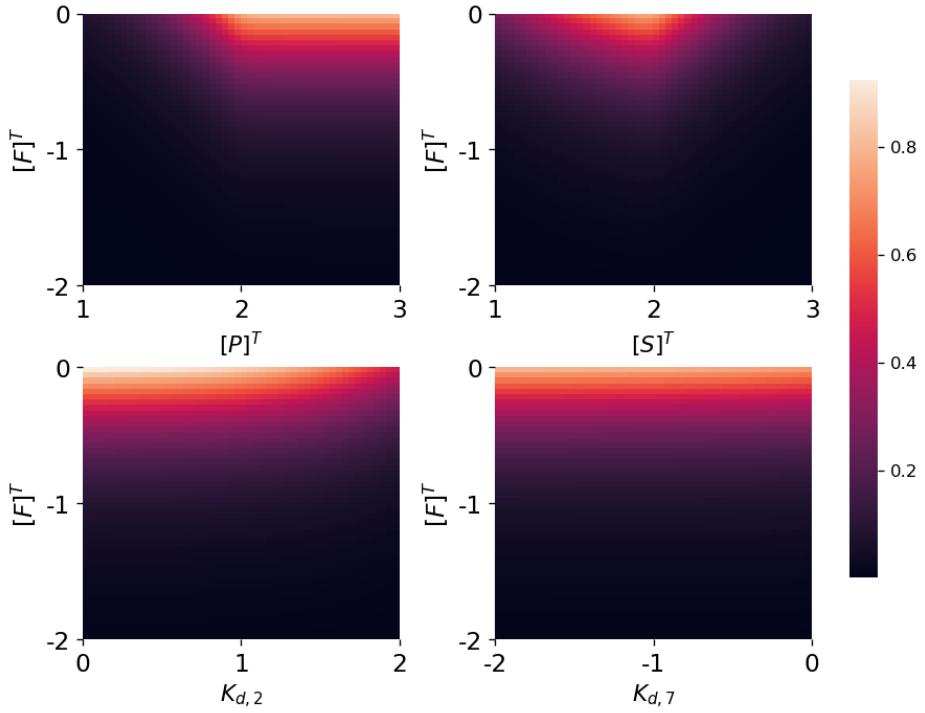

Figure 5.7: Amplitude of the steady state concentration of $pF \cdot S \cdot pP$ with units μM as indicated by the colour bar for different pairings of the parameters present in the steady state, in particular those pairings involving $[F]^T$. The x and y axes show the logarithm base 10 of the μM value of the parameter labelled.

complex rapidly increases. From the top left subplot, it can be observed that the concentration of the ternary complex is greatest when both $[F]^T$ and $[P]^T$ take their highest values, however the same is not true when $[F]^T$ is varied with

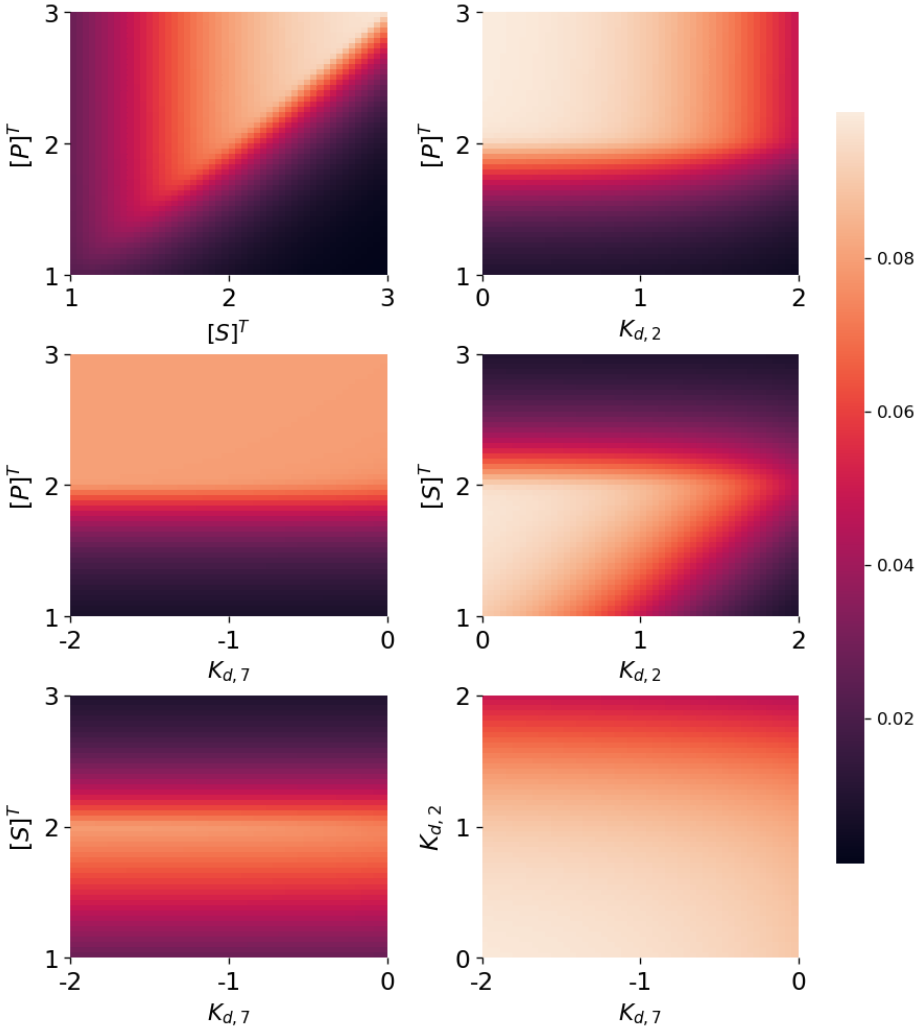

Figure 5.8: Amplitude of the steady state concentration of $pF \cdot S \cdot pP$ with units μM as indicated by the colour bar for different pairings of the parameters present in the steady state, in particular those pairings not involving $[F]^T$. The x and y axes show the logarithm base 10 of the μM value of the parameter labelled.

$[S]^T$, where there is an optimal concentration of Shp2_C of approximately 10^2 μM which yields the highest concentration of $pF \cdot S \cdot pP$. Ternary complex production appears to be greatest when $[F]^T$ and $[P]^T$ take the maximal values in their ranges and $[S]^T$ takes values in the middle of its range. The bottom row of the same figure implies that $K_{d,2}$ should take lower values in order for more ternary complex to form, which is unsurprising since this condition means a higher affin-

ity between the receptor and Shp2_C. The constant $K_{d,7}$, however, seems to have very little effect on $[pF \cdot S \cdot pP]^*$, which is likely due to the much lower μM range for this constant, seen in Table 5.5.

The lack of variation in the concentration of the ternary complex caused by $K_{d,7}$ remains in place when this constant is varied with the other parameters and initial concentrations, seen in Figure 5.8. In each subplot with $K_{d,7}$ on the x-axis, there is little noticeable change in the concentration of the ternary complex as one moves along the x-axis. $[S]^T$ again optimally takes the value 10^2 μM when varied with $K_{d,7}$ in terms of generating the maximal concentration of $pF \cdot S \cdot pP$, and $K_{d,2}$ again takes its lowest values for the same condition. Similarly to when $[P]^T$ is varied with $[F]^T$, when it is varied with $K_{d,7}$, the maximal concentrations of the ternary complex appear when $[P]^T \geq 10^2$ μM, as can be seen in the left hand subplot of the middle row of Figure 5.8. When $[S]^T$ and $[P]^T$ are varied together and $[F]^T$ is kept constant, the concentration of ternary complex in steady state is greatest when $[S]^T$ and $[P]^T$ are both maximal. This is seen in the top left subplot of the figure, and it can also be noted that there is a large region of parameter space where $[S]^T$ is large but $[P]^T$ is small, where the ternary complex concentration is minimal. However when the opposite is true, that $[P]^T$ is large but $[S]^T$ is small, there are many combinations of these two concentrations which still yield a relatively large concentration of $pF \cdot S \cdot pP$. The remaining subplots in Figure 5.8 show $K_{d,2}$ on the x-axis and one of the initial concentrations, $[S]^T$ or $[P]^T$ on the y-axis. The dynamics for $K_{d,2}$ versus $[P]^T$ are as would be expected, where $[pF \cdot S \cdot pP]^*$ is greatest when $K_{d,2}$ is lowest and $[P]^T$ is highest. However, interestingly, the dynamics are different when $K_{d,2}$ is varied with $[S]^T$, where $K_{d,2}$ should take its lowest values and $[S]^T$ should also take its lowest values in order to generate the highest concentration of ternary complex.

The colour bar associated with Figure 5.7 has a much larger scale, reaching concentrations of at least 0.8 μM of the ternary complex whereas the concentration in Figure 5.8 only reaches values of around 0.08 μM. In Figure 5.8, $[F]^T$ is fixed at 10^{-1} μM and hence it can be concluded that, as expected, the steady state concentration of the ternary complex is greatest when $[F]^T$ is largest. Another way in which the importance of the parameters to the output $[pF \cdot S \cdot pP]^*$ (along with the steady state concentration of the other variables) can be quantified, is

by conducting a Sobol sensitivity analysis, which is presented in the following section.

5.3.2 Global sensitivity analysis

It is of interest to further explore how the steady state concentrations of the variables present in steady state (5.20) depend on the parameters present in the same steady state. These parameters are $[F]^T$, $[S]^T$, $[P]^T$, $K_{d,2}$ and $K_{d,7}$, as in the previous section. The ranges considered for each parameter when constructing the Saltelli sample (Zhang *et al.*, 2015b) to be used in the Sobol sensitivity analysis are given in Table 5.5 and are the same ranges as those used to plot Figures 5.7 and 5.8.

Having constructed a sample of the parameter values in Table 5.5, the steady state expressions (5.20) and (5.21) - (5.23) were then solved numerically using *Paramotopy* for each set of parameter values. The Sobol sensitivity analysis (see Section 2.4) was then carried out using the SALib package in Python in order to find the total-order Sobol indices for each of the parameters listed in Table 5.5 with respect to each of the steady state concentrations for the variables pF, S, $pF \cdot S$, pP, $S \cdot pP$ and $pF \cdot S \cdot pP$. The results of this analysis are plotted in Figure 5.9, which shows the total-order Sobol index for each parameter (different coloured bars) with respect to each steady state variable (groupings on the x-axis).

The grouping on the far right of the x-axis of Figure 5.9 shows the total-order Sobol indices for each of the five parameters with respect to the steady state concentration of the ternary complex. In line with the results in the previous section, the bar for the total-order Sobol index for the parameter $K_{d,7}$ is not visible for the ternary complex output, or the outputs for any of the other variables, indicating that this parameter is unimportant in the amplitude of the steady state concentrations. $K_{d,2}$ is also relatively unimportant, where the Sobol index for this parameter comprises only approximately 5% of the total variation in $[pF \cdot S \cdot pP]^*$. The three initial concentrations are all reasonably important in determining the concentration of the ternary complex in steady state, where, $[F]^T$ is the most important, corroborating the result of the previous section. The

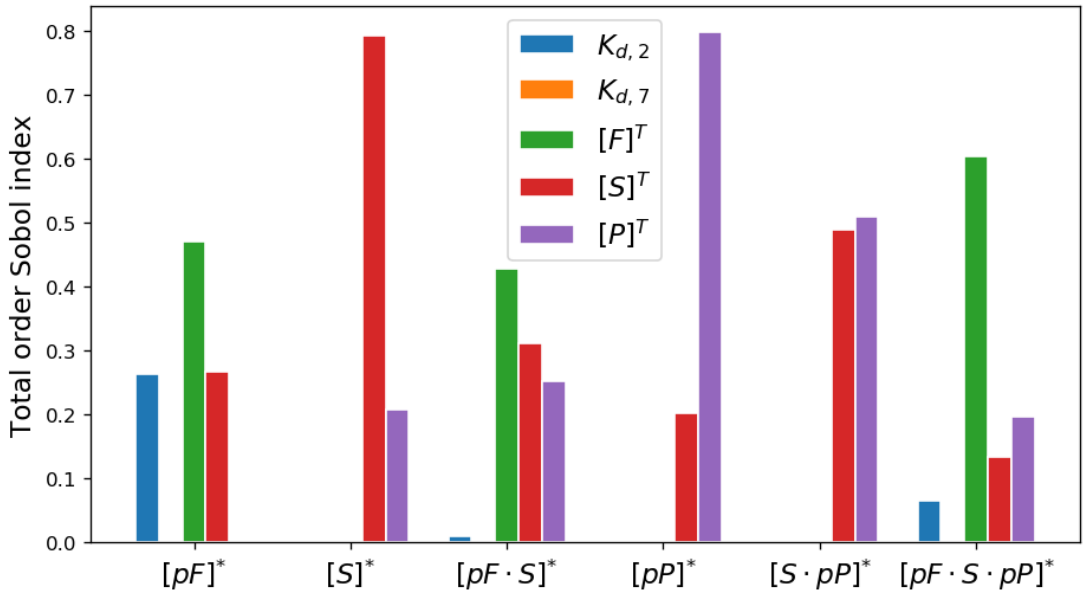

Figure 5.9: Bar charts of the total-order Sobol indices for each parameter (represented by different coloured bars as given in the legend) with respect to each steady state variable (groupings on the x-axis) in the steady state involving the ternary complex $pF \cdot S \cdot pP$ from the FGFR2 mathematical model.

initial concentration of Plcγ1 appears to be slightly more important than the initial concentration of Shp2_C with regards to the steady state concentration of the ternary complex, given the ranges considered here for the parameter values.

The importance of the parameters with relation to the steady state concentrations of the other variables are all intuitive, where for example $[P]^T$ and $[S]^T$ are the most important parameters for the outputs $[pP]^*$ and $[S]^*$ in steady state. These parameters are almost equally as important as one another when considering the output $[S \cdot pP]^*$, as expected. $[F]^T$ is an important parameter for both of the steady state outputs in which FGFR2 is present, $[pF]^*$ and $[pF \cdot S]^*$. Interestingly $[P]^T$ is also important for the output $[pF \cdot S]^*$, presumably because both pFGFR2 and Shp2_C can also bind with Plcγ1.

Clearly for certain combinations of parameter values, in particular when the initial concentrations of the three species are high, a reasonably high concentration of ternary complex can be produced and maintained in the late time dynamics of the system. It is then of interest to determine the stability of the steady state

involving the ternary complex, and hence a stability analysis is carried out in Section 5.3.3.

5.3.3 Stability analysis

In this section, the stability (see Section 2.3.2) of steady state (5.20) is explored for different combinations of parameter values in the model. Due to the complexity of the solutions found by solving Equations (5.20) together with Equations (5.21) - (5.23), analytic stability tests, such as computing symbolically the eigenvalues of the Jacobian matrix, or using the Routh-Hurwitz criteria (Niu & Wang, 2008), are not employed here. Instead, the stability was assessed numerically, for varying rate constants and initial concentrations, by finding numeric values for the steady state concentration of each species, using the method introduced in Section 5.3.1. These numeric values were then substituted into the Jacobian matrix for the system and the eigenvalues computed. In order to simplify the analysis, it was firstly noted that the original system of 14 variables could be reduced to a system of 11 variables by considering the three conservation equations, Equations (5.15) - (5.17). In particular, these equations could be rearranged so that $[F]$, $[S]$ and $[P]$ could be written in terms of the fixed initial concentrations $[F]^T$, $[S]^T$ and $[P]^T$, and the other variables in the system as,

$$\begin{aligned}[F] = [F]^T &- [pF] - [pF \cdot S] - [pF \cdot P] - [pF \cdot pP] - [pF \cdot S \cdot P] \\ &- [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S],\end{aligned} \tag{5.26}$$

$$\begin{aligned}[S] = [S]^T &- [pF \cdot S] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] \\ &- [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S], \text{ and},\end{aligned} \tag{5.27}$$

$$\begin{aligned}[P] = [P]^T &- [pF \cdot P] - [pF \cdot pP] - [pP] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] \\ &- [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S].\end{aligned} \tag{5.28}$$

The variables $[F]$, $[S]$ and $[P]$ could then be implicitly tracked by solving the ODEs for the remaining variables only, where $[F]$, $[S]$ and $[P]$ in the remaining equations were substituted by the right hand side of Equations (5.26) - (5.28). This resulted in the system of equations given by (5.29) - (5.39).

$$\begin{aligned}\frac{d[pF]}{dt} &= k_1([F]^T - [pF] - [pF \cdot S] - [pF \cdot P] - [pF \cdot pP] - [pF \cdot S \cdot P] \\ &- [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) + k_{-2}([pF \cdot S] \\ &+ [pF \cdot S \cdot P] + [pF \cdot S \cdot pP]) - k_{+2}[pF]([S]^T - [pF \cdot S] - [pF \cdot S \cdot P] \\ &- [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) + k_{-3}([pF \cdot P] \\ &+ [pF \cdot P \cdot S]) - k_{+3}[pF]([P]^T - [pF \cdot P] - [pF \cdot pP] - [pP] - [S \cdot pP] \\ &- [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) \\ &+ k_5([pF \cdot pP] + [pF \cdot pP \cdot S]) \end{aligned} \tag{5.29}$$

$$\begin{aligned}\frac{d[pF \cdot S]}{dt} &= k_{+2}[pF]([S]^T - [pF \cdot S] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] \\ &- [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) - k_{-2}[pF \cdot S] - k_{+7}[pF \cdot S][pP] \\ &+ k_{-7}[pF \cdot S \cdot pP] \end{aligned} \tag{5.30}$$

$$\begin{aligned}\frac{d[pF \cdot P]}{dt} &= k_{+3}[pF]([P]^T - [pF \cdot P] - [pF \cdot pP] - [pP] - [S \cdot P] - [S \cdot pP] \\ &- [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) \\ &- k_{-3}[pF \cdot P] - k_4[pF \cdot P] \end{aligned} \tag{5.31}$$

$$\frac{d[pF \cdot pP]}{dt} = k_4[pF \cdot P] - k_5[pF \cdot pP] \tag{5.32}$$

$$\begin{aligned}\frac{d[pP]}{dt} &= k_5[pF \cdot pP] - k_{+7}[pP]([S]^T - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] \\ &- [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) + k_{-7}([S \cdot pP] \\ &+ [pF \cdot S \cdot pP]) \end{aligned} \tag{5.33}$$

$$\begin{aligned}\frac{d[S \cdot P]}{dt} &= k_{+6}([S]^T - [pF \cdot S] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] \\ &- [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])([P]^T - [pF \cdot P] - [pF \cdot pP] - [pP] \\ &- [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] \\ &- [pF \cdot pP \cdot S]) - k_{-6}[S \cdot P] - k_{+2}[pF][S \cdot P] + k_{-2}[pF \cdot S \cdot P] \\ &- k_{+3}[pF][S \cdot P] + k_{-3}[pF \cdot P \cdot S] \end{aligned} \tag{5.34}$$

$$\frac{d[S \cdot pP]}{dt} = k_{+7}[pP]([S]^T - [pF \cdot S] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S]$$

$$- [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) - k_{-7}[S \cdot pP] - k_{+2}[pF][S \cdot pP]$$
$$+ k_{-2}[pF \cdot S \cdot pP] + k_5[pF \cdot pP \cdot S] \tag{5.35}$$

$$\frac{d[pF \cdot S \cdot P]}{dt} = k_{+2}[pF][S \cdot P] - k_{-2}[pF \cdot S \cdot P] \tag{5.36}$$

$$\frac{d[pF \cdot P \cdot S]}{dt} = k_{+3}[pF][S \cdot P] - k_{-3}[pF \cdot P \cdot S] - k_4[pF \cdot P \cdot S] \tag{5.37}$$

$$\frac{d[pF \cdot S \cdot pP]}{dt} = k_{+7}[pF \cdot S][pP] - k_{-7}[pF \cdot S \cdot pP] + k_{+2}[pF][S \cdot pP]$$
$$- k_{-2}[pF \cdot S \cdot pP] \tag{5.38}$$

$$\frac{d[pF \cdot pP \cdot S]}{dt} = -k_5[pF \cdot pP \cdot S] + k_4[pF \cdot P \cdot S] \tag{5.39}$$

Denoting by f_1 to f_{11} the right hand sides of Equations (5.29) - (5.39), respectively, the Jacobian matrix for the system of ODEs is given by

$$\mathbf{J} = \begin{pmatrix} \frac{\partial f_1}{\partial [pF]} & \frac{\partial f_1}{\partial [pF \cdot S]} & \cdots & \frac{\partial f_1}{\partial [pF \cdot S \cdot pP]} & \frac{\partial f_1}{\partial [pF \cdot pP \cdot S]} \\ \frac{\partial f_2}{\partial [pF]} & \frac{\partial f_2}{\partial [pF \cdot S]} & \cdots & \frac{\partial f_2}{\partial [pF \cdot S \cdot pP]} & \frac{\partial f_2}{\partial [pF \cdot pP \cdot S]} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial f_{11}}{\partial [pF]} & \frac{\partial f_{11}}{\partial [pF \cdot S]} & \cdots & \frac{\partial f_{11}}{\partial [pF \cdot S \cdot pP]} & \frac{\partial f_{11}}{\partial [pF \cdot pP \cdot S]} \end{pmatrix}.$$

The partial derivatives comprising the Jacobian matrix are listed in Appendix C. The stability of a numerical steady state of the system could then be assessed by computing the eigenvalues of $\mathbf{J}$, evaluated at the steady state. In the case where the real parts of each of the 11 eigenvalues are negative, the steady state is stable, and otherwise, the steady state is unstable.

Firstly, the stability of the steady state given in Table 5.4, set 2 (see also Figure 5.5), obtained using the experimental initial conditions and K_d values (Tables 5.3 and 5.1) was assessed, where the parameters not used to generate the steady state were sampled from the feasible distributions given in Table 5.6. In particular, for 10^4 sampled parameter sets, where the parameters were sampled

as specified in Table 5.6, the eigenvalues of the Jacobian matrix were computed, using the steady state values for the variables as in Set 2 of Table 5.4. Of the 10^4 parameter sets, 68% yielded stable steady states, whereby all of the real parts of the eigenvalues of the Jacobian matrix were negative, and the remaining 32% yielded unstable steady states, with at least one positive real part of an eigenvalue. This is a positive result as it implies that for a large range of biologically feasible parameter values, the steady state involving the ternary complex $pF \cdot S \cdot pP$ is stable and hence is likely to be observed experimentally. Figure 5.10 shows histograms of the sampled parameter values for each parameter, coloured by the stability of the steady state found when using such parameter values. From Figure 5.10 it can be seen that several of the parameters appear to have very little effect on the stability of the steady state, namely $K_{d,6}$, k_1, k_{-2}, k_{-6} and k_5. The affinity between pF and P should in general be low in order for the steady state to be stable, as indicated by the parameters $K_{d,3}$ and k_{-3}. The steady state is often unstable for very low values of k_4, further confirming the notion that the phosphorylation of Plcγ1 whilst bound to phosphorylated FGFR2 should be a very fast reaction. Finally, in general k_{-7} should take low values in order for the steady state to be stable.

Parameter	Range	Unit
$K_{d,3}$	10^r, where $r \sim Unif(-2, 0)$	μM
$K_{d,6}$	10^r, where $r \sim Unif(-1, 1)$	μM
k_1, k_4	10^r, where $r \sim Unif(-3, 2)$	μM
k_{-2}, k_{-3}, k_{-6}, k_{-7}, k_5	10^r, where $r \sim Unif(-2, 1)$	μM

Table 5.6: Distributions used to sample each of the parameters *not* used to obtain the steady state of the FGFR2 mathematical model in which the ternary complex $pF \cdot S \cdot pP$ is present. These distributions were used to numerically assess the stability of the steady state.

In Figure 5.5, the mathematical model was solved under the conditions $k_{-2} = k_{-3} = k_{-6} = k_{-7} = k_5 = 10^{-1}$ s^{-1} and $k_1 = k_4 = 10^0$ s^{-1} (unlike in this section so far whereby each parameter has been sampled independently of the other parameters, as stated in Table 5.6), with the association rate constants fixed using

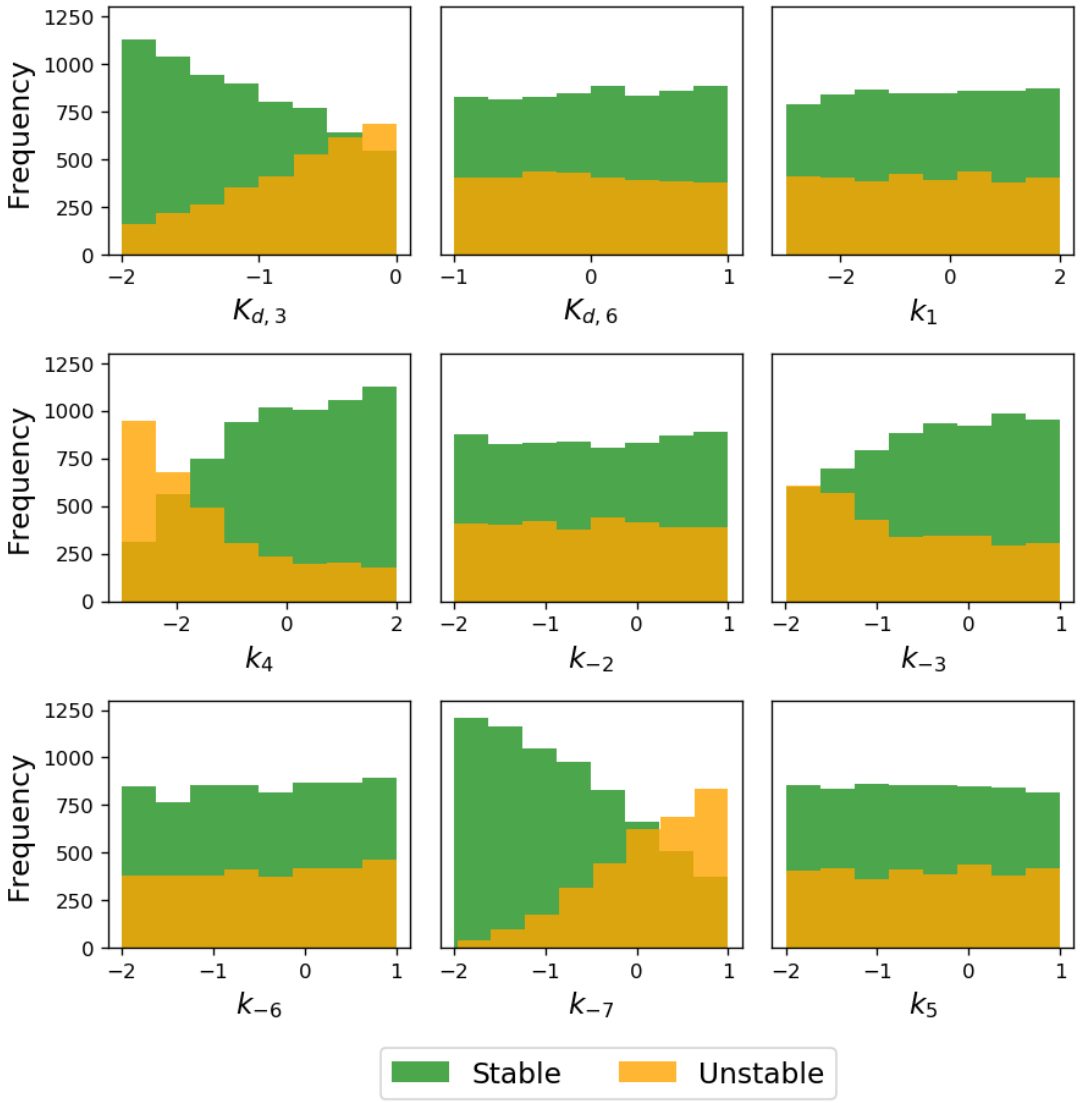

Figure 5.10: Histograms of the sampled parameter values for each parameter not used to generate the experimental steady state solution, coloured by the stability of the steady state when using such parameter values.

the K_d values in Table 5.1. Although this scenario may not be biologically realistic (*i.e.* rate constants for different reactions being equal), it is still an interesting case to study as it allows for the mathematical model given by Equations (5.29) - (5.39) to be written in terms of only one parameter. Specifically, setting $k_{-2} = k_{-3} = k_{-6} = k_{-7} = k_5 = k_m$ and $k_1 = k_4 = k_p$ and writing $k_{+i} = \frac{k_m}{K_{d,i}}$ for

$i \in \{2, 3, 6, 7\}$ everywhere in the model equations would recast the model in terms of the two unknown parameters k_m and k_p with units s^{-1}, since the K_d values for each reaction, and the initial concentrations for each species, can be fixed at the experimentally derived/used values. Then, dividing through by k_m results in a system of ODEs with only one unknown parameter, namely $\frac{k_p}{k_m}$, where time has been rescaled so that the derivatives are with respect to $\tau = k_m t$. This system of ODEs is given in Appendix D.

In this specific case, one can assess the stability of the biologically feasible and relevant steady state in terms of the unitless parameter $\frac{k_p}{k_m}$ only. This assessment was carried out using the same numeric eigenvalue method and for 10^4 sampled values of $\frac{k_p}{k_m}$, where $\frac{k_p}{k_m} = 10^r$, with $r \sim Unif(-4, 4)$. The 10^4 sampled parameter values are plotted as a histogram in Figure 5.11, coloured by the stability of the steady state found when using such parameter values. As in Figure 5.10, the colours representing stable and unstable parameter values in Figure 5.11 are semi-transparent so that one could see if there was any overlap between the histograms of each colour. Interestingly here, there is no such overlap, and hence there is a value of the parameter $\frac{k_p}{k_m}$ for which the steady state switches from unstable to stable, and this value is approximately $\frac{k_p}{k_m} = 10^{-1}$ (the maximum sampled value resulting in an unstable steady state is 0.0970 and the minimum sampled value resulting in a stable steady state is 0.0978). In the theory of dynamical systems, this switch in stability of a steady state upon varying a parameter value is known as a *bifurcation* and the parameter value at which the switch occurs is known as the *bifurcation point* (Trefethen *et al.*, 2017). This bifurcation point implies that the steady state seen in Figure 5.5 will be stable for all parameter values where $\frac{k_p}{k_m} > 10^{-1} \implies k_p > 10^{-1} k_m$ (approximately), *i.e.* k_p should be greater than a tenth of k_m for stability.

Returning to the more biologically likely scenario in which the unbinding rates are not forced to be equal to each other, and neither the phosphorylation rates, more numerical values of the steady state given by Equations (5.20) together with Equations (5.21) - (5.23) are now considered. Although it is encouraging that the steady state seen in Figure 5.5 is stable for a wide range of parameter values, given that there may be some experimental error in determining the K_d values and initial concentrations of molecules, it is also of interest to test the stability

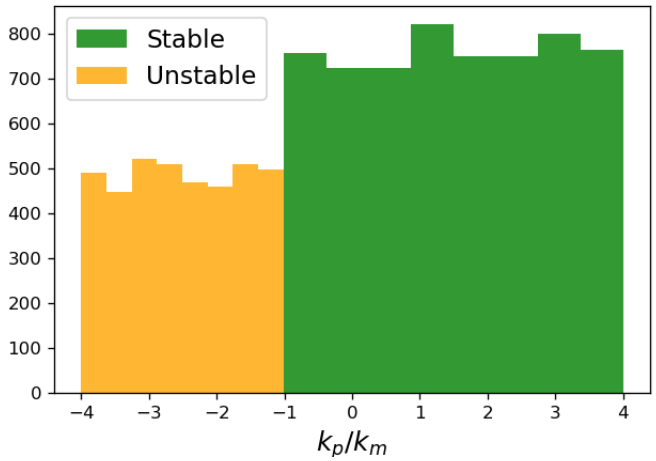

Figure 5.11: Histogram of the sampled values of $\frac{k_p}{k_m}$, coloured by the stability of the steady state when using such parameter values.

of further steady states, using slight variations of the parameters present in the steady state equations. To this end, 10^3 numeric steady states were computed using *Paramotopy*, for sampled parameters $[F]^T$, $[P]^T$, $[S]^T$, $K_{d,2}$ and $K_{d,7}$, where the parameters were sampled uniformly across the 5-dimensional space, as specified in Table 5.5. For each of these steady states, the stability was assessed using the numerical eigenvalue method, where again 10^4 sets of the parameters given in Table 5.6 were used per steady state and were sampled from the distributions given in the same table. For each of the 10^3 steady states, the percentage of stable steady states was computed (out of the 10^4 sampled parameter sets) and a histogram of these percentages is given in Figure 5.12. It can be seen that, for all of the 10^3 steady states computed, using values in the region of the experimentally determined/used values, the steady state is always stable for at least some of the parameter sets tested. In fact, the minimum value of the percentage of stable steady states is 7.29% however most steady states are stable for a much greater percentage of sampled parameters. The median value of the percentages is 66.3% for example, and a quarter of the percentages take value larger than 89% implying that the steady state is more often stable than unstable for the parameter ranges considered here.

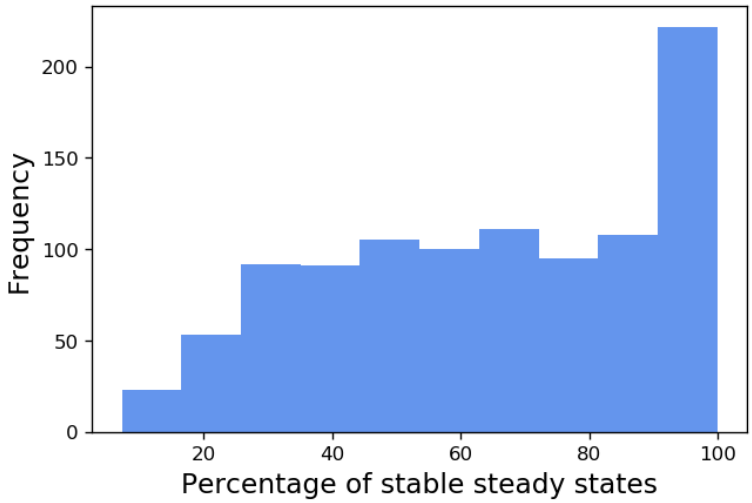

Figure 5.12: Histogram of the percentage of stable steady states for the 10^3 numerical steady states where the stability for each steady state was assessed using 10^4 sampled parameter sets from the distributions given in Table 5.6.

Having computed the stability of 10^3 steady states in the region of the experimental steady state, the dependence on the parameters in the model for stability can be further clarified. To this end, similar histograms to those in Figure 5.10 are plotted in Figure 5.13, where the histograms are now the parameter values sampled for each of the 10^3 steady states combined, again coloured by stability as in the figure legend. This figure confirms the fact that k_1, k_{-2} and k_5 do not influence the stability of the steady state, however $K_{d,6}$ and k_{-6} are seen to have a moderate effect on stability. In general, the steady state is found to be stable most often when $K_{d,6}$ takes larger values, indicating that the affinity between S and P is low.

5.4 Discussion

In this chapter, a deterministic mathematical model of the theoretical formation of a ternary complex between the proteins pFGFR2, Shp2_C and pPlcγ1 has been developed. Each of these proteins contribute to cell signalling through different pathways such as the MAPK, JAK-STAT, PI3K-AKT and Plcγ pathways, where FGFR2 can be the receptor at the head of such pathways. Shp2 is well known

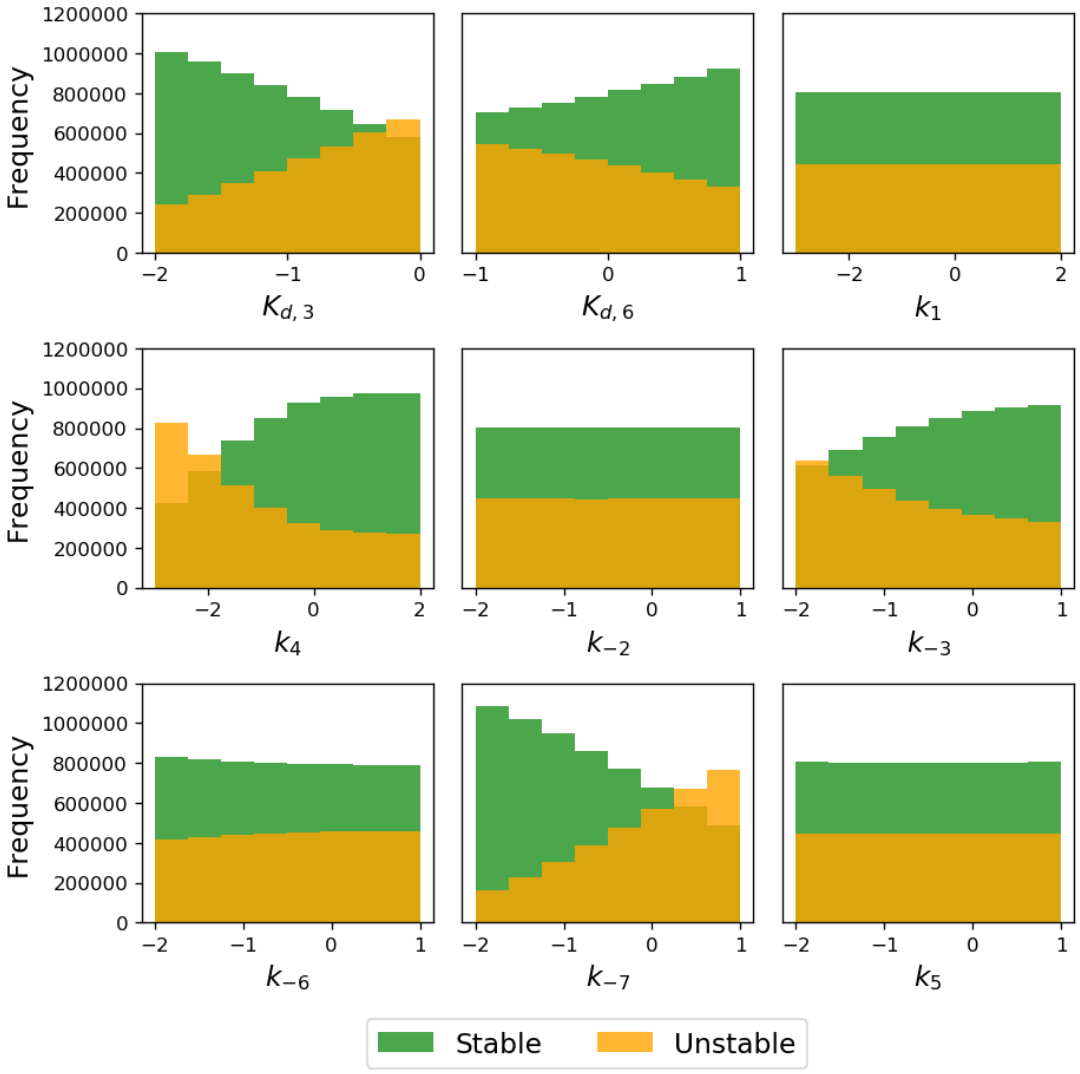

Figure 5.13: Histograms of the combined sampled parameter values for each parameter not used to generate the steady state solution, for each of the 10^3 numeric steady states, coloured by the stability of the steady state when using such parameter values.

as a signal enhancing protein in signalling cascades, however, to date, it has not been known to engage with Plcγ1 and play a role in the Plcγ pathway. Dr. Chi-Chuan Lin and other members of the group of molecular and cellular biology at the University of Leeds however, have observed, through experimental work, that pFGFR2, $Shp2_C$ and pPlcγ1 co-locate within LLPS droplets. It was postulated

therefore, that the three proteins form a ternary complex, and the aim of this chapter was to confirm, through a mechanistic mathematical analysis, that this ternary complex formation is indeed possible. Such ternary complex and LLPS droplet formation between the three proteins is important biologically, as it revels a way in which signalling is maintained in a crowded cellular environment.

Given that the LLPS droplets were observed in an *in vitro* experimental setup, the mathematical model was developed based on the underlying reactions thought to occur in the experiments. The ODE model therefore was built using mass action kinetics and considering binding and unbinding of the three proteins with one another, as well as phosphorylation reactions for the receptor and Plcγ1. Given that both Shp2$_{\mathrm{C}}$ and Plcγ1 can bind with pFGFR2 and that Plcγ1 can exist in a phosphorylated or unphosphorylated state, there were four ternary complexes which could form via the reactions underlying the mathematical model. Through analysis of the steady states of the model however, it was found that there was only one biologically relevant steady state in which all three of pFGFR2, Shp2$_{\mathrm{C}}$ and Plcγ1 were present and in this steady state, only one of the four ternary complexes prevailed, namely $pF{\cdot}S{\cdot}pP$. Encouragingly, this is the ternary complex which the experimentalists had hypothesised would form in the LLPS droplets. The steady state was found implicitly using *Mathematica*, however considering also the conservation equations for the system in steady state, one could write six equations in six variables to be solved for the steady state.

The steady state concentration of the ternary complex, as well as the other complexes present in the steady state, was then explored for varying parameter values and initial conditions, where these values were varied around those found/used experimentally. A method known as parameter homotopy continuation, carried out using *Paramotopy*, was used in order to solve efficiently and numerically, the system of implicit steady state equations. It was found through solving of the steady state equations as well as through Sobol sensitivity analysis, that the initial concentrations of FGFR2, Shp2$_{\mathrm{C}}$ and Plcγ1 were particularly important to the concentration of the ternary complex at steady state. The parameter $K_{d,2}$ (indicating the strength of binding between pF and S) had little importance to the concentration of any species in steady state, and the param-

eter $K_{d,7}$ (indicating the strength of binding between S and pP) had almost no importance to the same outputs, for the ranges considered here.

Finally, although it is encouraging that a steady state was found in which the theorised ternary complex existed, it is only meaningful if the steady state were to be stable for at least some biologically relevant parameter values. Hence, in Section 5.3.3 (see also Appendices C and D) a numerical stability analysis of the steady state is carried out. The stability of the experimentally defined steady state, as well as other numerical steady states in the region of the experimental, was analysed for varying parameter values, for parameters present in the Jacobian matrix for the system. Specifically, the steady states were found numerically using the parameter homotopy continuation method in *Paramotopy*, and then the eigenvalues of the Jacobian matrix were computed numerically in *Python*. A steady state is considered stable if the real parts of all of the eigenvalues are negative, and it was found through the stability analysis that the biologically relevant steady state of the system is indeed stable for a large range of parameter values.

The work carried out in this chapter gives confirmation of a biological hypothesis which would be difficult to prove experimentally. Currently, the mathematical model extends only up to the point of ternary complex formation, and hence a continuation of this work could be to include also the aggregation of such ternary complexes to simulate the formation of the LLPS droplets. Given appropriate quantitative data, the model could even be parameterised, using for example the ABC-SMC method used in Chapter 4, and could then be used to predict droplet formation in different concentration regimes.

Chapter 6

Statistical analysis of EGFR inhibition

The epidermal growth factor receptor is a transmembrane receptor tyrosine kinase and is a member of the ErbB family of receptors. The ErbB family has four protein members in human cells, where EGFR, also known as HER-1, was the first to be discovered (Carpenter & Cohen, 1976). As the name suggests, growth factor receptors play a role in the growth, survival, differentiation and division of cells (Herbst, 2004). The EGFR is comprised of an extracellular ligand-binding domain, a transmembrane domain, and an intracellular tyrosine kinase domain. Ligands, small proteins which diffuse in the extracellular medium, can bind to the extracellular domain of EGFR and this process initiates the classical EGFR signalling mechanism. There are at least seven different ligands known to bind with EGFR, including EGF, which is specific to EGFR, whereas some of the other known ligands can also bind other members of the ErbB family (Harris *et al.*, 2003). Upon ligand binding, EGFR undergoes a conformational change in the extracellular domain, which allows two ligand-bound receptors to form a dimer when they diffuse through the cell membrane into close proximity with one another. This ligand-induced dimerisation is followed by trans-autophosphorylation of particular residues on the intracellular domains of each receptor in the dimer (Lemmon *et al.*, 2014; Schlessinger, 2002). Ligand binding and dimerisation are the first steps of classical EGFR signalling which induce a cascade of intracellular

phosphorylation events, ultimately promoting cell growth, survival, differentiation or division.

One such intracellular signalling cascade is the MAPK pathway (see Chapter 1), which can be induced by EGFR-EGF binding and subsequent receptor dimerisation. There are many proteins involved in this pathway, as well as different branches initiating different cellular outcomes and complex negative feedback mechanisms (McKay & Morrison, 2007). Lake *et al.* (2016) however, give a brief description of the key events in the MAPK pathway which is summarised here. The phosphorylated residues on the intracellular domains of EGF receptors in a ligand-induced dimer act as docking sites for other intracellular proteins, in particular, those with a certain domain known as the Src homology 2 (SH2) domain, one example of which is Grb2. Grb2 can then recruit a protein known as SOS, which in turn recruits Ras-GTP proteins and catalyses their activation. Upon activation of Ras-GTP, Raf proteins are also recruited and activated. In the same fashion, Raf proteins are capable of activating MEK which can then activate ERK, where all of the activating events are achieved by phosphorylation of protein kinase domains. The end result of these signalling events is dependent on the intracellular conditions and the availability of ERK targets. ERK has multiple phosphorylation targets some of which are transcription factors which promote the production of other proteins and some are other cytoplasmic proteins such as RSK. A depiction of the MAPK pathway is shown in Figure 6.1. The proteins produced by gene transcription as a result of the activation of the MAPK pathway are important in regulation of the cell cycle.

Given its importance in the cell cycle, dysregulation of the MAPK pathway can have implications in human disease. In particular, and of interest here, is the role of EGFR and the MAPK signalling pathway in different types of cancer. Since one of the predominant roles of EGFR is to promote the division of cells, it is easy to see how aberrant amplification of EGFR could lead to tumour growth. In fact, many authors (Nicholson *et al.*, 2001; Normanno *et al.*, 2006) have reported a strong correlation between the amplification level of EGFR and prognosis for patients with with many types of cancer including head and neck, ovarian, cervical, bladder and oesophageal. Increased expression of EGFR was the initial discovery which caused EGFR to become a major clinical target

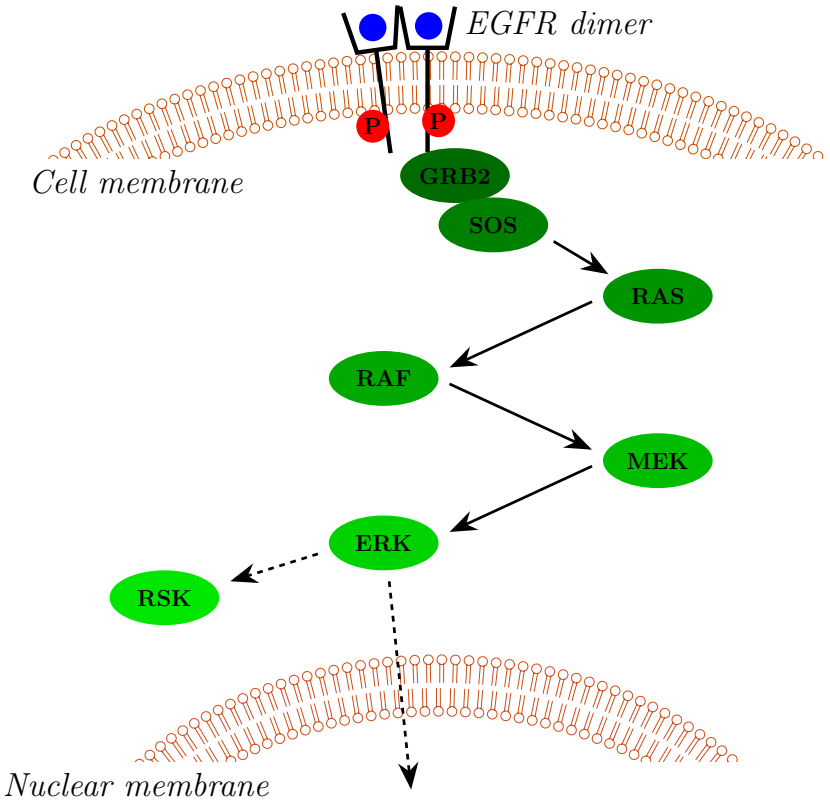

Figure 6.1: A diagram of the receptor-ligand initiated RAS-RAF-MEK-ERK signalling pathway, known as the traditional, or classical MAPK pathway, based on Pratilas & Solit (2010) and Lake *et al.* (2016). The arrows indicate the direction of the phosphorylation cascade, where the protein ERK can either phosphorylate other cytoplasmic proteins or can move to the nucleus to initiate protein synthesis.

in oncology. More recently however, EGFR has been discovered in different mutated forms, whereby even without the amplification of the protein, the mutations can cause EGFR to become constitutively active (da Cunha Santos *et al.*, 2011; Roberts & Der, 2007). This means that ligand binding is no longer required for activation of the MAPK pathway and hence ligand concentration in the extracellular medium, a factor which can control the level of MAPK signalling, is no longer relevant. The MAPK pathway therefore is constantly stimulated in cells

with mutant copies of EGFR, hence resulting again in increased cell division and potentially tumour growth.

The focus here will be on EGFR mutations in a specific type of lung cancer known as non-small cell lung cancer (NSCLC) which comprises approximately 80% of lung cancers, where lung cancer is the leading cause of cancer death worldwide (Bade & Cruz, 2020; Sharma *et al.*, 2007). The mutations come in different forms and are commonly found in exons (sections) of the tyrosine kinase domain of EGFR. One of the most common EGFR mutations found in NSCLC, is the L858R amino acid exchange mutation, in which a Leucine amino acid is replaced by Arginine at codon 858 in exon 21 of EGFR. This mutation can be treated with tyrosine kinase inhibitors (TKIs) such as Gefitinib and Erlotinib, known as first-generation EGFR inhibitors, which bind reversibly to the ATP binding pocket of EGFR, thus blocking the binding of ATP which is required for receptor autophosphorylation and further downstream signalling (Cataldo *et al.*, 2011). Often however, after an initially promising course of treatment with Gefitinib or Erlotinib, acquired drug resistance occurs, which in approximately 50% of cases is due to the development of a secondary EGFR mutation known as T790M (Günther *et al.*, 2016). Another type of EGFR mutation is the exon 20 insertion mutation, in which 1 to 7 amino acids are inserted into the exon 20 of the tyrosine kinase domain of EGFR (Vyse & Huang, 2019). The response to first-generation and second-generation (used to treat the T790M mutation) inhibitors is poor in patients with exon 20 insertion mutants and there are currently no recommended therapies specific for this type of mutation, although a review of current clinical data surrounding exon 20 insertion mutant treatment is given by Baraibar *et al.* (2020). It is therefore of huge clinical importance to understand better the mechanisms of inhibition of the exon 20 insertion mutations.

Mathematical modelling has been applied by many authors to study processes involving EGFR, ranging from ligand binding and dimerisation dynamics (Blinov *et al.*, 2006; Klein *et al.*, 2004; Kozer *et al.*, 2013b) to reactions of the MAPK pathway (Asthagiri & Lauffenburger, 2001; Tian & Song, 2012) and, more recently, interactions between EGFR and TKIs (Huang *et al.*, 2017; Zhai *et al.*, 2020). Claas *et al.* (2018a) used a compartmental mathematical model, calibrated via Bayesian inference to identify changes in receptor trafficking parameter values

upon cell treatment with different inhibitors but where here the inhibitors were targeting MEK and ERK rather than EGFR. Here, statistical analysis is employed to elucidate differences between eight potential EGFR TKIs being used in preclinical studies at AstraZeneca, using imaging data in wild type and mutant EGFR cell lines, where WT EGFR is the unmutated (natural) form of EGFR. To conclude the chapter, a review of the current literature surrounding mathematical modelling approaches with TKIs is given. The experimental data used in this chapter is explained in detail in Section 6.1.

6.1 Experimental data

A summary of the experiments is as follows. Two cell lines were used in the dose response and kinetic experiments: an isogenic H2073 cell line (WT EGFR) and an SVD H2073 cell line (exon 20 insertion mutant EGFR). For each of eight different EGFR tyrosine kinase inhibitor compounds (D1 - D8), each cell line was treated with the compound dissolved in dimethyl sulfoxide (DMSO) solvent and, separately, the DMSO solvent alone, as a control. Ten different inhibitor concentrations in the range 10^{-2} μM to 10^{1} μM were used, and five different time points (2, 4, 6, 24 and 48 hours post dose) were considered. The interest was in quantifying the abundance of four different proteins: total EGFR (tEGFR), phosphorylated EGFR (pEGFR), phosphorylated MAPK (pMAPK) (equal to the sum of phosphorylated ERK (pERK) and phosphorylated MEK (pMEK)) and phosphorylated RSK (pRSK) in each of three different cellular compartments: the plasma membrane (PM), the cytoplasm and the nucleus, in order to see the effect of the inhibitors on these proteins in the MAPK pathway. Each experimental well contained a number of cells (an equivalent number per well at the beginning of the experiments), and the abundance of a specific protein was quantified in terms of the fluorescence intensity corresponding to this protein (where different proteins were observed using different fluorescent tags which are visible in different colour channels) using fluorescence microscopy (imaging). For each cell line, inhibitor, inhibitor concentration, time point, protein and cellular compartment, the raw data is then the FI averaged over all cells in the well (*i.e.* a mean per cell FI). Two repeats were carried out for each individual experiment.

6.1.1 Data normalisation

The data is firstly normalised to the control (DMSO) data, so that background effects caused by the control are neglected in the inhibitor data. In particular, denoting the FI by fl, $fl(c, i, e, p, d, t, r)$ corresponds to the FI in cell line

$$c \in C = \{\text{WT}, \text{SVD}\},$$

stimulated with inhibitor

$$i \in I = \{\text{D1, D2, D3, D4, D5, D6, D7, D8, DMSO}\},$$

for protein

$$e \in E = \{\text{pEGFR, tEGFR, pMAPK, pRSK}\},$$

in cellular compartment

$$p \in P = \{\text{PM, Cytoplasm, Nucleus}\},$$

at inhibitor concentration (with units μM)

$$d \in D = \{9.98, 4.64, 2.16, 1.00, 0.48, 0.22, 0.10, 0.05, 0.02, 0.01\},$$

at time point (with units *hours post dose*)

$$t \in T = \{2, 4, 6, 24, 48\},$$

and repeat

$$r \in R = \{1, 2\}.$$

Each data point $data(c, i, e, p, d, t, r)$, was then computed as

$$data(c, i, e, p, d, t, r) = \frac{fl(c, i, e, p, d, t, r)}{fl(c, \text{DMSO}, e, p, d, t, r)}. \tag{6.1}$$

6.1.2 Identification of outliers

Through examination of the normalised experimental data, it was clear that there were some outliers in the data, which might skew any statistical analyses. An example of such a case is given in Figure 6.2, for the protein cytoplasmic pRSK, in the WT cell line treated with the inhibitor D4. It is clear from the scatter plot in the left hand subplot of this figure that the data point at time 2 hours, for the inhibitor concentration of 0.48 μM is an outlier. Although there is a range of concentrations, and it is expected that the higher concentrations will have a greater affect from baseline (plotted as a dashed line), one would still expect that the change from baseline would be gradual as the inhibitor concentration changes. Therefore the data point for the concentration 0.48 μM, a concentration in the middle of the range, is certainly unusual, particularly since it indicates a large increase from baseline, whereas biologically, a decrease from baseline would be expected at this time point. In order to remove such extreme data points from further analysis, the inter-quartile range (IQR) of the data at each time point was considered. A standard technique for identifying outliers is to multiply the IQR by 1.5 and add this value to the third quartile and subtract the same value from the first quartile. This gives two bounds for the data, and if a data point lies outside this range, it can be considered an outlier (Walfish, 2006). In the identification of outliers for this dataset however, a factor of 5 is chosen to multiply by the IQR, instead of the commonly used 1.5, so that fewer data points are removed, since there is an expected concentration effect. When considering the data for the proteins pEGFR, pMAPK and pRSK, a data point is only removed if it lies above the upper bound $\text{Q3} + (5 \times \text{IQR})$, and *not* if it lies below the lower bound $\text{Q1} - (5 \times \text{IQR})$. This is because it is expected biologically that the inhibitors should have the effect of reducing the amount of pEGFR, pMAPK and pRSK in the cells and hence, in order to be cautious when removing data points, data points are only removed if they indicate an increase in the amount of these proteins (*i.e.* are above the upper bound). Similarly, for tEGFR, since it is seen in the data that the inhibitors in general increase the amount of tEGFR in the cells, data points corresponding to this protein are only removed if they

are below the lower bound Q1 − (5 × IQR) and *not* if they are above the upper bound Q3 + (5 × IQR).

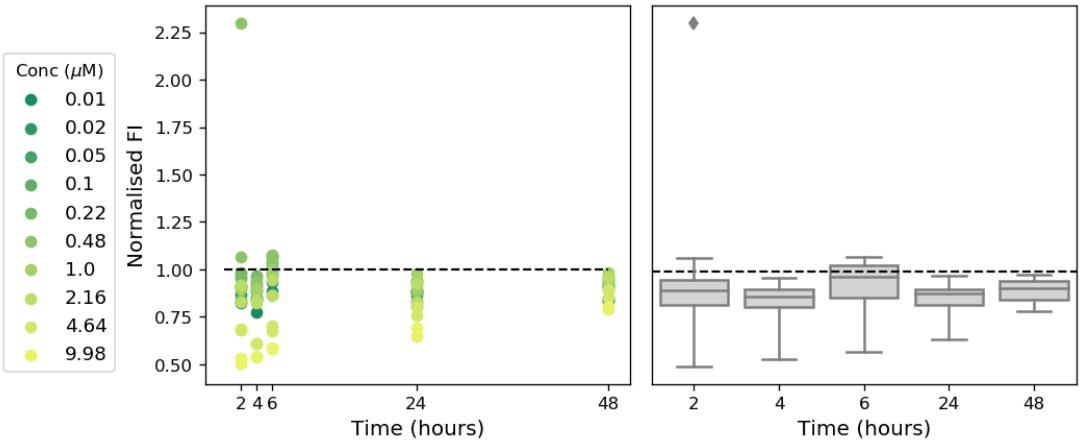

Figure 6.2: Left: Scatter plot of the normalised experimental data given by Equation (6.1), for cytoplasmic pRSK in the WT cell line treated with the inhibitor D4. The colour of the points represents the concentration of inhibitor with units μM as given in the legend and there are 20 data points at each time point corresponding to the 10 concentrations multiplied by 2 repeats of the experiment. **Right:** Box plots of the data at each time point, where individual data points are shown if they are computed as an outlier by the method explained in the text.

If a data point, $data(c, i, e, p, d, t, r)$ has been identified as an outlier, then the data points for the other proteins at each cellular compartment are also removed from further analyses, since these data points come from the same experiment which produced an outlier. For example, since the data point

$$data(\text{WT}, \text{D4}, \text{pRSK}, \text{Cytoplasm}, 0.48\ \mu\text{M}, 2 \text{ hours post dose}, 1),$$

was identified as an outlier in replicate 1, each of the data points

$$data(\text{WT}, \text{D4}, e, p, 0.48\ \mu\text{M}, 2 \text{ hours post dose}, 1),$$

for $e \in E$ and $p \in P$ were also removed for this experimental replicate. Different time points in the data correspond to different experimental wells, which explains why an outlier does not usually prevail over time. Altogether a total of 180 data points out of a total of 19140 were removed, corresponding to roughly 1% of the

data. Figure 6.3 shows histograms of the absolute difference between each pair of experimental replicates in the normalised data (given by Equation (6.1)) for the full dataset (*i.e.* for both cell lines, all inhibitor types and concentrations, all proteins at all cellular compartments and for all time points) with outliers included (left hand subplot) and with outliers removed (right hand subplot). In the right hand subplot of the figure, corresponding to the data with outliers removed, the absolute difference is only plotted for replicate pairs where *both* experimental replicates remain. It can be seen that when the outliers are removed from the data, every absolute difference is less than 1 and in fact over 95% of the differences are smaller than 0.2 and over 75% of the differences are smaller than 0.1.

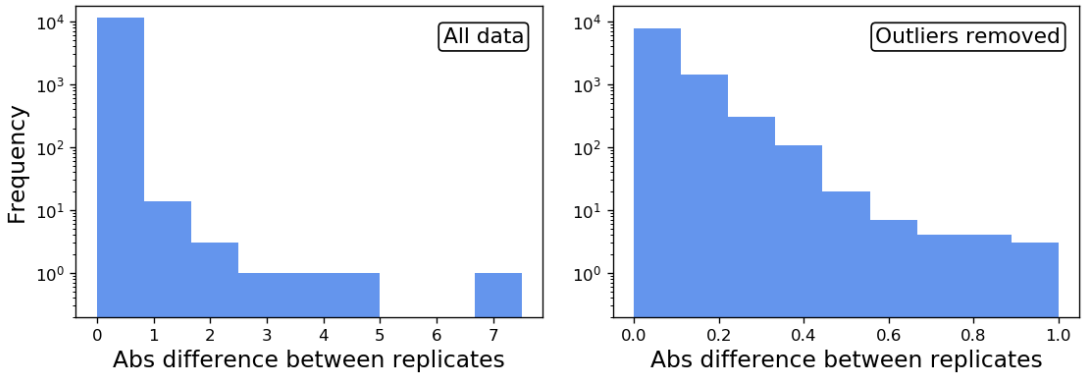

Figure 6.3: Histograms of the absolute differences between pairs of experimental replicates in the normalised data for the full dataset (**left**) and the data with outliers removed, where two experimental replicates remain (**right**).

Where there are still two repeats of the data, the mean of these replicates can be computed as,

$$\text{MFI} = \mu_{data}(c, i, e, p, d, t, r) = \frac{1}{2}\sum_{r=1}^{2} data(c, i, e, p, d, t, r), \tag{6.2}$$

and this MFI will be used in the statistical analysis in Section 6.2. Where there is only one replicate remaining (due to removal of outliers), the MFI is equal to the data point for this replicate.

6.1.3 Data visualisation

Given that the data has many dimensions, it cannot be displayed in a single plot and hence an example is given in Figure 6.4 for the SVD cell line under inhibition with the inhibitor D3. Figures of the remaining data for each cell line and inhibitor type can be found in Appendix E. Each subplot in Figure 6.4 shows the MFI (as defined by Equation (6.2)) for a single protein and cellular compartment and for the 10 different inhibitor concentrations, plotted as different colours as indicated by the figure legend. From the figure it can be seen that the inhibitor causes a decrease in pEGFR, pMAPK and pRSK at the earliest time points and in all three cellular compartments, corresponding with an increase in tEGFR. This is generally the expected behaviour since the inhibitors are designed to down-regulate the MAPK pathway and hence a decrease in the abundance of pEGFR, pMAPK and pRSK is promising. An increase in tEGFR however could imply that a feedback mechanism has come into place, whereby the cells are trying to account for the down-regulation in MAPK signalling by increasing the synthesis of EGFR and hence returning the signalling to the pre-dose level. Indeed, by the later time points, and for all inhibitor concentrations, most of these initial trends (down-regulation of pEGFR and pMAPK in particular) start to trend in the opposite direction as initially seen, even returning to baseline in some cases. This is a common theme in the data for several inhibitor types and both cell lines, where good examples can be seen in Figures E.3 (for the inhibitor D3 in the WT cell line) and E.15 (for the inhibitor D7 in the SVD cell line). In Figure E.3 for example, the same trend in pEGFR and pMAPK can be seen as in Figure 6.4, but also in this figure after an initial up-regulation, tEGFR returns to baseline and after an initial down-regulation, pRSK can be seen to be increasing. From comparing figures of the data alone it is difficult to determine any other trends in the data, such as differences between cell lines or inhibitor types, for the effect of the inhibitors on protein abundance, and hence in Section 6.2, statistical techniques are used to look for differences between groupings of the data.

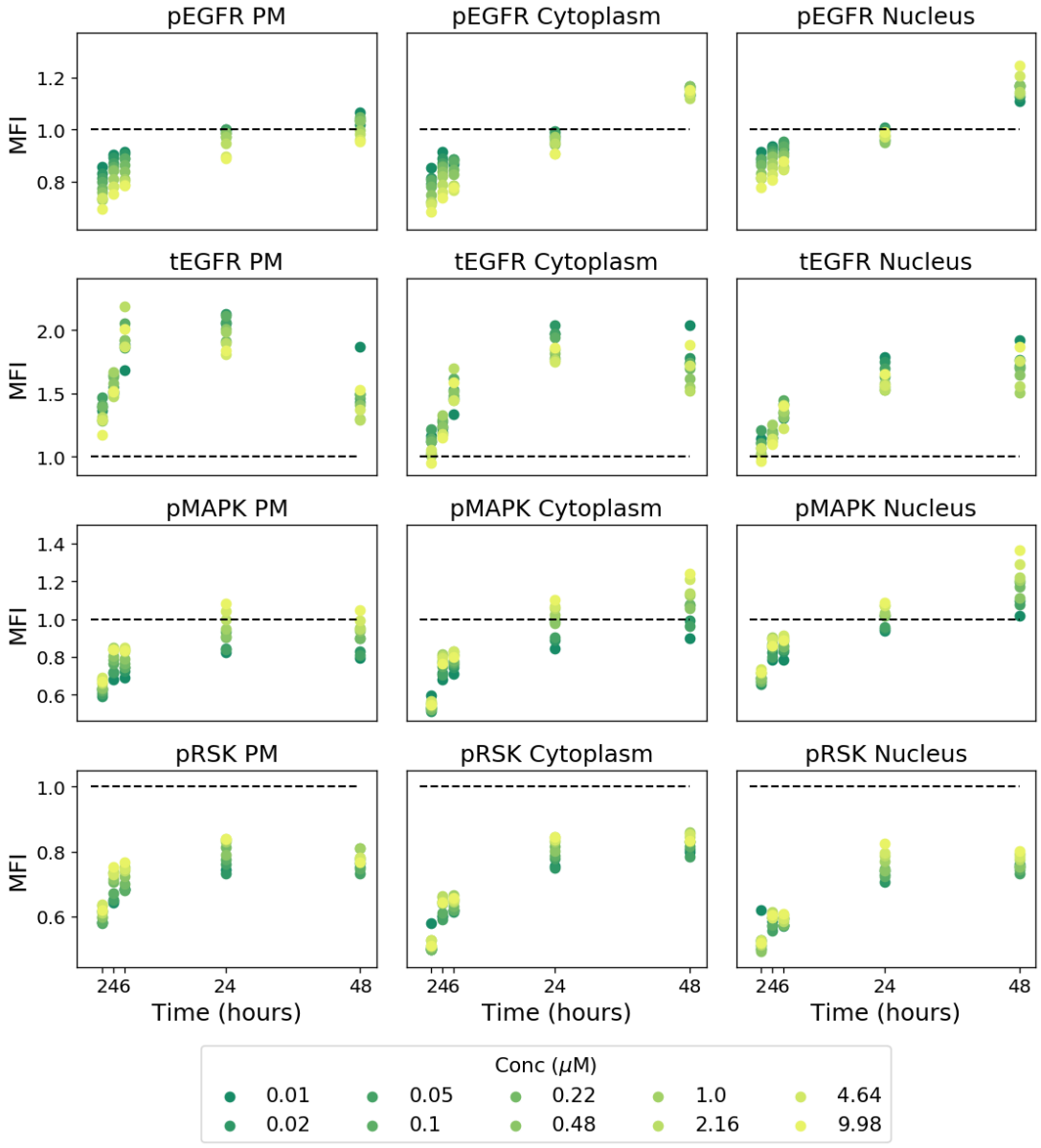

Figure 6.4: Figure of the MFI data in the SVD cell line under inhibition with inhibitor type D3. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

6.2 Statistical analysis

In this section, statistical analysis is carried out using the MFI data defined in Section 6.1.2 in order to find statistically significant differences between groupings. In particular, it was of interest to determine whether the different inhibitors (D1-D8) had differential effects on the levels of the proteins of interest. Furthermore, ideally, a TKI will inhibit only mutant EGFR, and have no effect on WT EGFR, since only the cancerous cells should be affected by the drug. Therefore it was also of interest to determine whether there were differences in the changes in protein levels caused by the inhibitors, between cell lines (WT and SVD).

6.2.1 Analysis of variance

Two-way analysis of variance (ANOVA) is a statistical method used to determine how the mean of a variable is affected by two categorical variables. A description of the method is given in Section 2.6.1. In this case, the two categorical variables (also known as independent variables) are cell line, with two levels (WT and SVD), and inhibitor, with eight levels (D1-D8). The dependent variable is the MFI for a particular protein at a particular cellular compartment, time point and inhibitor concentration. Although it is also possible to carry out higher order ANOVA, such as three- and four-way ANOVA, here only two-way ANOVA is employed since it is already expected that there should be some differences in the data points as a result of time point, and inhibitor concentration. An assumption of ANOVA is that the dependent variable is normally distributed, which can be tested using, for example, the Shapiro-Wilk test (Marques de Sá, 2003). However, here there are only, at most, two repeats of each individual experiment, and hence one cannot test for normality. Given that the dependent variable is repeats of exactly the same experiment, any variation in the data is due to experimental noise and hence, although it cannot be formally statistically tested, it is still assumed here that the data is normally distributed and ANOVA is used.

In the ANOVA, the effects of the two independent variables (cell line and inhibitor type) individually, on the dependent variable (MFI for a particular protein at a particular cellular compartment, time point and inhibitor concentration), are called the *main effects*. It can also be tested whether one independent variable

affects the dependent variable in the same way across all levels of the other independent variable. For example, one can ask questions such as: does the effect on the MFI of pEGFR in the cytoplasm at time 2 hours post dose with 1 μM inhibitor, caused by the inhibitor D1, depend on the cell line? By asking such questions, it is assumed in the ANOVA that an *interaction effect* between the two independent variables may be present. The results of the ANOVA identify which of the effects (main or interaction) are statistically significant to the values of the dependent variable. There are three null hypotheses which are tested for by using two-way ANOVA,

H_{01} : the means of both groups defined by cell line are equal,

H_{02} : the means of the 8 groups defined by inhibitor are equal, and

H_{03} : there is no interaction between inhibitor type and cell line,

where H_{01} and H_{02} correspond to the main effects and H_{03} corresponds to the interaction effect. An effect (main or interaction) is considered statistically significant (*i.e.* the corresponding null hypothesis is rejected) if the ANOVA results in a p-value for this effect less than a threshold α, where here $\alpha = 0.05$.

In order to reduce the total number of ANOVA runs required to analyse the data, only four of the ten inhibitor concentrations are considered, chosen to be representative of the whole range of concentrations. Thus, for the statistical analysis, $d \in D = \{9.98, 1.00, 0.10, 0.01\}$. There are five time points, and for each time point and inhibitor concentration, the four proteins, E, at the three cellular compartments, P, are considered. Thus in total $4 \times 5 \times 4 \times 3 = 240$ individual two-way ANOVA tests are run. The results of these analyses are visualised in Figure 6.5 where in each column of the figure there are 240 individual pixels, one for each protein and cellular compartment (y-axis of a subplot), time point (x-axis of a subplot) and inhibitor concentration (row of the figure). The columns represent the effects, where the first column is the cell line main effect, the second column is the inhibitor main effect and the third column is the interaction effect. A blue, asterisk-annotated pixel indicates that the null hypothesis for the effect in the column title was rejected (at the 5% level) whereas a black pixel indicates that it was accepted. A blue, asterisk-annotated pixel in the inhibitor column

for example, indicates that there is a significant difference between the mean values of the MFI (for that particular protein, cellular compartment, time point and inhibitor concentration) for at least two of the eight groupings defined by inhibitor type. This is important as it implies that at least one inhibitor type is acting differently from the others in its ability to down-regulate the MAPK pathway.

From Figure 6.5 it can firstly be noted that both the cell line and inhibitor main effects are statistically significant for many different pixels (*i.e.* they are asterisk-annotated and coloured in blue). The interaction effect is less significant overall, however there are some interesting groupings. For example, the interaction effect is often significant for the latest time points for the protein tEGFR. Figure 6.6 shows the disparity between cell lines for this protein at the PM after treatment with 1 μM of the different inhibitors. It is clear that tEGFR is up-regulated to a much greater extent in SVD cells than WT cells, after treatment with the inhibitors D4, D5 and D6. This could indicate that these three inhibitor types are inducing a stronger feedback effect in the SVD cell line than in the WT cell line. For the greatest two concentrations of inhibitor, the interaction effect is often also significant for pRSK across several time points and cellular compartments. Figure 6.7 shows the MFI data for pRSK in the cytoplasm at an inhibitor concentration of 1 μM in the WT and SVD cell lines, for each inhibitor type. From this figure (and other similar figures not presented here), it can be concluded that the significance of the interaction effect for pRSK is coming from the fact that the inhibitors D4, D5 and D6 have a lesser effect on the down-regulation of pRSK in the WT cell line than they do in the SVD cell line, whereas all other inhibitor types have a similar effect on pRSK in both cell lines. In general therefore, it appears than when the interaction effect is significant it is due to the inhibitors D4, D5 and D6 acting differently on the proteins pRSK and tEGFR between the two cell lines.

In the case of the cell line main effect (first column of Figure 6.5), it can be seen that the number of statistically significant differences between the mean of the two groups defined by cell line generally increases with both time and inhibitor concentration. In fact, for the latest two time points, and the highest two inhibitor concentrations, the cell line is a significant main effect for almost

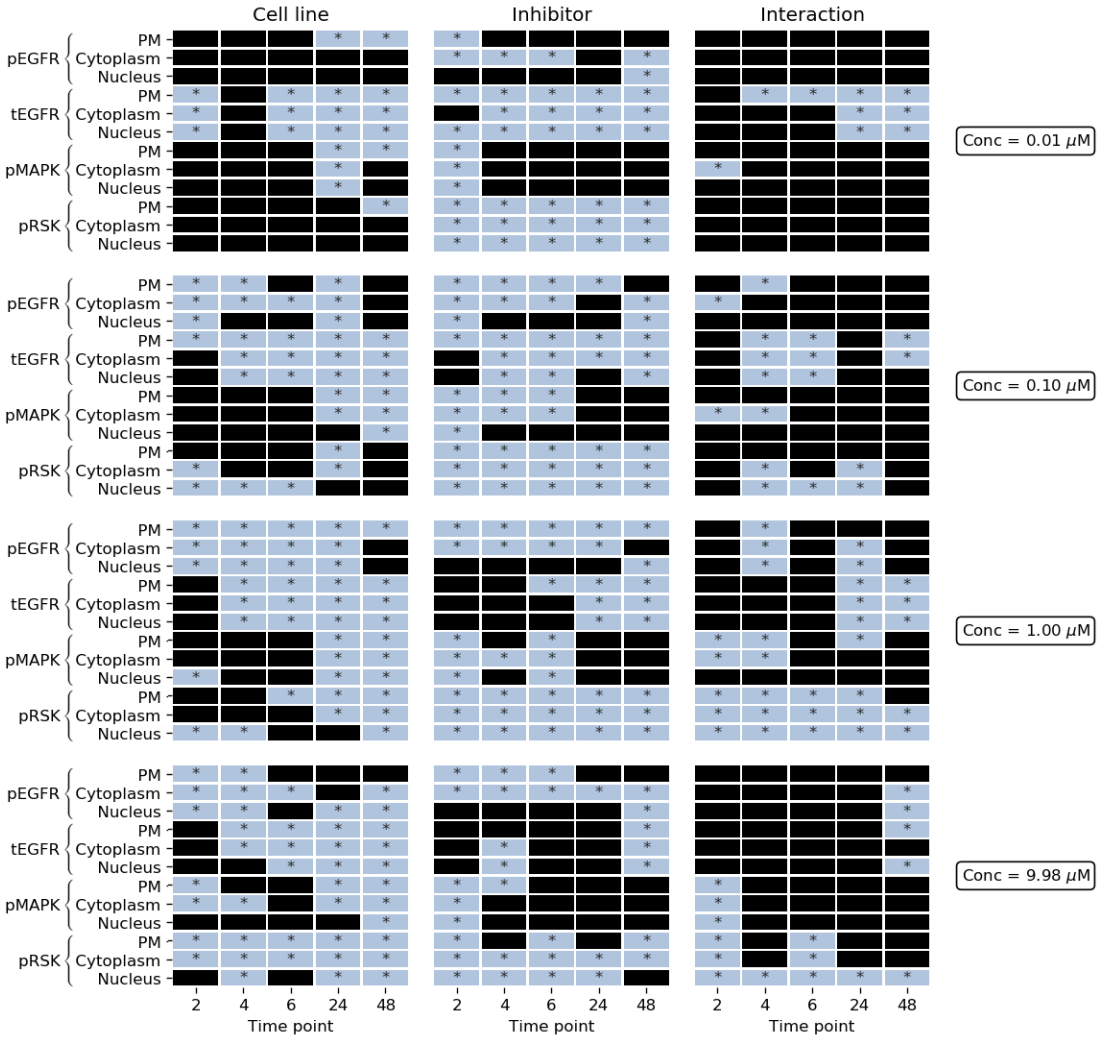

Figure 6.5: Results of the two-way ANOVA for each combination of protein, cellular compartment, time point and inhibitor concentration. A blue, asterisk-annotated pixel indicates that the null hypothesis corresponding to the effect in the column title was rejected (at the 5% level) whereas a black pixel indicates that it was accepted.

all proteins and cellular compartments, at the 5% level. As with the interaction effect, the difference in the data between cell lines (as the main effect) is most noticeable for the proteins tEGFR and pRSK. For example, by comparing Figures E.6 and E.14 which show the MFI data for inhibitor D6 in the WT and SVD cell lines respectively, one can see that tEGFR (in all cellular compartments) is

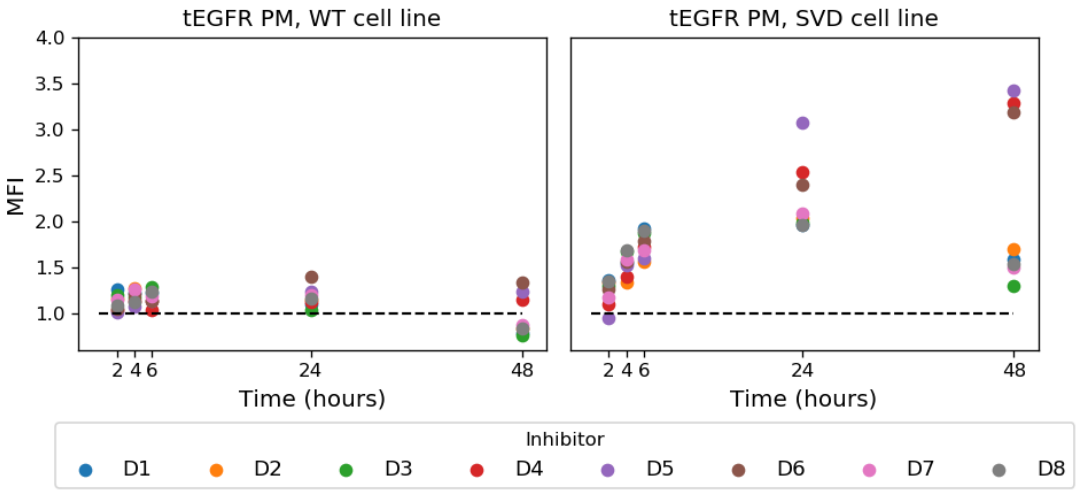

Figure 6.6: Scatter plots of the MFI data for tEGFR at the PM, under inhibition at a concentration of 1 μM of the TKI, in WT cells (**left**) and SVD cells (**right**). The colour of a point indicates the inhibitor type used, as given in the figure legend.

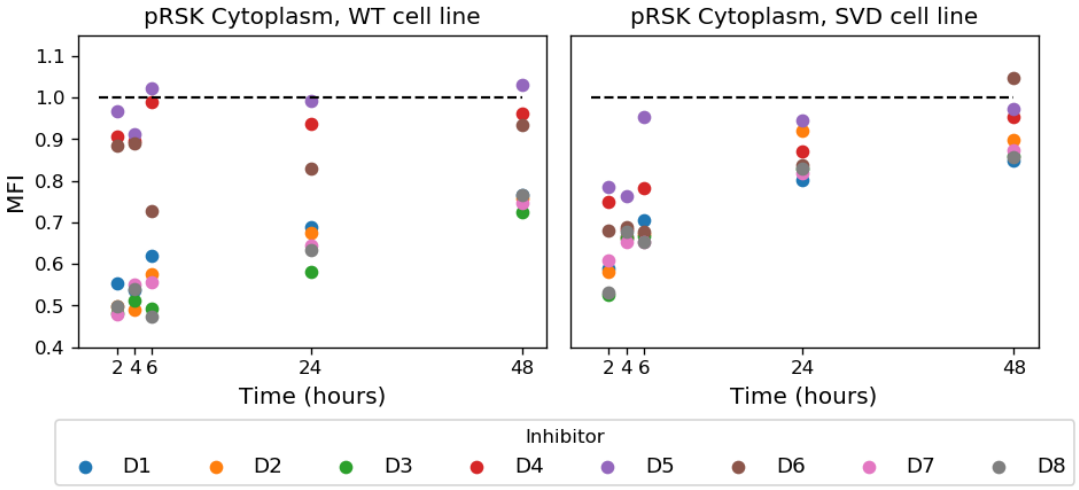

Figure 6.7: Scatter plots of the MFI data for pRSK at the cytoplasm, under inhibition at a concentration of 1 μM of the TKI, in WT cells (**left**) and SVD cells (**right**). The colour of a point indicates the inhibitor type used, as given in the figure legend.

much greater increased from baseline in the SVD cell line than in the WT cell line (note the different scales on the y-axis). As indicated by Figure 6.5, this is

particularly noticeable for the latest two time points whereby the MFI data for tEGFR in the SVD cell line reaches levels of around 3.5 and the corresponding data in the WT cell line only reaches levels of around 1.6. For both cell lines and all inhibitors, pMAPK is initially down-regulated in all three compartments, but by time 48 hours, the level has in most cases returned to baseline or even above baseline where this trend is seen more so in the SVD cells (for example compare the pMAPK rows of Figures E.1 and E.9, which correspond to inhibitor D1, particularly the subplots corresponding to the cytoplasm and the nucleus). It is difficult to see a trend in the MFI data for pEGFR in terms of a cell line or inhibitor effect by examining figures of the data alone.

Also interesting is the effect of inhibitor type on the MFI for each protein and cellular compartment. It can be seen from Figure 6.5 that there are a number of proteins, cellular compartments, time points and inhibitor concentrations for which the inhibitor type has a significant effect, indicating that for at least two inhibitor types there is a significant difference between the means of the data. The ANOVA result can only determine *if* there is a significant difference between groups, *i.e.* whether or not to reject the null hypothesis, but it cannot tell us which groups are statistically significantly different, in the case where an effect has more than two levels (which is the case for the inhibitor type since there are 8 levels). For this, one can use a post-hoc analysis, as is carried out for the inhibitor types in Section 6.2.2.

6.2.2 Post-hoc analysis

Tukey's honest significant difference (HSD) test is a post-hoc analysis to simultaneously compare the means of pairwise groupings and the method is described in Section 2.6.2. When a significant difference is identified by the ANOVA for the inhibitor effect, Tukey's HSD test can then be used to find which of the means of the data for the $\binom{8}{2} = 28$ possible pairs of inhibitors have a significant difference. At first glance it may seem that one could just perform 28 individual t-tests here, however, as explained by Barnette & McLean (2005), the drawback of this method is the accumulation of type 1 errors, where a type 1 error (or "false positive") is the rejection of a null hypothesis when it is actually true. For

a t-test with significance level α (*i.e.* the p-value is α), the probability of a type 1 error is α, and when carrying out n t-tests simultaneously, the probability of a type 1 error is $1-(1-\alpha)^n$ (Lee & Lee, 2018). Hence for the 28 possible pairs of inhibitors here, if individual t-tests were used to compare the means of the data for each pair at a significance level of $\alpha = 0.05$, the probability of at least one type 1 error would be $1-(1-0.05)^{28} = 0.76$, as opposed to the desired 0.05. Tukey's HSD test overcomes this problem by adjusting the p-value for multiple testing, and hence is used here to compare the data for the different pairs of inhibitors. In Figure 6.5, one can observe 38, 44, 39 and 30 statistically significant results (blue pixels) for the inhibitor effect for the concentrations 0.01, 0.10, 1.00 and 9.98 μM, respectively, meaning one must carry out Tukey's HSD test, $38+44+39+30 = 151$ times in total. The results of this analysis are summarised in Figure 6.8, where each column (and colour) represents a different concentration of inhibitor. All possible pairs of the eight inhibitors are listed on the y-axis, and a bar indicates the frequency of statistically significant differences between the mean of the data for the two inhibitors in the pairing, across all 4 proteins, 3 cellular compartments and 5 time points. As indicated by the ANOVA results in the previous section, the maximal frequencies for the bar charts are 38, 44, 39 and 30 for the concentrations 0.01, 0.10, 1.00 and 9.98 μM, respectively.

From Figure 6.8, the inhibitors which are most frequently different from other inhibitors can be identified. For example, the inhibitor D3 has a high frequency of significant differences from most of the other inhibitors, particularly at the lowest concentrations. In general, as the concentration of inhibitor increases (moving from left to right in Figure 6.8), the frequency of significant differences between inhibitor types decreases, indicating that the inhibitors have a very similar effect on the protein levels at high concentrations.

The results of Tukey's HSD test can also be represented as a network, as in Figures 6.9 - 6.12 where each of the four figures shows the results for one of the four inhibitor concentrations studied here. There are 12 subplots in each figure, one for each protein and time point, and within each subplot there are 8 nodes, where node 1 represents inhibitor type D1 and so on. Two nodes are connected if, for that concentration, protein and time point, the means of the data groups defined by the two inhibitor types (labelled on the nodes) are statistically

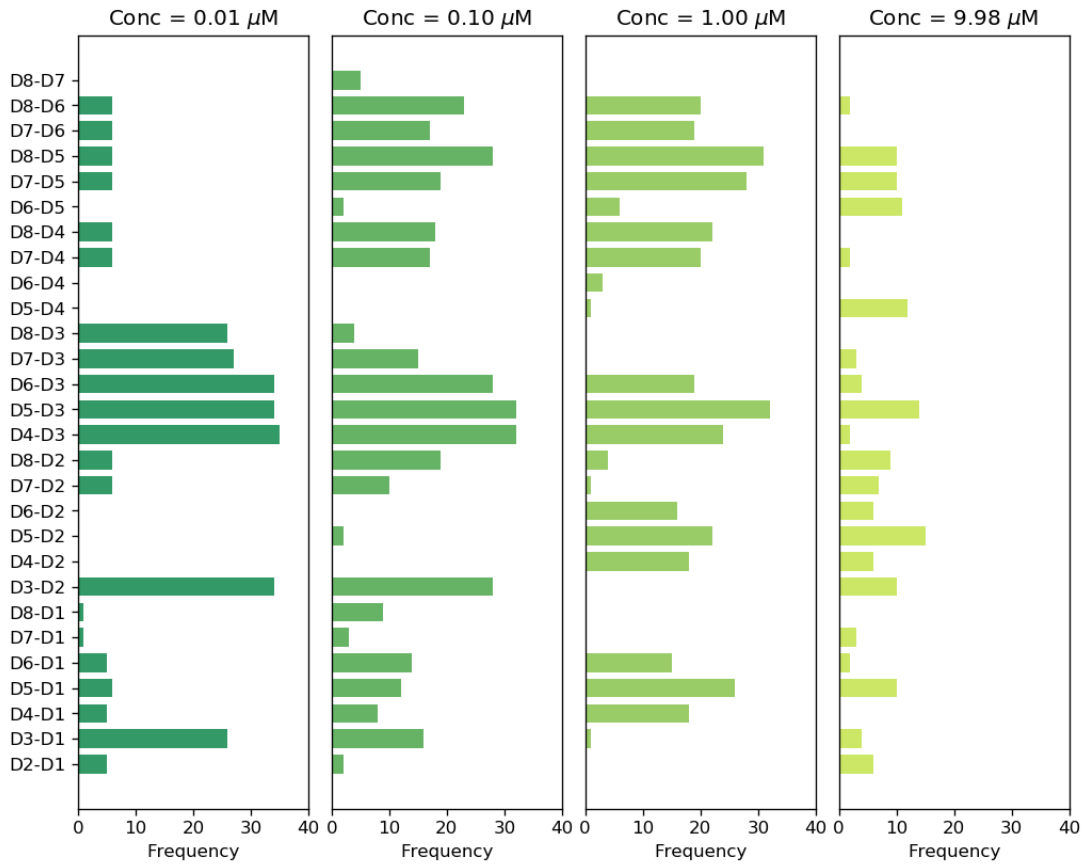

Figure 6.8: Bar plots of the results of Tukey's HSD test for each of the four inhibitor concentrations considered, from 0.01 μM (**far left**) to 9.98 μM (**far right**). Each bar chart shows the frequency of a significant difference between the inhibitor pairs on the y-axis over all time points, proteins and cellular compartments.

significantly different. Finally, the nodes are connected in up to three different colours, where the colours represent the different cellular compartments, listed in the figure legend.

Starting by looking at the bar chart for the lowest concentration of inhibitor, 0.01 μM (Figure 6.8, left hand panel), it is obvious that D3 is frequently statistically significantly different from the other inhibitor types. From the network figure for this concentration, Figure 6.9, one can identify that the differences between D3 and the other inhibitors are very common for the protein pRSK, across all cellular compartments and time points, as well as for the protein pMAPK

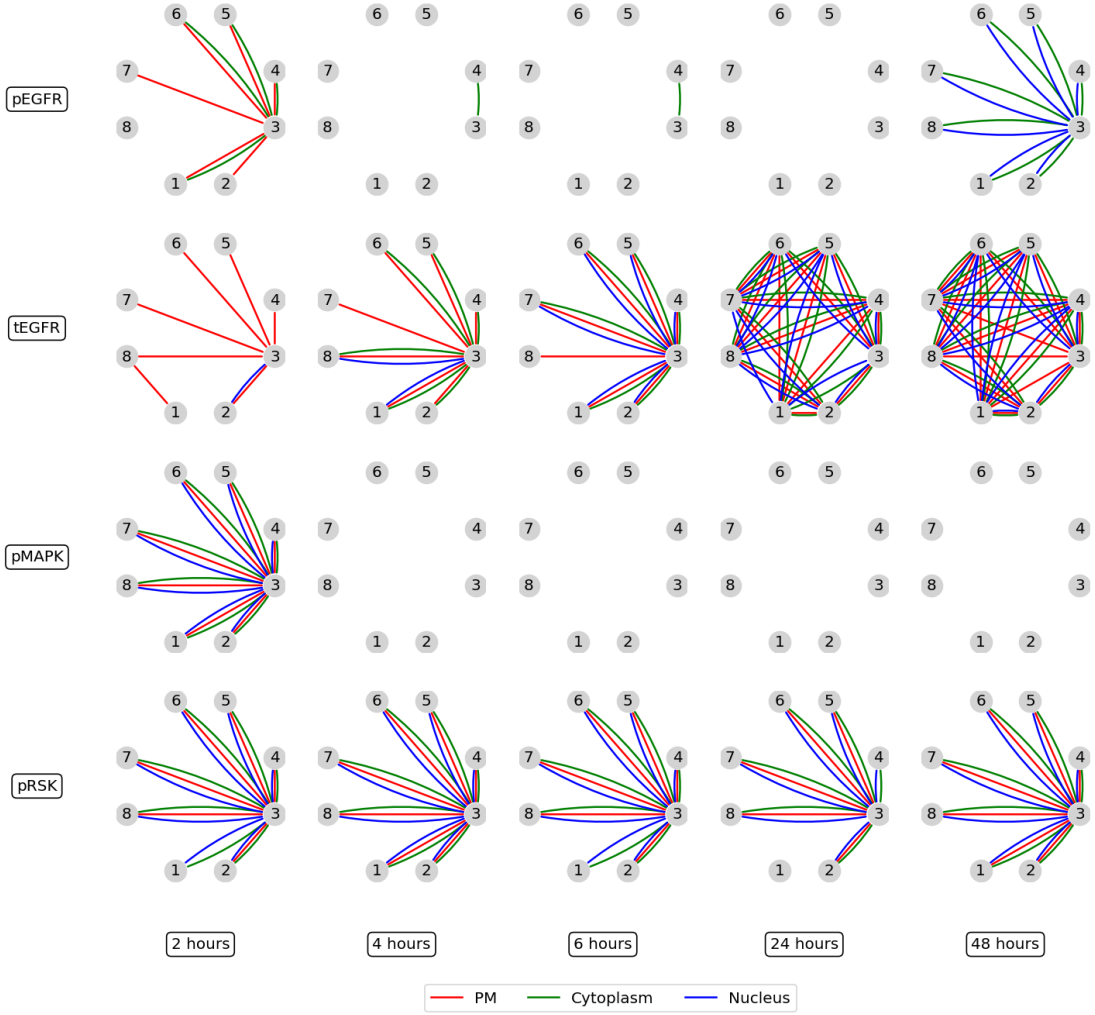

Figure 6.9: Network representation of significant differences between means of data groupings defined by inhibitor type, elucidated by Tukey's HSD test at inhibitor concentration 0.01 μM for each combination of protein and time point. The nodes 1 to 8 represent the inhibitor types D1 - D8, respectively, and they are connected if there is a significant difference between the means of the data groupings defined by this pair of inhibitors. The colour of the connecting line represents the cellular compartment as indicated by the legend.

at time 2 hours post dose in all cellular compartments. In particular, it can be observed from Figure 6.13, that pRSK is down-regulated much more so by D3 than all other inhibitor types in both WT and SVD cells, for this concentration of inhibitor. This disparity between inhibitor types is most pronounced at earlier

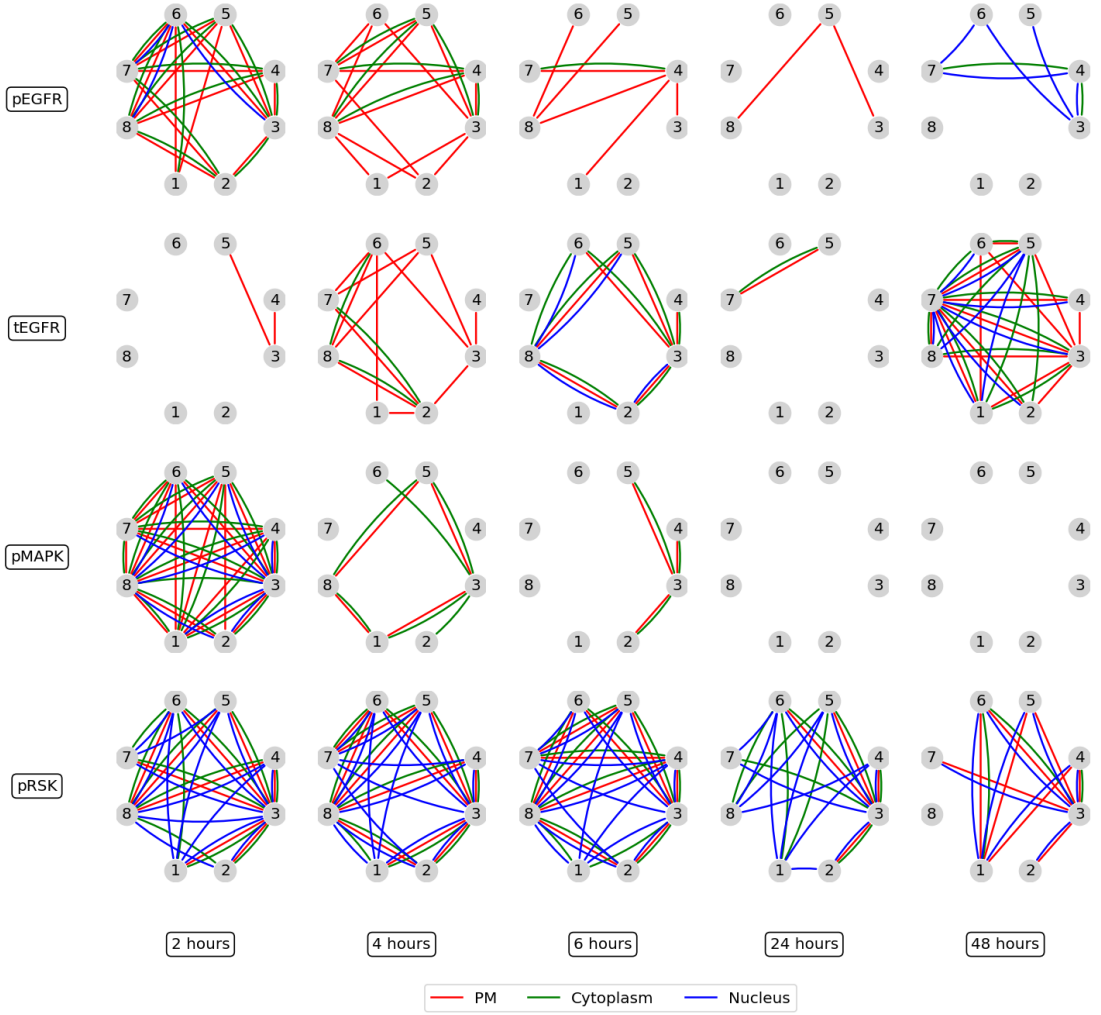

Figure 6.10: Network representation of significant differences between means of data groupings defined by inhibitor type, elucidated by Tukey's HSD test at inhibitor concentration 0.10 μM for each combination of protein and time point. The nodes 1 to 8 represent the inhibitor types D1 - D8, respectively, and they are connected if there is a significant difference between the means of the data groupings defined by this pair of inhibitors. The colour of the connecting line represents the cellular compartment as indicated by the legend.

time points, and in the WT cell line.

There is also a greater up-regulation of tEGFR by D3 than most other inhibitors in SVD cells and there is some evidence of the same trend at early time points in WT cells, as seen in Figure 6.14. This result can also be identified

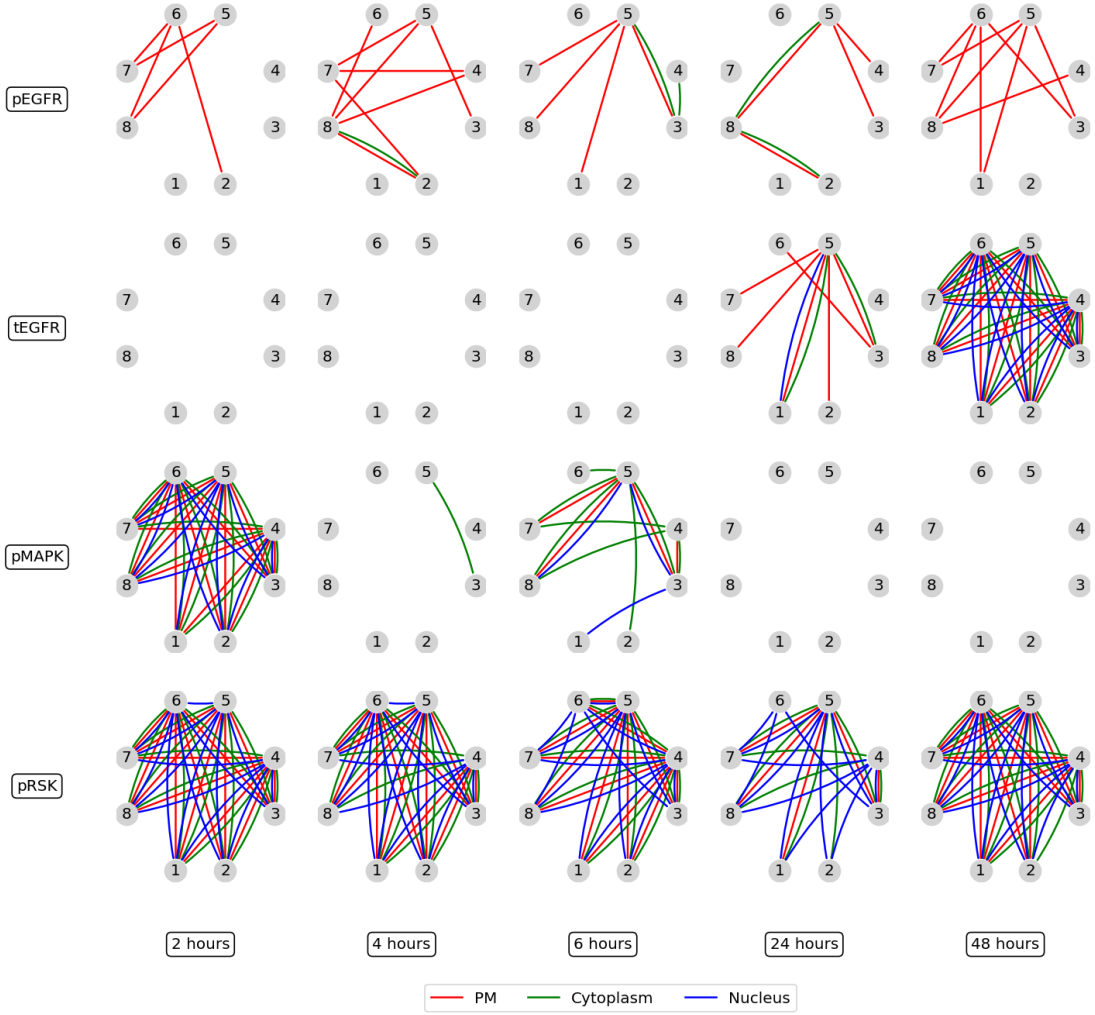

Figure 6.11: Network representation of significant differences between means of data groupings defined by inhibitor type, elucidated by Tukey's HSD test at inhibitor concentration 1.00 μM for each combination of protein and time point. The nodes 1 to 8 represent the inhibitor types D1 - D8, respectively, and they are connected if there is a significant difference between the means of the data groupings defined by this pair of inhibitors. The colour of the connecting line represents the cellular compartment as indicated by the legend.

from Figure 6.9, where in the second row it is clear that D3 has significant differences with almost all other inhibitors at most time points. For later time points this is true for all cellular compartments, however at the earliest time point (2 hours) this is mostly true only at the PM. Biologically, this is intuitive since,

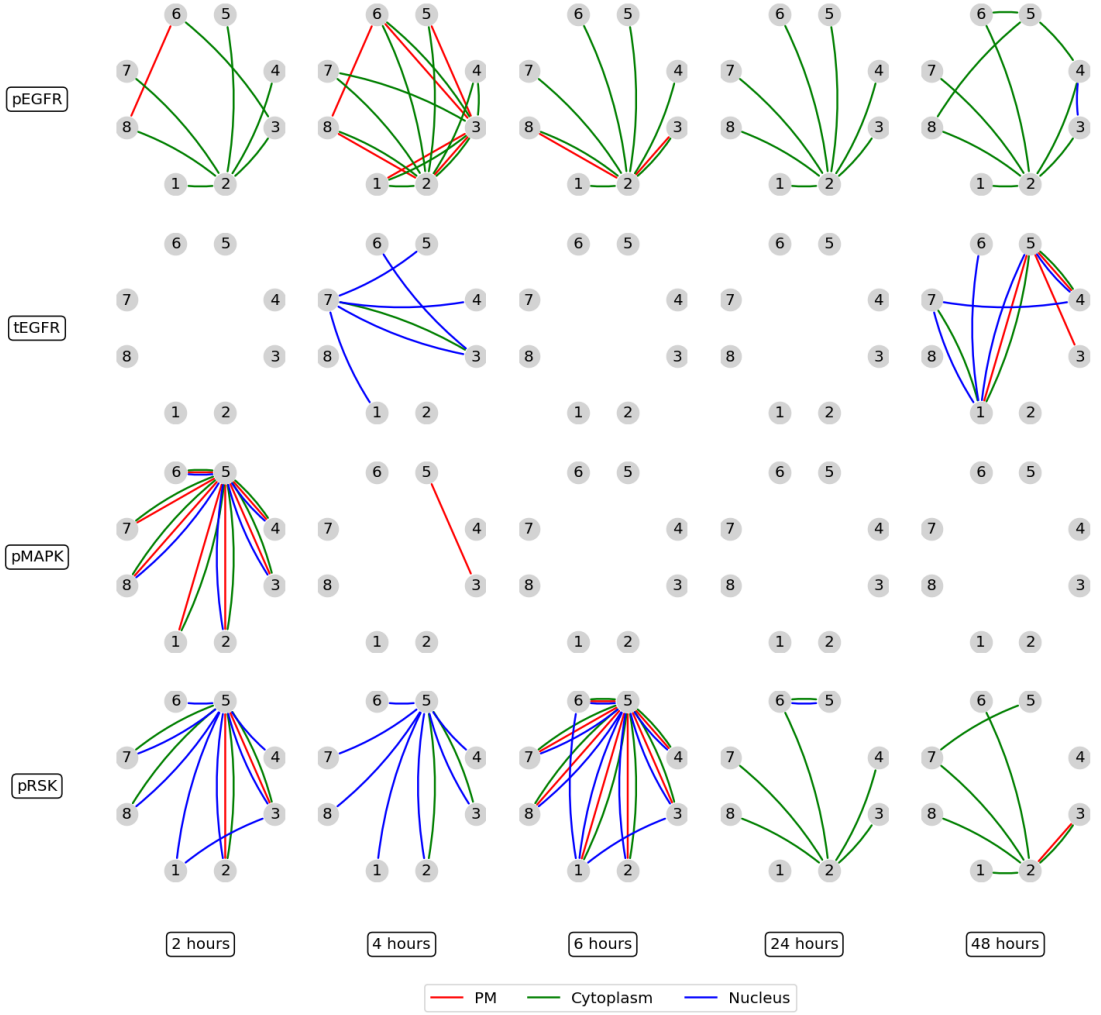

Figure 6.12: Network representation of significant differences between means of data groupings defined by inhibitor type, elucidated by Tukey's HSD test at inhibitor concentration 9.98 μM for each combination of protein and time point. The nodes 1 to 8 represent the inhibitor types D1 - D8, respectively, and they are connected if there is a significant difference between the means of the data groupings defined by this pair of inhibitors. The colour of the connecting line represents the cellular compartment as indicated by the legend.

when new receptors are synthesised, they are delivered to the PM. From Figure 6.14 it can be seen that D7 and D8 are also having a relatively large effect on the up-regulation of tEGFR in SVD cells. This explains the significant differences between D7/8 and the other inhibitors seen in the first subplot of Figure 6.8.

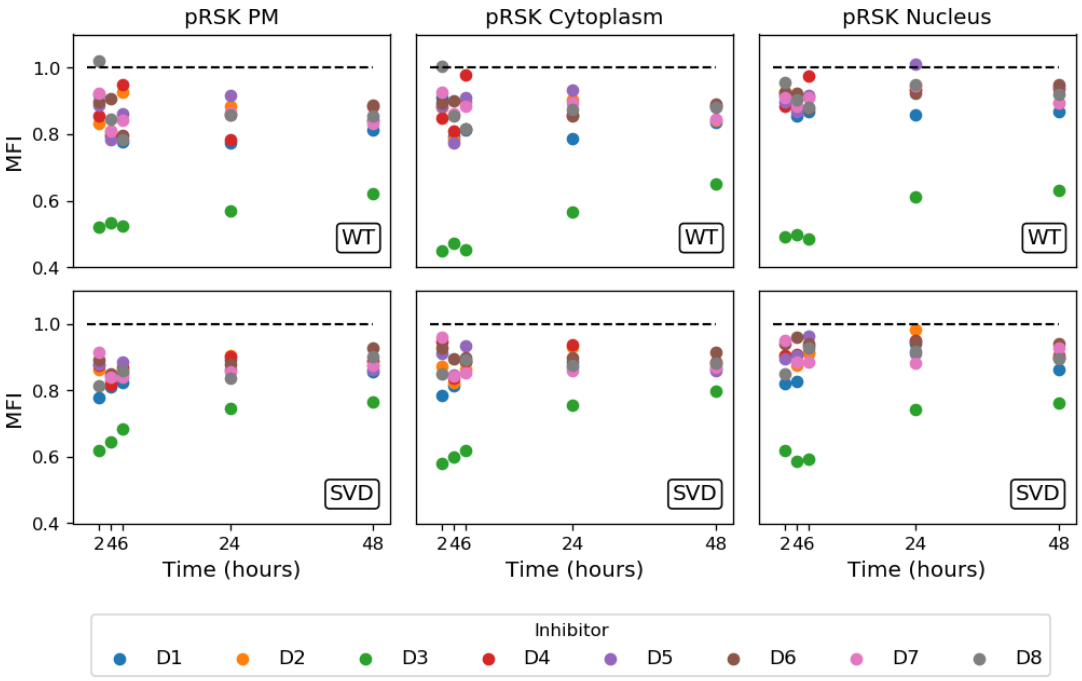

Figure 6.13: Scatter plots of the MFI data for pRSK in the WT cell line (**top row**) and SVD cell line (**bottom row**), and for each cellular compartment (columns of the figure). The colour of a point represents the inhibitor type as given in the legend, at a concentration of 0.01 μM.

There is also a relatively high frequency associated with the pairings involving D1 in Figure 6.8, first subplot. It appears that D1 acts similarly to D3 at the lowest inhibitor concentration, down-regulating pRSK and up-regulating tEGFR more than the other inhibitors, but to a lesser extent than D3, hence the lesser frequency for these pairs than pairs involving D3. This can be observed in Figures 6.13 and 6.14. From Figure 6.9, there is little difference between the inhibitor types on the down-regulation of pEGFR or pMAPK at the lowest inhibitor concentration, other than the inhibitor D3 being statistically significantly different from the other inhibitors at select time points (2 hours post dose for pEGFR and pMAPK and also 48 hours post dose for pEGFR). In all cases, this is because the inhibitor D3 is having a significantly greater effect than the other inhibitors.

From the second subplot in Figure 6.8 for the concentration 0.10 μM a similar trend can be seen as for the lower concentration, where D3 is still frequently

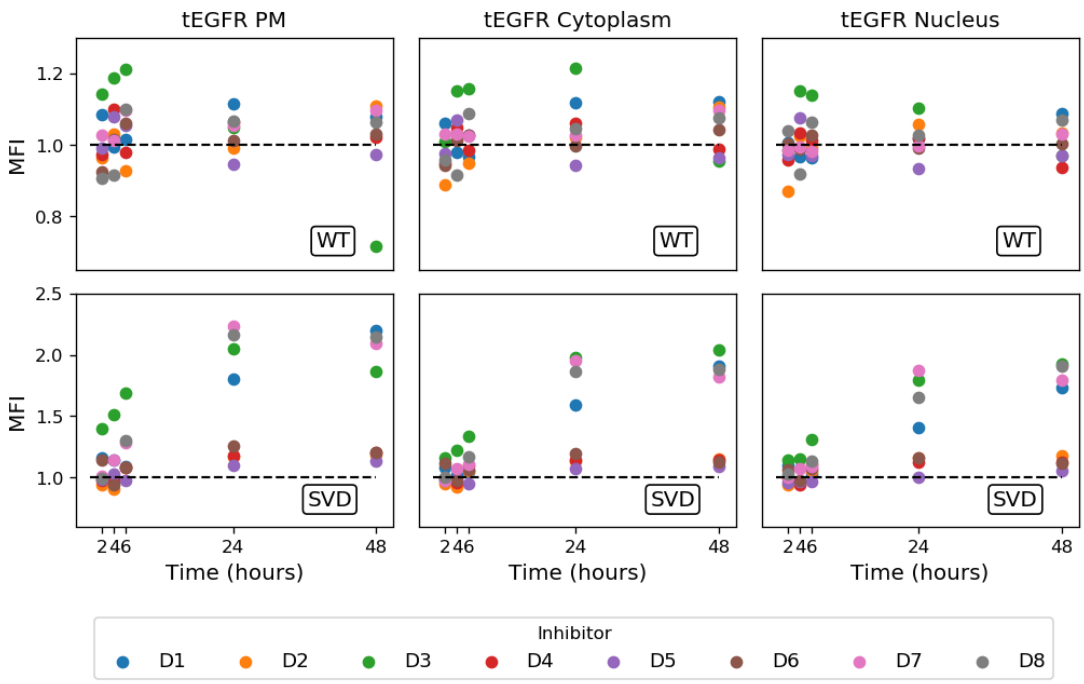

Figure 6.14: Scatter plots of the MFI data for tEGFR in the WT cell line (**top row**) and SVD cell line (**bottom row**), and for each cellular compartment (columns of the figure). The colour of a point represents the inhibitor type as given in the legend, at a concentration of 0.01 μM.

significantly different from all other inhibitors, but now less frequently for the inhibitors D1, D7 and D8. This is particularly evident in the pRSK row of Figure 6.10. In fact it can be seen that D1, D7 and D8 also have relatively frequent differences with the inhibitors D2, D4, D5 and D6. From the data it can be observed that this is because D1, D7 and D8 are beginning to equalise with D3 in their ability to down-regulate pRSK (see Figure 6.15) and to some extent to up-regulate tEGFR (data not shown), at this concentration of inhibitor.

It is also found that pEGFR and pMAPK are more significantly down-regulated at early time points than by the lower concentration of inhibitor, and this down-regulation is most pronounced for the inhibitors D1, D3, D7 and D8. This is shown in Figure 6.16 for pEGFR (data not shown for pMAPK). By time 48 hours however, there is an up-regulation of these proteins as compared to the DMSO control, again more significantly for the inhibitors D1, D3, D7 and D8. The other

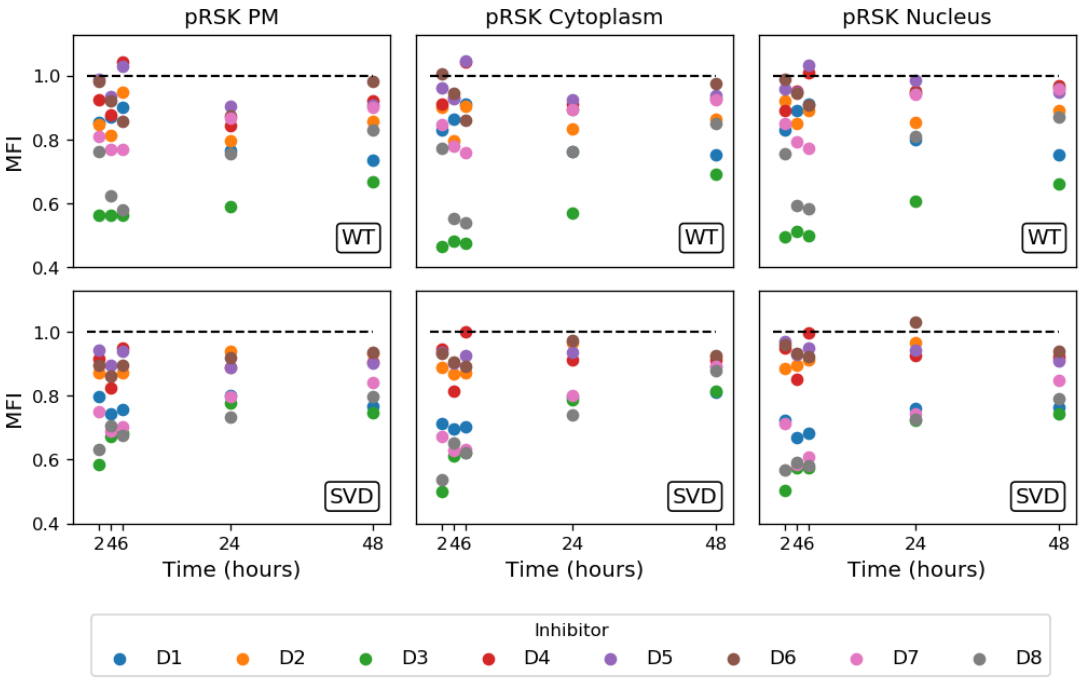

Figure 6.15: Scatter plots of the MFI data for pRSK in the WT cell line (**top row**) and SVD cell line (**bottom row**), and for each cellular compartment (columns of the figure). The colour of a point represents the inhibitor type as given in the legend, at a concentration of 0.10 μM.

inhibitors follow a similar trend to these four, but all data values remain closer to baseline (DMSO control).

At an inhibitor concentration of 1.00 μM, from Figure 6.8, one can see that there are almost no significant differences between any pairings of the inhibitors D1, D3, D7 and D8 across all proteins, compartments and time points. Similarly there are very few significant differences between D2 and any of D1, D3, D7 and D8. All five of these inhibitors have frequent differences with the inhibitors D4, D5 and D6, allowing to separate the inhibitor types into two distinct groups at this concentration. This grouping is coming mostly from differences in down-regulation of pRSK, where D1, D2, D3, D7 and D8 all down-regulate pRSK significantly more so than D4, D5 and D6 and where this is more noticeable in the WT cells than the SVD cells. This can be confirmed by analysis of Figure 6.11 where almost all of the significant differences (connections between nodes)

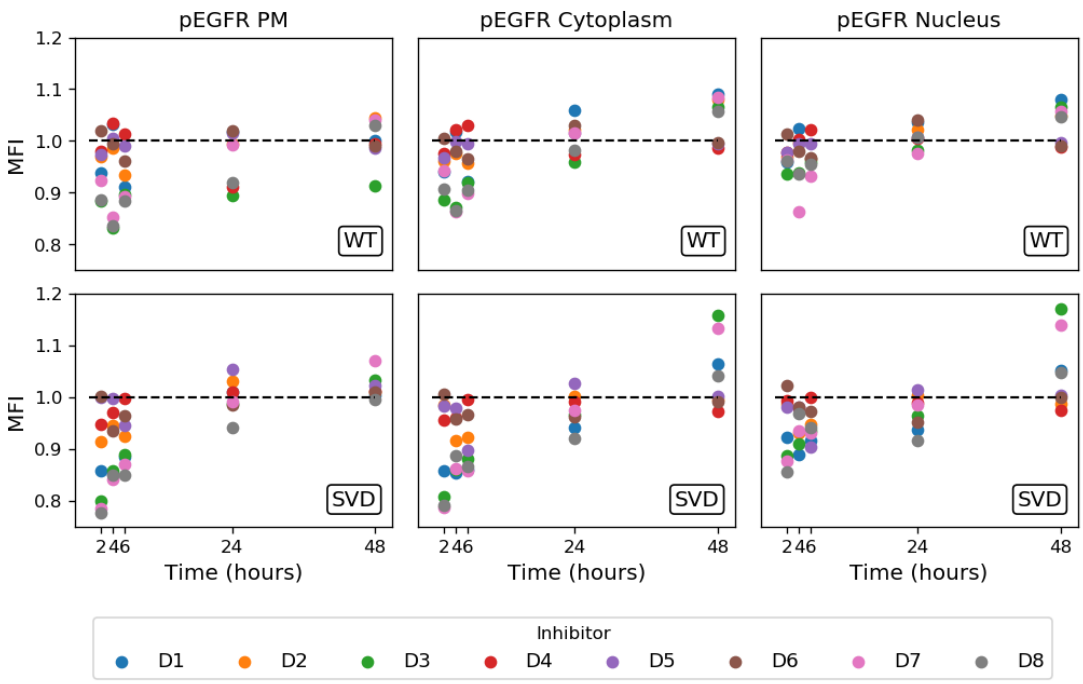

Figure 6.16: Scatter plots of the MFI data for pEGFR in the WT cell line (**top row**) and SVD cell line (**bottom row**), and for each cellular compartment (columns of the figure). The colour of a point represents the inhibitor type as given in the legend, at a concentration of 0.10 μM.

are occurring in the row corresponding to pRSK. Almost all other significant differences seen in Figure 6.11 are occurring at the plasma membrane, and most frequently for the protein pEGFR. Interestingly, at this concentration, D4, D5 and D6 are having the largest effect on up-regulating tEGFR by the latest time points, in both WT and SVD. It is difficult to see trends in the inhibitors for the proteins pEGFR and pMAPK, although at early time points there is evidence from the figures to suggest that pMAPK is down-regulated more so by D1, D3, D7 and D8. This is evident from Figure 6.11 where, similarly to the lower inhibitor concentrations, a relatively large number of significant differences is seen for pMAPK at time 2 hours.

Finally, at an inhibitor concentration of 9.98 μM, from Figures 6.8 and 6.12, D5 is found to be most frequently significantly different from the other inhibitors. From inspection of the data, this is predominantly due to the fact that D5 has

a lesser effect on down-regulation of pRSK than the other inhibitors in the WT cell line (see Figure 6.17, top row). From Figures 6.12 and 6.17 one can note that this is the case in particular for the time points 2, 4 and 6 hours and less so at later time points. D5 also appears to have a large effect on the up-regulation of tEGFR by the latest time point (see second row of Figure 6.12).

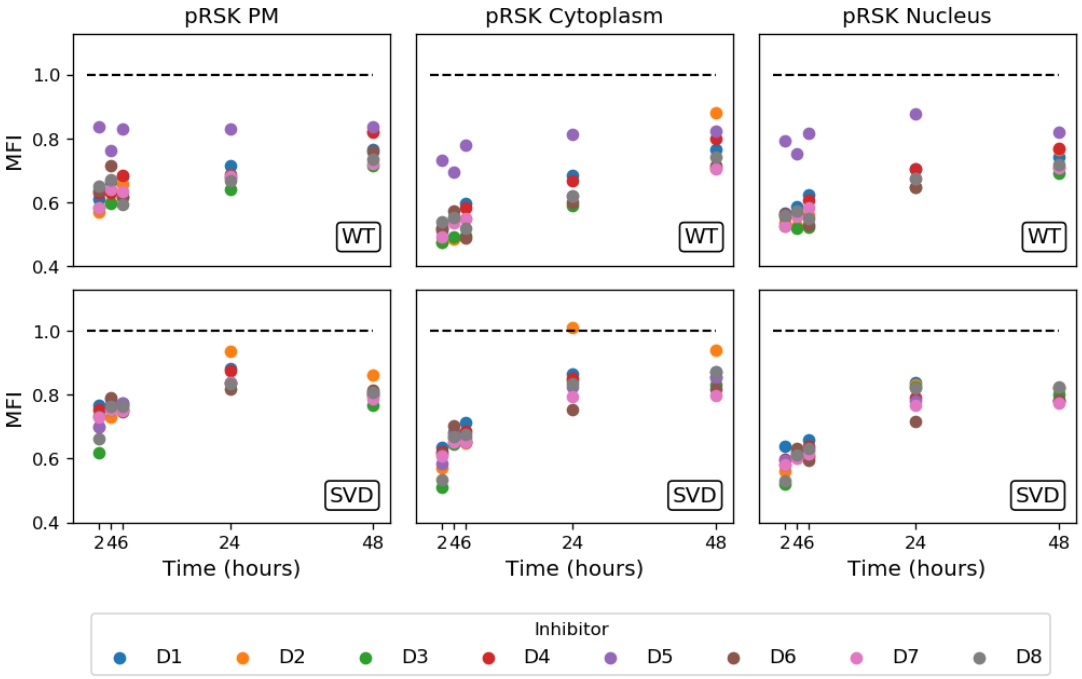

Figure 6.17: Scatter plots of the MFI data for pRSK in the WT cell line (**top row**) and SVD cell line (**bottom row**), and for each cellular compartment (columns of the figure). The colour of a point represents the inhibitor type as given in the legend, at a concentration of 9.98 μM.

From the bar chart, it can be noted that D2 also has some significant differences with all other inhibitors. Looking at the data figures, one can conclude that this is since D2 causes a large up-regulation of pEGFR in the cytoplasm in both WT and SVD cells which is not seen when the cells are treated with the other inhibitors (see Figures E.2 and E.10, top row, middle subplots). D2 also has less of an effect on the down-regulation of pRSK in the cytoplasm in the SVD cell line at the later time points (see Figure 6.17, bottom row).

In general, from the pairwise analysis of the inhibitor types performed here,

it is found that D3 has a large effect on altering protein levels in the MAPK pathway even at the lowest concentrations, and with increasing concentration, the other inhibitor types begin to act in the same way. D1, D2, D7 and D8 have the next largest effects for lower concentrations and only at high concentrations do D4, D5 and D6 have a similar effect. From examining the data figures and the network figures, one can conclude that most of the significant differences between inhibitor types are due to changes in the level of pRSK and tEGFR, where these proteins are down-regulated and up-regulated, respectively, by the inhibitors. Overall, there are very few significant differences between the inhibitor types for the protein pMAPK at any time point other than 2 hours.

To better visualise the differences between inhibitor types on pRSK and tEGFR, the data can be plotted as a violin plot. An example of this is seen in Figure 6.18 for pRSK in the cytoplasm at time 2 hours. For this particular protein, time point and cellular compartment, the data for all 8 inhibitors (columns in each subplot), and 10 concentrations is plotted as violin plots (top subplot) and scatter plots coloured by concentration (bottom subplots). The centre subplots show the correlations between the data points for all concentrations between each inhibitor type. From the top subplot, it is clear that the inhibitor D3 is significantly different from the others, where D3 has a large effect on pRSK for all concentrations, but the other inhibitors only have the same effect at higher concentrations. One can also see how the data distributions vary between cell line, where the WT data distribution is plotted on the left hand side of a violin plot, and the SVD data on the right. There is a statistically significant difference between the means of the WT and SVD data, by the result of a Student's t-test at the 5% level, for the inhibitor D3 only (annotated by an asterisk below the violin plot). Also of interest in this figure, is the negative correlation between the D3 data and the data for all other inhibitors (for all concentrations). One would expect that the correlations would almost always be positive, indicating that larger concentrations of any inhibitor type lead to greater effects on the proteins. However there are some reasonably large negative correlations for pairs involving D3, particularly in the WT cells, which suggests that D3 has the largest effect on the proteins at lower concentrations, rather than higher. In fact this

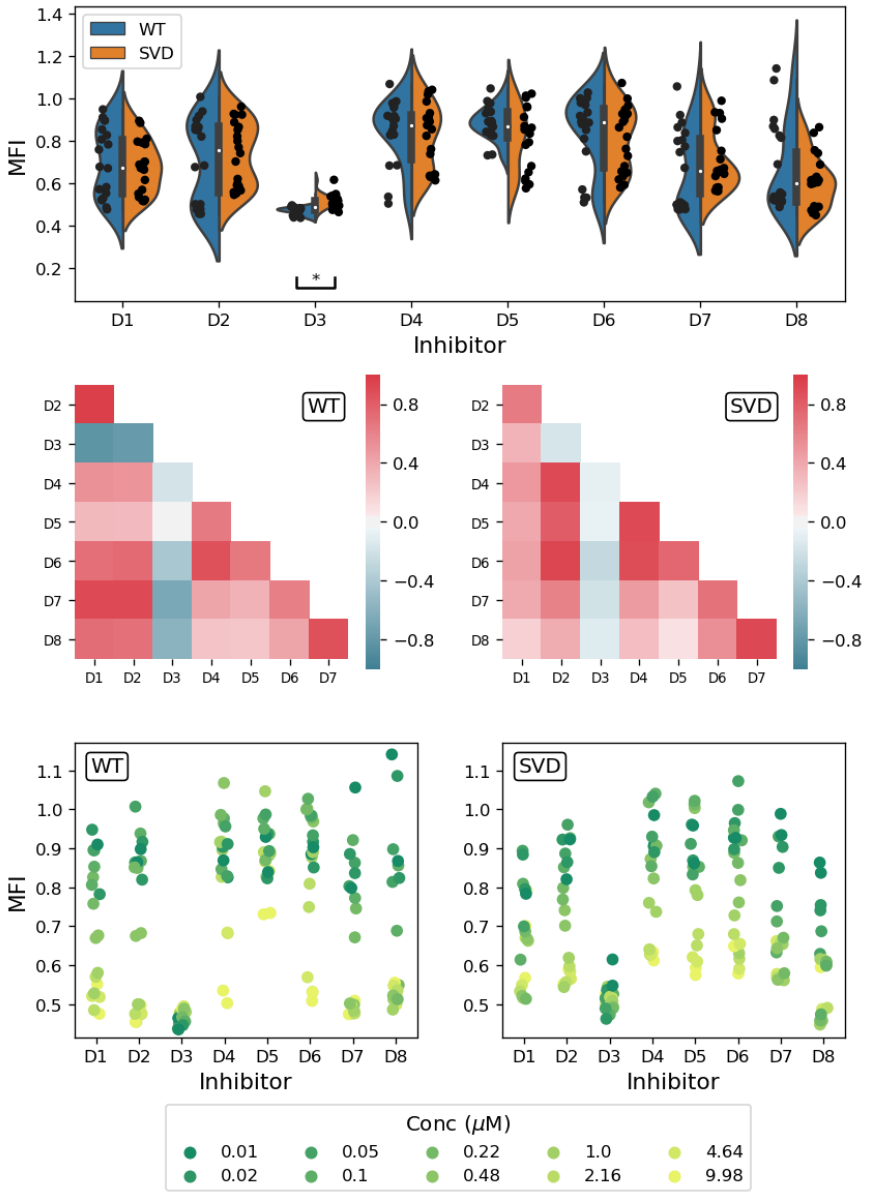

Figure 6.18: Violin plots (**top row**), correlation plots (**middle row**), and scatter plots (**bottom row**) of the data distributions of pRSK in the cytoplasm at time 2 hours. In the violin plots, an asterisk beneath an individual inhibitor plot indicates that the means of the WT and SVD data for this inhibitor type are statistically significantly different (Student's t-test, 5% level). In the scatter plots, the colour of a point represents the inhibitor concentration as given in the legend.

trend in the D3 concentrations is apparent for all time points and cellular compartments for pRSK, as can be seen in Figure 6.19. A similar pattern is found for pMAPK under inhibition with D3, but the reverse (expected) behaviour is found for pEGFR. It is difficult to see any real concentration effect for tEGFR. This negative correlation between inhibitor concentrations is also found in some isolated cases for other pairs of inhibitors, but is most noticeable for pairings involving the inhibitor D3. It is noted however, from Figures such as the violin plot on the top row of Figure 6.18, that the data points for the full range of concentrations for D3 are much more clustered than the corresponding data points for the other inhibitors. This implies that D3 has a very similar effect on the cells at every concentration in the range.

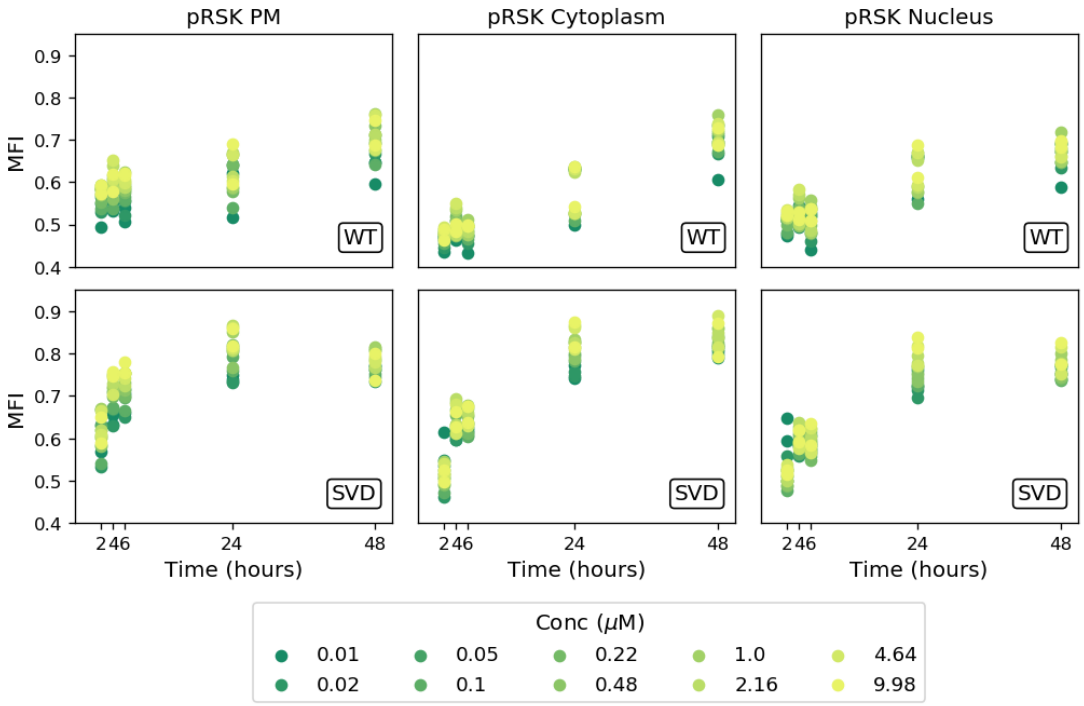

Figure 6.19: Scatter plots of the MFI data for pRSK in the WT cell line (**top row**) and SVD cell line (**bottom row**), and for each cellular compartment (columns of the figure), when treated with the inhibitor D3. The colour of a point represents the inhibitor concentration with units μM as given in the legend.

6.2.3 Principal component analysis

The statistical methods used in this chapter so far have been *univariate*, where each protein, at each cellular compartment, is treated as a single variable and the analysis is carried out for each of the 12 (4 proteins at 3 cellular compartments) variables separately. In order to get an overall picture of the data and to explore further the correlations and potential redundancies in the data, in this section, PCA is carried out with the full dataset. As introduced by Chatfield & Collins (1981) and explained in Section 2.6.3, PCA is a method used to transform a set of variables into a new set of uncorrelated variables, which are linear combinations of the original variables. The new uncorrelated variables are known as principal components (PC), where the first PC is the direction in the data along which the samples show the most variation (Kassambara, 2017). The second PC is then the second most important direction in the data in terms of the variance, and so on. Mathematically, the PCs describing the dataset are the eigenvectors of the standardised sample covariance matrix, and the sum of the corresponding eigenvalues will be equal to the number of variables in the original dataset (here 12). It is generally accepted that if a PC has a corresponding eigenvalue greater than 1, meaning that this PC accounts for more variance than any of the original variables, it is retained for further analysis, where the retained PCs should account for most of the variation in the data (Kassambara, 2017). Where the number of retained PCs is significantly less than the number of original variables, one can conclude that there is redundancy in the original data, since each data point can be well represented by fewer variables.

Here, PCA is used to get an overall picture of the data. The original variables are each of the 12 combinations of protein and cellular compartment, and the data for each cell line, inhibitor, inhibitor concentration and time point is considered together. In particular, the dataset is firstly transposed such that each of the variables becomes a column and there is a row for each combination of cell line, inhibitor, concentration and time point. A row with a missing value for any of the variables, due to the removal of outliers in Section 6.1.2, was removed, resulting in a dataset with 12 variables and 1575 samples (rows). The correlation between each of the original variables is plotted in Figure 6.20, where it can be seen that

there is a strong positive correlation between the three cellular compartments for all four proteins. There are some other reasonably large positive correlations between pairs of variables involving pEGFR, pMAPK and pRSK, where tEGFR is the least correlated with the other variables. One might assume that there should be negative correlations present between the tEGFR variables and the variables relating to the other proteins, since in general there is an increase of tEGFR from baseline and a decrease in the other proteins from baseline. The lack of negative correlations can be explained since, in many of the data figures (see Appendix E), the decrease in pEGFR, pMAPK and pRSK is maximal at time point 2 hours and then there is a gradual increase back to baseline or above baseline, whereas the increase in tEGFR is more gradual, starting from 2 hours up to around 6 or 24 hours. Therefore, for the earlier time points these variables would be positively correlated as the data is increasing together, until the tEGFR begins to decrease back to baseline, causing a negative correlation at later time points, hence resulting in an "overall" correlation of approximately zero. To explore further the data, PCA was then carried out using the *FactoMineR* package in R.

Figure 6.21 shows a scree plot of the percentage contributions of each of the 12 determined PCs to the overall variance in the data. It can be seen that the first four PCs account for a total of over 95% of the variability, and from the results of the PCA, each of these PCs have an eigenvalue greater than 1, whereas the eigenvalues associated with all other PCs are less than 1. This indicates that the data can be represented well by only four variables as opposed to the original 12, and hence there is a lot of redundancy in the data. Interestingly, the first PC accounts for almost 50% of the variability alone, and approximately 70% of the data can be represented well by the first two PCs.

To see how each of the original variables correlates with the first two PCs, and how well each variable is represented by these PCs, one can plot a correlation circle, as in Figure 6.22. The length of a vector (arrow) on the plot indicates how well this variable is represented by the first two PCs. This quality of representation is also given by the colour of a vector, where the colour shows the cos2 value as given by the colour bar. The cos2 stands for the "squared cosine" which is an index such that values closer to 1 indicate that the variable is better represented by the PC and values closer to 0 indicate lower quality representation.

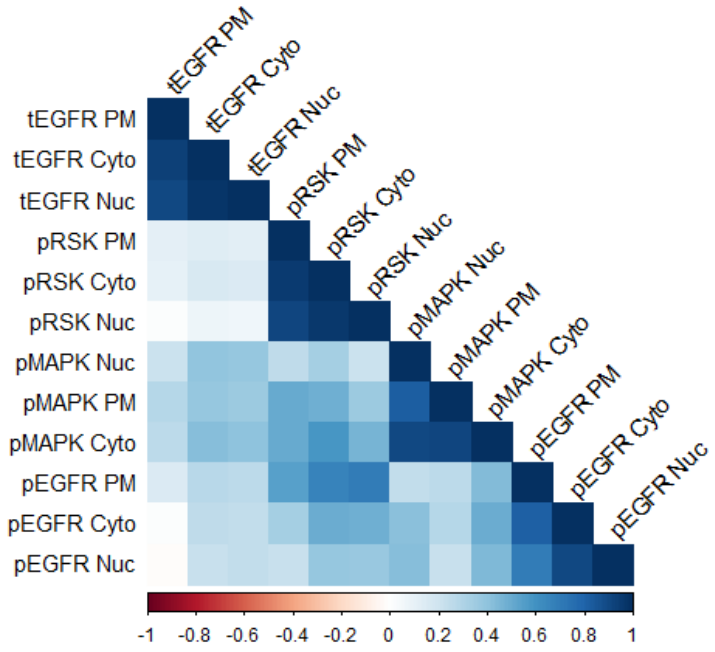

Figure 6.20: Visualisation of the correlation matrix as a heatmap of the Pearson correlation coefficients (see Definition 12 of Chapter 2) between each pair of variables in the data, for both cell lines, all inhibitor types, all concentrations and all time points combined.

From the correlation circle, some clear groupings between the variables can be seen, where the three cellular compartments for each protein tend to be close together, indicating some redundancy in the data between cellular compartments. tEGFR and pRSK are the best represented variables by the PCs 1 and 2, which is unsurprising since these are the proteins on which the inhibitors seem to have the largest effect. This is confirmed by Figure 6.23 which shows pairwise scatter plots of the data for each protein, combined for cell line, inhibitor type, concentration of inhibitor, time point and cellular compartment. Particularly for tEGFR, the range of the normalised data is much larger than the range of the data for the other proteins. Figure 6.22 also indicates a difference, along PC 2, between tEGFR and the other proteins. Also interesting to note is how much each original variable contributes to each of the PCs. The contribution is given as a percentage, such that each PC is fully explained by the sum of the contributions from each original variable. A heatmap of the percentage contribution to each of the first four PCs

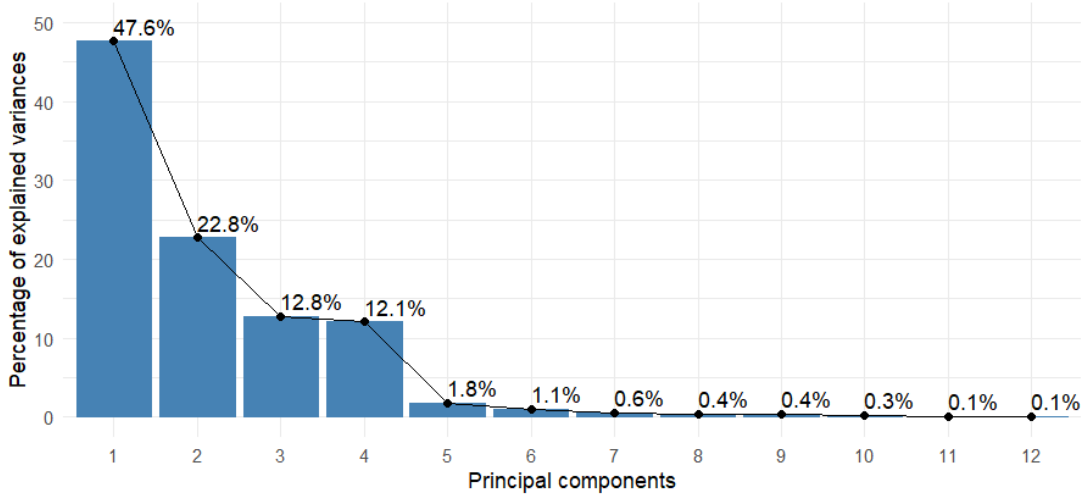

Figure 6.21: Scree plot of the percentage contribution of each of the 12 principal components, computed through the PCA, to the total variance in the data.

by the original variables is given in Figure 6.24. One can see that all of the pEGFR, pMAPK and pRSK variables contribute roughly equally to the first PC, and the tEGFR variables contribute less. On the other hand, the tEGFR variables contribute the most to the second PC with some contribution coming also from pRSK. The lower quality of representation seen in Figure 6.22 for pEGFR and pMAPK can be explained since these variables contribute more to the third and fourth PCs, respectively.

As well as plotting the *variables* on the correlation circle, one can also plot the individual data points as mapped to the first two PCs. In such a plot, the data points can be coloured according to any groupings in the data, and hence here the points are coloured, in separate plots, to show the cell line, the time point, the inhibitor type and the concentration of inhibitor. Figure 6.25 shows the variables and data points grouped by cell line (left hand side) and time point (right hand side), in terms of the first two PCs. From this figure, clear differences can be seen between the two cell lines, and the 5 time points. Figure 6.26 is a similar figure but where the groupings are the inhibitor type (left hand side) and the inhibitor concentration (right hand side). Here, any differences between these groups are much less defined. It should be noted that, according to the quality of

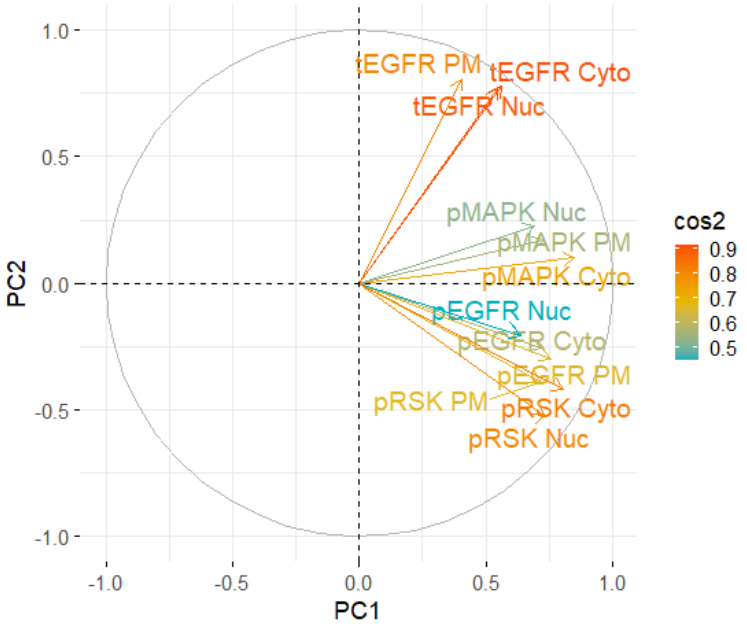

Figure 6.22: Correlation circle between the original variables and the first two principal components identified by the PCA. The colour of a coordinate arrow represents the quality of representation of that variable in terms of the cos2. The higher the cos2, the better that variable is represented by the PCs.

representation in Figure 6.22, the groupings in Figures 6.25 and 6.26 are defined mostly by the original variables tEGFR and pRSK and less so by pEGFR and pMAPK. These figures give an overall picture of where the variability in the whole dataset is the greatest, and one can conclude that, in line with the ANOVA results from Section 6.2.1, the inhibitors have differing effects on the proteins depending on the cell line. From Figure 6.26 however, it can be seen that as an overall effect, the inhibitor type is not very significant.

6.2.4 Analysis of concentration and cellular compartment

In the analysis so far in this chapter, the inhibitor concentration and cellular compartment of the proteins of interest have not been explicitly discussed. In the PCA it was found that there appeared to be some redundancy in the data between the cellular compartments, *i.e.* the cellular compartment of a protein does not seem

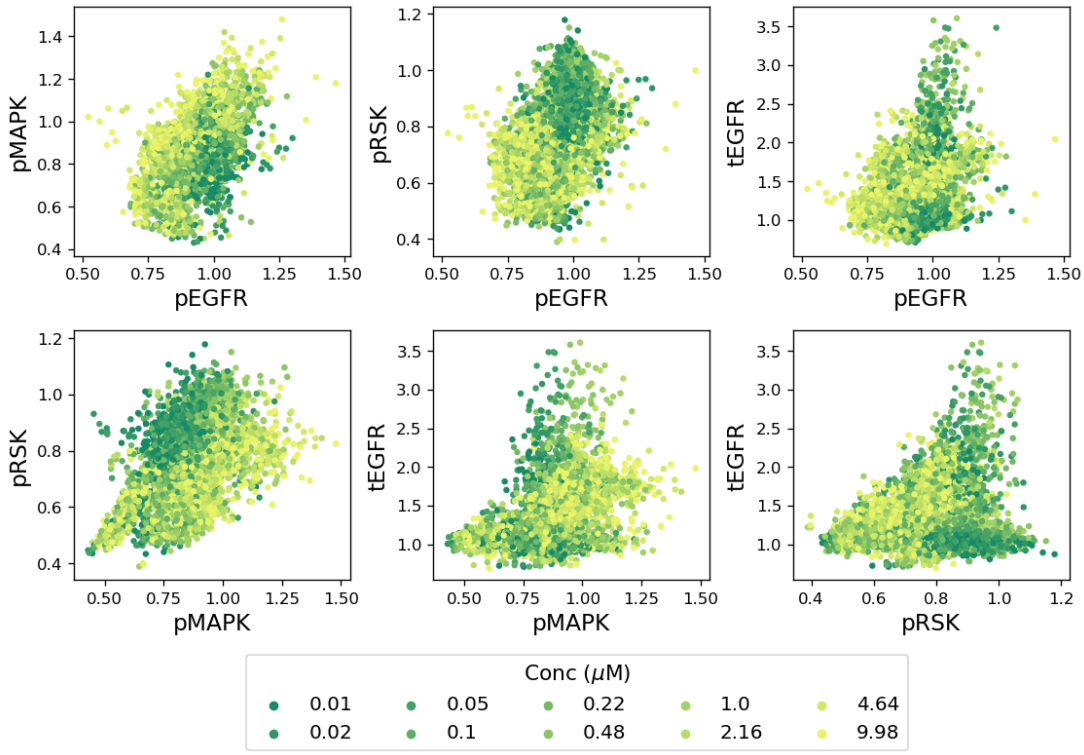

Figure 6.23: Scatter plots of the MFI data for each pair of proteins. The data here is combined for both cell lines, all inhibitor types and concentrations, all time points and all cellular compartments. The points are coloured by inhibitor concentration as indicated by the legend.

to have a great effect on the abundance of the protein. Here one-way ANOVA is used to assess this hypothesis further. In particular, a one-way ANOVA is carried out for each combination of cell line, inhibitor type, inhibitor concentration, time point and protein, with cellular compartment as the independent variable and MFI as the dependent variable, giving a total of $2 \times 8 \times 10 \times 5 \times 4 = 3200$ analyses, where now all 10 inhibitor concentrations are considered. As a specific example, one of the 3200 analyses would use the MFI data for pEGFR in the WT cell line, for inhibitor D1 at a concentration of 0.01 μM and at time point 2 hours. The ANOVA would then test whether there are any statistically significant differences between the means of the groups of the data (where the mean is taken over the experimental replicates) defined by cellular compartment. Similarly to in Section 6.2.2, if cellular compartment is found to be significant at the

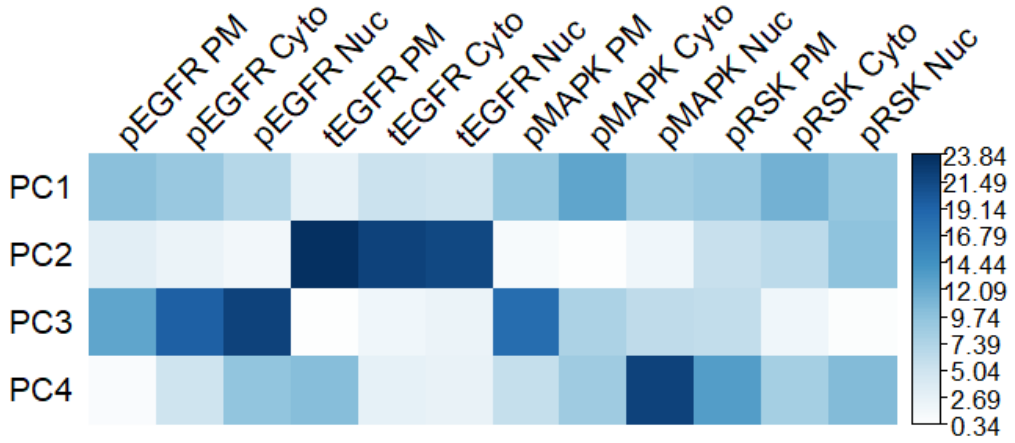

Figure 6.24: Visualisation of the percentage contributions from each original variable in the dataset (columns) to each of the first four PCs (rows), plotted as a heatmap where the colour of a pixel indicates the percentage as given by the colour bar.

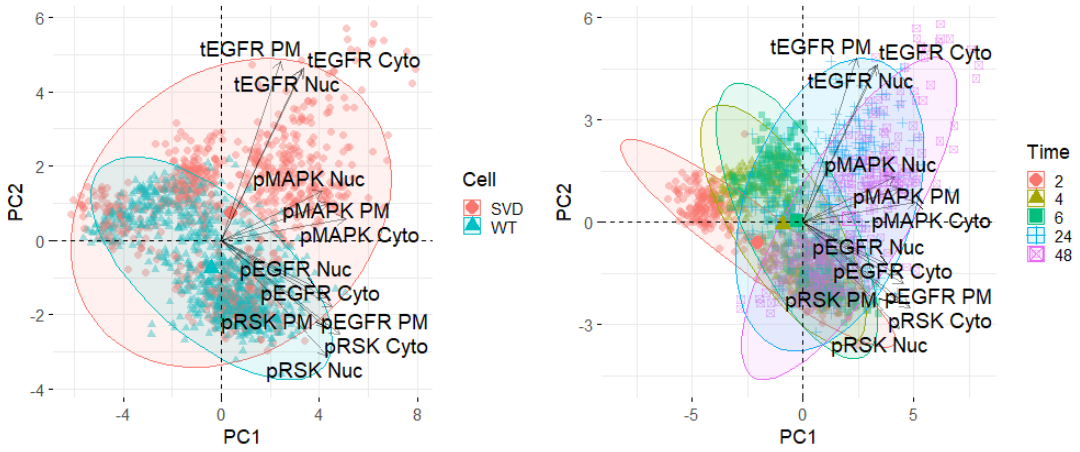

Figure 6.25: Correlation circles between the original variables and the first two principal components (on the x-axis and y-axis, respectively) identified by the PCA. Scatter plots of the data points are overlayed, coloured by cell line (**left hand side**) and time point (**right hand side**).

5% level (*i.e.* there is a statistically significant difference between the means of the data for at least one pair of cellular compartments), Tukey's HSD test is used to determine which pairs of cellular compartments are statistically significantly

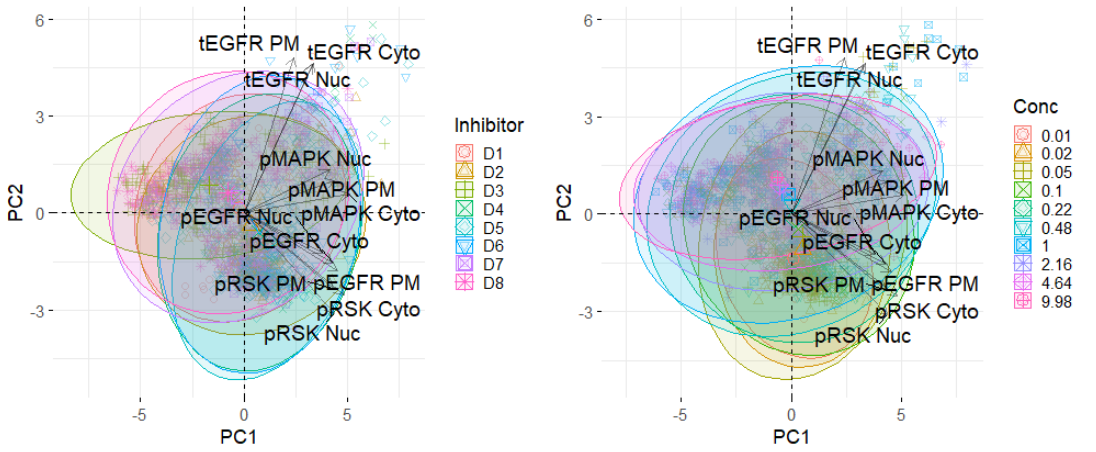

Figure 6.26: Correlation circles between the original variables and the first two principal components (on the x-axis and y-axis, respectively) identified by the PCA. Scatter plots of the data points are overlayed, coloured by inhibitor type (**left hand side**) and inhibitor concentration (**right hand side**).

different from one another. In total, out of the 3200 ANOVA tests, cellular compartment was only a significant effect for 699 combinations of the other variables, which is roughly 22% of the time. This result agrees with the PCA result, that the abundance of each protein is often approximately equally distributed between the three cellular compartments. Figure 6.27 (left hand side) shows a heatmap of the frequency of statistically significant differences between pairs of cellular compartments as a result of Tukey's HSD test at the 5% level. From the figure one can determine that when cellular compartment *is* a significant effect, this is due, most often, to differences between the protein abundances at the PM and the nucleus. Protein abundances in the cytoplasm and nucleus are rarely significantly different from one another. There could be some experimental error contributing to this apparent homogeneity between cellular compartments, for example it may be difficult to determine, from the FI, if a protein is in the cytoplasm or in the nucleus.

A similar analysis can be employed to study the effect of inhibitor concentration and to determine which concentrations are most often statistically significantly different from one another. Here the one-way ANOVA is carried out for each combination of cell line, inhibitor type, time point, protein and cellu-

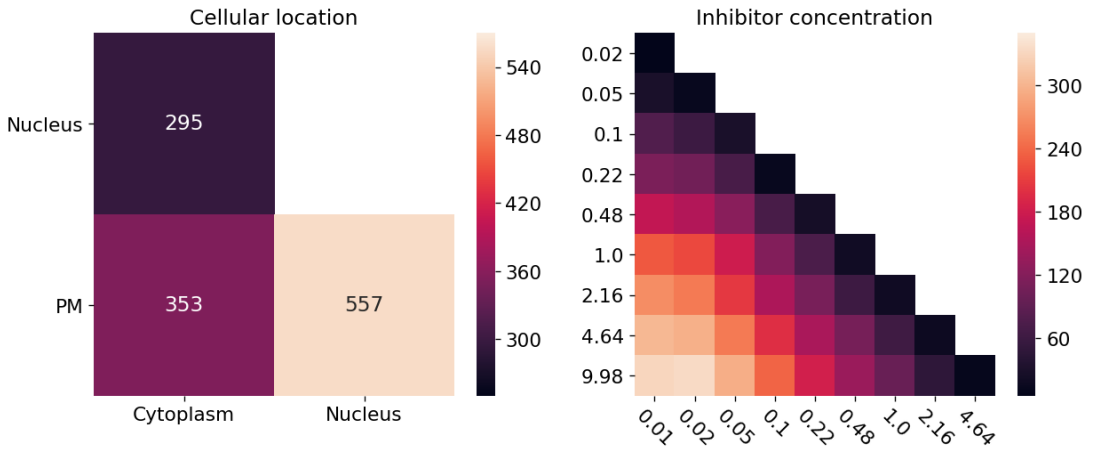

Figure 6.27: Heatmaps of the frequency of statistically significant differences between cellular compartment pairings (**left hand side**) and inhibitor concentrations (**right hand side**) as a result of Tukey's HSD test at the 5% level, applied after a one-way ANOVA with cellular compartment or inhibitor concentration as the independent variable (main effect).

lar compartment, with inhibitor concentration as the independent variable and MFI as the dependent variable, giving a total of $2 \times 8 \times 5 \times 4 \times 3 = 960$ analyses. Again, Tukey's HSD test is then used at the 5% level to determine pairs of inhibitor concentrations which are statistically significantly different from one another. Out of the 960 ANOVA tests, inhibitor concentration was a significant effect 56% of the time. This is a relatively low value, considering that one would expect that with a higher concentration of inhibitor there should be a greater effect on the protein levels. However in 44% of cases, there is no significant difference between even the highest and lowest concentrations, which are 10^2 μM and 10^{-2} μM, respectively. It can be observed from figures such as Figure 6.18, that all concentrations of inhibitor D3 have a very similar effect on the proteins, however this concentration clustering must also be present in the effect of other inhibitor types on the proteins. In the 56% of cases where there *is* a statistically significant difference between at least one pair of concentrations, the frequency of these differences for each pair is plotted as a heatmap in Figure 6.27 (right hand side), summed over all combinations of the other variables. The result is as would be expected, whereby the largest number of statistically significant differ-

ences occur between the concentrations with the largest difference, for example, 10^2 μM and 10^{-2} μM. There are very few statistically significant differences between consecutive pairs of inhibitor concentrations. This result validates the use of only four inhibitor concentrations, equally spaced across the logarithmic range of concentration values, for the two-way ANOVA used in Section 6.2.1.

6.2.5 Summary of the statistical analysis

In summary, from the statistical analysis carried out in this chapter, the cell line appears to be significant on the effect of the TKIs on the abundance of the proteins of interest. In general, the TKIs have a larger effect on the SVD cell line, in particular on the up-regulation of EGFR (tEGFR increase with respect to baseline). In terms of the efficacy of the TKIs, this is not a desirable result, since the aim of the TKIs is to inhibit the MAPK pathway, and an up-regulation of the EGFR would serve to further initiate the pathway. There are some differences in the response to the TKI depending on the inhibitor type also, particularly at low concentrations, where D3 seems to have a much larger effect than the other inhibitors. The other inhibitors reach similar levels of inhibition as D3 for higher concentrations. A general trend in both cell lines for most inhibitor types and concentrations, is that initially they have the expected effect on the proteins, in terms of up- or down-regulation with respect to the control, but by the latest time points, the protein abundances are returning to the level of the control, or in some cases even moving in the opposite direction as would be expected. This implies that upon treatment of the cells with the TKIs, some feedback mechanisms are coming into place, for example, up-regulating the EGFR levels, which is seen in the data. Given that the TKIs bind irreversibly to EGFR, the MAPK pathway can be re-initiated by the cells increasing EGFR synthesis so much that all of the TKI is depleted. This is presumably what is happening in the data by time 48 hours. This problem could potentially be overcome by dosing with the inhibitors at a high enough concentration that any newly synthesised EGFR is also rapidly inhibited, or by dosing at regular time intervals, such as once per day.

In order to determine, mechanistically, why there are some differences in the effect of the inhibitors between cell lines and inhibitor types, particularly at low

concentrations, one could use a mathematical modelling approach, similar to those employed in Chapters 4 and 5 of this thesis. In particular, some distinct groupings of data could be used, in conjunction with a mathematical model to determine, for example, how the parameters of such a model differ when the model is calibrated using the different datasets. This approach was taken by Claas *et al.* (2018b) who used a mathematical model of RTK trafficking to determine how the trafficking was "reprogrammed" upon inhibition of the MAPK pathway. In this paper, the authors do not explicitly include the inhibitors in the model, however this approach could also be taken and one could aim to identify parameters such as the binding rate of the inhibitor to the receptor, which might elucidate mechanistically how the inhibitor types and cell lines differ from one another.

In order to use such calibration methods with the data presented in this chapter, it would be necessary to firstly understand, and be able to model, the control (DMSO) data. This is since the data is normalised, in Section 6.1.1, by dividing the inhibitor dosed data by the DMSO data, and so the model should be normalised in the same way in order to be comparable to the data. However, when examining the DMSO data, there were some clear trends which could not be explained. For example, the DMSO data for the SVD cell line is plotted in Figure 6.28 and it can be seen that for the proteins, pEGFR, tEGFR and pRSK, there is an initial rise in fluorescence (proportional to copy number), followed by a peak and then a decline. These dynamics are typical of growth factor stimulated cells (Kholodenko *et al.*, 1999; Schoeberl *et al.*, 2002; Shankaran *et al.*, 2008), however for the data here, the cells are not treated explicitly with growth factor. Even though it is likely that there is some growth factor present in the serum that the cells were grown in, the dynamics of the protein phosphorylation seen in the data do not match with other examples from the literature, where the peak here is of the order hours, and in several examples from the literature, it is of order minutes or even seconds (Kholodenko *et al.*, 1999; Shankaran *et al.*, 2008). The DMSO profile for pMAPK also does not match with the theory of growth factor stimulation, since here the trend in the data is a decline followed by an incline, which would not be expected. For these reasons, as well as the overall lack of significant differences in the effect of the inhibitor types on the proteins of interest, no further mathematical modelling was carried out using the data

here. Instead, in Section 6.3, a short review of the current literature surrounding mathematical modelling of the interaction between RTKs and TKIs is given.

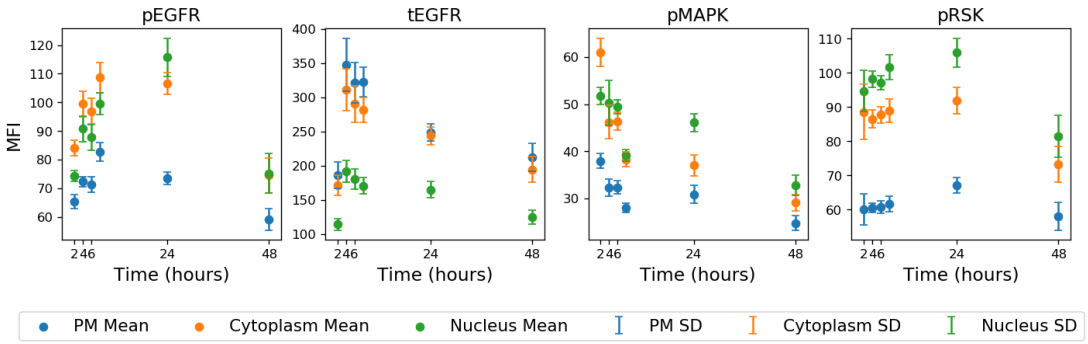

Figure 6.28: Plots of the control (DMSO) MFI data in the SVD cell line, for each of the four proteins and three cellular compartments.

6.3 Review of RTK inhibition modelling

First generation, reversible binding EGFR inhibitors such as Gefitinib and Erlotinib have been in use to treat patients with NSCLC worldwide since the mid to late 2000s. More recently, third generation, irreversible EGFR inhibitors have been developed, and there is an ongoing focus to develop new, more effective versions of such inhibitors. Osimertinib is one such example of a third generation TKI, developed by AstraZeneca, which was approved for use by the FDA in 2015 (Al-Quteimat & Amer, 2020). Given that all of these inhibitors have been developed relatively recently, there is not an extensive amount of mathematical modelling of the interaction between inhibitors and receptors in the literature. Here however, two general classes of mathematical modelling from the literature will be discussed, enzyme kinetic modelling and combination therapy modelling.

Enzyme kinetics, which is the study of enzyme catalysed reactions, focussing on the rate of product formation is commonly used in the development of new drugs. A mathematical formalism known as Michaelis-Menten kinetics was first introduced by Michaelis and Menten in their 1913 paper, "Michaelis, L., and Menten, M. L. (1913) *Die Kinetik der Invertinwirkung.* Biochem. Z. 49, 333–369", which has been translated from German to English by Johnson & Goody (2011).

An overview of Michaelis-Menten kinetics is as follows, based on Johnson & Goody (2011) and Rogers & Gibon (2009). When an enzyme and a substrate come together, the enzyme catalyses the reaction, where the substrate becomes the product and traditionally the enzyme is unchanged by the reaction, however the kinetics can also be used where the substrate and enzyme together form a product. The reactions underlying the traditional approach are,

$$E + S \underset{k_r}{\overset{k_f}{\rightleftharpoons}} ES \xrightarrow{k_{cat}} E + P,$$

where E denotes the enzyme, S the substrate and P the product. There is an initial reversible binding of the substrate to the enzyme, followed by an irreversible formation of the product, with rate k_{cat}. The aim of Michaelis-Menten kinetics is to find an equation for the rate of formation of the product P. Under the assumption of mass action kinetics, one can write the following four ODEs,

$$\begin{aligned}
\frac{d[E]}{dt} &= -k_f[E][S] + k_r[ES] + k_{cat}[ES], \\
\frac{d[S]}{dt} &= -k_f[E][S] + k_r[ES], \\
\frac{d[ES]}{dt} &= k_f[E][S] - k_r[ES] - k_{cat}[ES], \text{ and} \\
\frac{d[P]}{dt} &= k_{cat}[ES],
\end{aligned} \tag{6.3}$$

to describe the time dynamics of the concentrations of each of the four species. Given that the system is closed, and there is no degradation or synthesis of any species, the total concentration of enzyme is conserved, and hence $[E] + [ES] = [E]_0$, where $[E]_0$ is the initial concentration of enzyme. An approximation known as the "Quasi-steady-state approximation" is then made, which says that the concentration of the intermediate complex, ES, remains constant for a considerable period of time. This approximation is valid under the assumption that S is large relative to E and that k_{cat} is small relative to k_f and k_r, and hence the intermediate complex ES is formed rapidly and remains at an approximately constant value for a long period of time. It is within this quasi-steady-state period of time that Michaelis-Menten kinetics applies. Under the quasi-steady-state

approximation, one can write that

$$\frac{d[ES]}{dt} = 0,$$

and hence

$$\begin{aligned} k_f[E][S] &= k_r[ES] + k_{cat}[ES] \\ &= (k_r + k_{cat})[ES]. \end{aligned} \tag{6.4}$$

Combining Equation (6.4) with the expression for total enzyme, gives

$$k_f([E]_0 - [ES])[S] = (k_r + k_{cat})[ES].$$

Rearranging and simplifying yields

$$[ES] = \frac{[E]_0[S]}{k_M + [S]},$$

where

$$k_M = \frac{k_r + k_{cat}}{k_f},$$

and is known as the Michaelis constant. Finally, a formula for the rate of product formation, v, can be obtained by substituting the expression for $[ES]$ into the differential Equation (6.3), where

$$\begin{aligned} \frac{d[P]}{dt} = v &= k_{cat}[ES] \\ &= \frac{k_{cat}[E]_0[S]}{k_M + [S]} \\ &= \frac{V_{max}[S]}{k_M + [S]}, \end{aligned} \tag{6.5}$$

with $V_{max} = k_{cat}[E]_0$, *i.e.* the fastest rate of product formation which happens at a saturating substrate concentration. Equation (6.5) is useful when compared to data from enzyme assays in which the initial rate of reaction is measured for varying substrate concentration (see Figure 6.29). Methods such as nonlinear

regression can then be used to fit the kinetic parameters k_{cat}, V_{max} and k_M with the data. Provided that k_{cat} is small relative to k_r and k_f, the Michaelis constant k_M can be thought of as a proxy to the dissociation constant for a given enzyme, where lower k_M values indicate higher affinity of the enzyme for the substrate. Equation (6.5) is widely used, and in particular, Zhai *et al.* (2020), Schwartz *et al.* (2014) and Yates *et al.* (2016) have all used this equation and variations of the equation to compare inhibitors for EGFR.

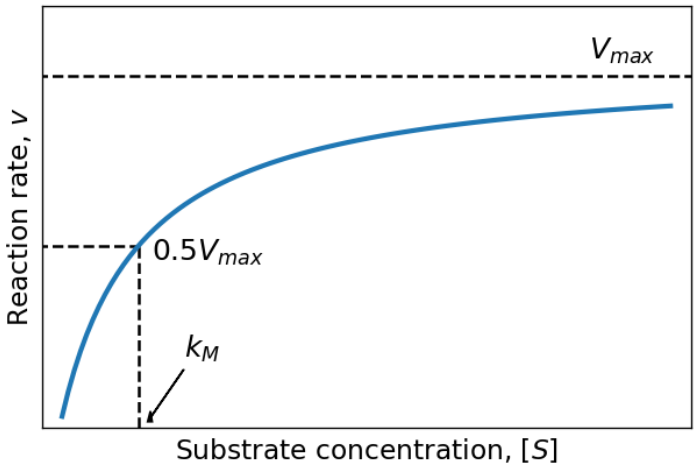

Figure 6.29: A plot of substrate concentration, $[S]$, against the product formation rate, v, under Michaelis-Menten kinetics, where the parameters of Equation (6.5) are annotated (inspired by a similar figure from Rogers & Gibon (2009)).

A recent paper by Zhai *et al.* (2020) from AstraZeneca, combines experimental fluorescence imaging with mathematical modelling to examine the therapeutic selectivity of Osimertinib. The paper employs a Michaelis-Menten enzyme kinetic mathematical model and calibrates the mathematical model for different EGFR constructs; WT, L858R mutant and L858R/T790M mutant. The kinetic mechanism of action of Osimertinib is explored in relation to the parameters of the mathematical model, since it is seen in the data that Osimertinib inhibits the double mutant and single mutant to a greater extent than the WT. The authors determine values for the binding and inactivation parameters in a two step binding model, and show that Osimertinib both binds faster to, and inactivates faster, the mutant proteins than the WT. Both initial binding rates and steady

state kinetic parameters are identified by combining the model and data. A similar approach was taken by Schwartz *et al.* (2014), in an earlier paper, in which an ODE model is used to define inhibitor potency in terms of the contributions from reversible binding and unbinding reactions and the irreversible inactivation reaction. The authors here consider third-generation EGFR inhibitors such as Dacomitinib and Afatinib, developed prior to Osimertinib.

Another recent paper from AstraZeneca, by Yates *et al.* (2016) explores both the pharmacokinetic (PK) and pharmacodynamic (PD) properties of Osimertinib and how they relate to a reduction in tumour growth, using a mouse xenograft model. The PK model is necessary in the mouse model since Osimertinib has been found to be metabolised into an active metabolite in mice. The PD model involved the reversible binding of the drug to pEGFR and the further irreversible inactivation of the protein. The model was calibrated using pEGFR data over a time course of several days. A further step of tumour growth was added to the model, where the volume of a tumour was modelled, such that the tumour grows with a rate proportional to the concentration of pEGFR. The modelling explored different dosing regimes, such as daily dosing and intermittent dosing. Experimentally it was found that pEGFR returns to baseline in about 72 hours following a single oral dose of Osimertinib. After dosing once daily for 14 days however, tumour growth did not restart for 2 weeks. This was captured in the mathematical model by considering a pEGFR "pool" where after 1 dose only some of the pEGFR is inhibited but after 2 weeks of dosing much more inhibition occurs and the pEGFR is regenerated slowly. The model fitted well to experimental data and again, biologically relevant parameters were determined. Studies such as those described here allow for insights into how well certain drugs inhibit the desired proteins, and how different drugs, and different cell lines, can be compared with one another.

A second class of mathematical models, prevalent in the literature for EGFR inhibition, concern "combination therapy". This is a term used to describe the treatment of a patient with more than one type of drug, simultaneously. The approach has been shown to be effective since it targets multiple pathways synergistically or additively, and has the added benefit of preventing drug resistance (Mokhtari *et al.*, 2017). An early paper on the topic of combination therapy comes

from Araujo *et al.* (2005) and is a mathematical modelling study which extends the well known EGFR signalling model developed by Kholodenko *et al.* (1999). The model by Kholodenko *et al.* (1999) involves all of the early steps in EGFR signalling, such as ligand binding and dimerisation of the receptor, and some intracellular protein phosphorylation reactions. Araujo *et al.* (2005) build upon this model by including the inhibition of EGFR alone, as well as the combined inhibition of EGFR and further downstream proteins, Shc and Plcγ. The inhibition is modelled implicitly, by involving a pre-multiplier in each of the forward phosphorylation reactions for the receptor, Shc and Plcγ, such that the pre-multiplier takes a positive value less than 1. The lower the value of the pre-multiplier, the greater the effect of the inhibitor. Through simulation of the mathematical model at varying values of each of the three pre-multipliers, the authors firstly find that although inhibition of the receptor alone has a large effect on the down-regulation of the phosphorylated receptor, it is only when the pre-multiplier takes values less than 0.5 that a significant effect on the down-regulation of downstream proteins in the pathway is observed. When multiple proteins in the pathway are inhibited (the receptor and at least one downstream protein), the signal attenuation is much greater. The authors also discuss "additive" and "synergistic" effects of inhibition when using multiple inhibitors simultaneously. An additive effect corresponds to the effect which would be expected if one added together the effects of using the multiple inhibitors independently, and a synergistic effect implies that the effect caused by using multiple inhibitors simultaneously is greater than the additive effect. They compare this hypothetical simulation with the simulation upon inhibiting both proteins simultaneously and find that the effect is actually synergistic (super-additive), *i.e.* the signal attenuation is greater than the effect of the sum of the individual attenuations. This theoretical mathematical work clearly indicates that combination therapy could be useful in the inhibition of EGFR initiated signalling, however further experimental work would be required to develop such downstream inhibitors and then to check the levels of toxicity when such inhibitors are used in combination.

A similar, more recent paper on the mathematical modelling of combination therapies comes from Huang *et al.* (2017) who, similarly to Araujo *et al.* (2005), extend upon a mathematical model taken from the literature, where here they

choose the mathematical model by Hornberg *et al.* (2005). This model is again a large network model involving close to 150 reactions, but Huang *et al.* (2017) choose to study only a subset of the reactions, yielding the pathway seen in Figure 6.30. This work extends the ODE model presented by Hornberg *et al.* (2005) to include three potential inhibitors, for each of EGFR, BRaf and MEK. The inhibitors are this time included explicitly in the model, where they bind to, and inactivate, the target proteins, indicated by a "D" in a coloured box in Figure 6.30. The model is validated by comparing simulations with previous experimental data from the literature. The authors use a value known as the "combination index" to determine the efficacy of different combinations of inhibitors. The authors state that the combination index (CI) can be computed as follows,

$$\mathrm{CI} = \frac{(D)_1}{(Dx)_1} + \frac{(D)_2}{(Dx)_2},$$

where $(D)_1$ and $(D)_2$ are the *combination* doses of the two drugs (1 and 2) which yield 50% efficacy and $(Dx)_1$ and $(Dx)_2$ are the *single* doses for each drug that yield the same effect. The value of the combination index then implies how effective the drug pairing is, where $\mathrm{CI} < 1$ implies the combination is synergistic, $\mathrm{CI} = 1$ implies additivity and $\mathrm{CI} > 1$ implies antagonism. Huang *et al.* (2017) are also able to validate their model by comparing the calculated CIs with reported CIs from the literature and thus propose that the model could be used to predict CIs for new drug combinations.

Another paper which focuses on combination therapy and this time includes both mathematical and experimental results, is by Misale *et al.* (2015). The authors use a colorectal cancer (CRC) patient derived cell line to study the effects of different drug combinations, specifically those for EGFR and MEK. Unlike in the previous papers, and in the data used in this chapter, the EGFR inhibitors here are not TKIs. Instead they are EGFR antibodies, such as Cetuximab and Panitumumab, which, rather than inhibiting EGFR phosphorylation, work further upstream and compete with EGF for binding to the ligand-binding domain of EGFR. When the receptor is bound to an inhibitor, dimerisation and subsequent activation of the receptor does not occur. Misale *et al.* (2015) tested the time to progression (TTP) of the disease (usually a clinical parameter) *in vitro* by defining

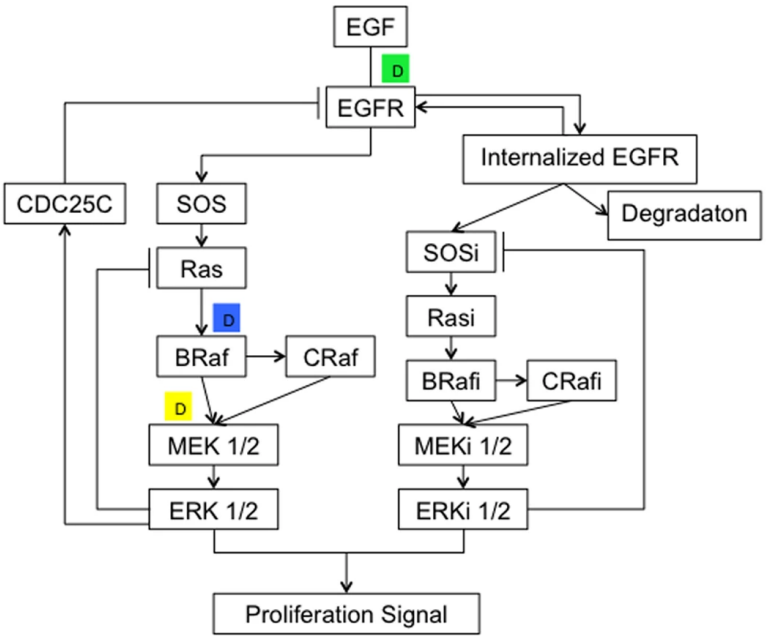

Figure 6.30: A figure of the reactions underlying the mathematical model by Huang *et al.* (2017), where letter "D" in a coloured box indicates a reaction which is targeted with an inhibitor in the model. Figure taken from Huang *et al.* (2017).

it as the time until cells treated with an inhibitor begin to divide exponentially, and hence have become resistant to the drug. Upon treatment with Cetuximab (an EGFR antibody) alone, it was found that the resistance defined by the TTP occurred within 80 to 120 days post treatment initiation. However, when the cells were treated with a combination of inhibitors for EGFR and MEK, no drug resistance occurred up to 6 months post treatment initiation. The authors of the paper assume that drug resistance may occur due to a population of cells which are drug resistant *pre* treatment initiation. A mathematical model is used to test this hypothesis, where the output of the model is the volume of a tumour over time and is denoted $V(t)$. The model assumes independent exponential growth of two populations of cells, sensitive cells with initial volume a and resistant cells with initial volume c. The sensitive cells have decline rate $b < 0$ and the resistant cells have growth rate $d > 0$, so that the model is

$$V(t) = a\exp(bt) + c\exp(dt).$$

The model was fitted to experimental data in which CRC cells were treated with Cetuximab and the parameters were estimated such that the model gave a good fit to the data. It was estimated that around 2% of the cells in the experiment were resistant to Cetuximab at the start of treatment. The authors then went on to fit a more complex exponential growth model to data in which there were phases of treatment and absence of treatment, and the parameters were again estimated. To model the combination therapy, since no resistance was observed in the data over a time period of 6 months, the model included only the sensitive cells, hence

$$V(t) = a\exp(bt),$$

and again the model was able to fit the data well.

A fourth paper relating to combination therapies for treating signalling pathways is by Klinger *et al.* (2013), and the authors here come to a similar conclusion as others, that a combination therapy of EGFR and MEK inhibitors could be a good candidate for blocking cell signalling. Similarly to the work by Misale *et al.* (2015), this paper combines both experimental and mathematical results in CRC cell lines, here using a network based model. Six CRC cell lines were treated with combinations of inhibitors and growth factors and the phosphorylation of signalling molecules was quantified. This data was used to calibrate parameters of mathematical models which were then simulated under different theoretical scenarios to predict effects of drug combinations.

The papers reviewed here describe two main classes of EGFR inhibition modelling; enzyme kinetics modelling and combination therapy modelling. Given that EGFR TKIs and other MAPK inhibitors are relatively new and promising cancer therapies, there is still scope for more mathematical modelling in this area.

6.4 Discussion

In this chapter, EGFR inhibition has been explored, specifically tyrosine kinase inhibition, which is a current treatment for many different types of cancer. The role of the inhibitors is to inactivate the signalling receptor, by blocking the

phosphorylation of downstream proteins in the MAPK pathway (amongst others), which leads to less protein translation and ultimately less cell division, hence a reduction in tumour growth. Ideally, such an inhibitor would preferentially inhibit mutant varieties of EGFR, so that only the cancerous cells are treated, and healthy cells are left to function normally. Given the emergence of drug resistance when patients are treated with first-generation reversible EGFR inhibitors, in this chapter, third-generation irreversible inhibitors are discussed and analysed. There is still an ongoing need for the development of more inhibitors for EGFR, other RTKs and further downstream proteins.

The data used in this chapter are the fluorescence intensity of antibodies for four proteins of interest in the MAPK pathway, pEGFR, tEGFR, pMAPK and pRSK, when the cells are treated with one of eight different EGFR TKIs, being used in preclinical studies at AstraZeneca. The experiments were carried out in both a WT EGFR cell line and a mutant (SVD) EGFR cell line. In Section 6.1, the experimental data is introduced, presented and normalised, and any data points considered outliers are removed prior to further analysis. Then in Section 6.2, a thorough statistical analysis of the data is employed, whereby ANOVA and post-hoc analyses are used to find statistically significant differences between groupings of the data, be it by cell line, or inhibitor type. In Section 6.2.3, a principal component analysis is carried out to give a more general overview of the data, identify any redundancies in the data and confirm any groupings identified in the previous analyses.

The results of the ANOVA indicated that for several combinations of protein, cellular compartment and time point, the MFI was affected differentially depending on the cell line. In particular, from analysis of the data, it can be seen that upon treatment with the inhibitors, there is a general increase in tEGFR over the first 24 hours, where this up-regulation is more pronounced in the SVD cell line. This is indicative of a feedback effect coming into place, whereby the cells try to return to their levels of signalling prior to inhibitor addition by up-regulating the signalling protein. There is not a very noticeable difference in the phosphorylation levels of the other three proteins, between the two cell lines however. This is not an encouraging result given that an inhibitor to be used to treat patients should preferentially bind mutant EGFR over the WT. In general however, from

the experimental data, the effects of the inhibitors on both cell lines and all proteins are relatively small. The largest effect is an approximately 3.5 fold increase in tEGFR at the PM in the SVD cell line under some concentrations of some inhibitors, however most effects are less than a 2 fold change from baseline (see Appendix E). The inhibitor type was also often significant in the change in the levels of the proteins of interest, and the pairs of inhibitors which contributed to this significance were found using Tukey's HSD test. It was found that there were some differences between inhibitors at the lowest inhibitor concentrations, where D3 had a noticeably larger effect on the up-regulation of tEGFR and down-regulation of pRSK than the other inhibitor types. Other inhibitors then begin to have a similar effect to D3 as the concentration increases, and, by the largest concentration of inhibitor, all inhibitor types have a very similar effect. It should be noted that the inhibitors had the largest effect on the proteins tEGFR and pRSK and that significantly different groupings in the data for pEGFR and pMAPK were less common.

It is clear from plots of the data, as well as the results of the ANOVA that there is some level of redundancy in the data. This is since, data for a single protein is very similar across cellular compartments. From the ANOVA (see Figure 6.5) it can be seen that if there is a significant main effect for a protein and time point at a specific cellular compartment, it is often also significant for the other two cellular compartments. From figures of the data (comparing columns of the figures in Appendix E) this can also be seen, where the data for one protein and cellular compartment looks very similar, in terms of trend and value, to the data for the same protein in different cellular compartments. This redundancy was further confirmed in Section 6.2.3, where a correlation matrix of the data for both cell lines, all inhibitor types, concentrations and time points is plotted in Figure 6.20. There is a very strong positive correlation between each pair of the three cellular compartments for each of the four proteins individually. The results of the principal component analysis for the whole dataset also showed this result, where the coordinates of the original variables on the first two principal components are grouped by protein type in Figure 6.22, *i.e.* the cellular compartments are clustered together for a single protein. The PCA also corroborates the ANOVA result, that cell line is reasonably significant within the data, since in Figure 6.25

it can be seen that the data mapped to the first two PCs forms well defined groups per cell line. The inhibitor type, however, does not appear so significant when the data is pooled, seen in Figure 6.26.

Had the statistical analysis have revealed specific inhibitor types which had a significantly different effect from other types on all proteins, and cellular compartments, then it would have been interesting to study these differences further using a mechanistic mathematical model. Given that the data does not show many very large effects due to the inhibitors, and that the inhibitor data is very similar between different types, further modelling has not been carried out in this chapter. In order to use the data for modelling it would be beneficial to see larger effects and have data with less redundancy between cellular compartments. The control data, where the cells were treated with DMSO alone, also shows trends which are currently unexplained, which also limits the data for further analysis. Using this *type* of data, one can imagine constructing an ODE based trafficking model, where each of the proteins are trafficked with some rate between the cellular compartments upon addition of inhibitor. Inhibitor binding and inactivation of the proteins could be added as reactions in the model and different feedback loops could be explored. The model parameters could then be calibrated using experimental data of the type discussed in this chapter, with the aim of revealing specific reactions which differ in rate between inhibitor types. This sort of mathematical analysis could help to select inhibitor types for further development. The statistical analysis carried out in this chapter is an example of an exploratory analysis to find initial differences between groupings and to test for the efficacy of inhibitors in terms of their effect on different EGFR constructs. The work done here can also pick out any redundancies in the data so that only the most relevant and different groupings are carried through to a mathematical model.

Finally, in Section 6.3 of this chapter, a review of the current literature surrounding mathematical modelling of EGFR inhibition is given. Two main topics are discussed; enzyme kinetics and combination therapy. Enzyme kinetics, under the Michaelis-Menten formulation, is often used to compare different inhibitors for the same protein, by estimating parameters of the model using experimental data. The parameters indicate how well an inhibitor binds to a protein and to what extent the inhibitor inactivates the protein. Combination therapy is an

emerging method of treatment in which EGFR (or another RTK) is inhibited simultaneously with another protein, further downstream in the signalling pathway. Some of these combination therapies have been shown, mostly *in vitro* or theoretically through the use of a mathematical model, to have a greater effect on signal attenuation and to minimise the chance of drug resistance. There is still a lot of clinical work to be done in this area. In general, this chapter has shown how statistical analyses and mathematical modelling can be used to aid decisions in the field of small molecule therapies for RTKs.

Chapter 7

Concluding remarks

In this thesis, mathematical and statistical techniques have been used to study intracellular biological processes, specifically signalling pathways at the head of which is a receptor molecule. Stimulation of signalling pathways, by other extracellular proteins, can push a cell to a fate such as division, death or migration, which are vital processes in the maintenance of a healthy population of cells. These pathways are also important to understand from the perspective of treatment of human disease, since it is often found in diseases such as cancer and autoimmune disorders, that the signalling is dysregulated. In the context of cancer for example, this dysregulation can mean that cells are dividing faster and more often than is necessary, causing the formation of tumours. In Chapters 4 - 6 of this thesis therefore, different signalling pathways are analysed mathematically in order to better understand their role in human disease and to give insight into how signalling dysregulation can be controlled. Firstly, in Chapter 3 a more general stochastic model of the competition between multiple receptor types for a common ligand type is introduced, which can be applicable to many cell types and signalling pathways.

In Chapter 3, the primary reactions at the head of a general signalling pathway are explored in terms of a stochastic mathematical model. In particular, the reversible reactions between two receptor types which can bind with a common ligand type are modelled as a continuous-time Markov quasi-birth-and-death process. This model is purely theoretical and is not compared with experimental data, however a stochastic approach is used here in order to simulate biological

scenarios in which there are a low number of receptors and/or ligands per cell and hence random fluctuations in copy numbers are important. Two stochastic descriptors are analysed, both of which can be important in the eventual cellular fate which the signalling initiates, namely the steady state distribution of the process, and the time-scales of receptor-ligand complex formation of each type. A matrix analytic approach is firstly used to compute these statistics, however this method is limited by its computational time, which is related to the state space of the process, which increases with increasing number of molecules of each type. Approximate methods of computation of the two descriptors are therefore developed in order to save on computational time, which are based on the independence of the processes for each receptor type, when there is a large number of ligands in the system, compared with the numbers of receptors. This approach was found, through numerical comparison with the analytic method, to be very accurate in many biologically relevant parameter regimes. A limitation of the method however, is that the accuracy decreases when the competition between the two receptor types is very high, such as when there are few ligands compared with the number of each receptor type or when the affinity of one receptor to the ligand is great. Further work could be carried out based on the ideas discussed in this chapter, such as extending the methods to more stochastic descriptors, or employing the methods developed here to other scenarios outside of molecular biology.

After a general introduction to the first steps in a signalling pathway is given in Chapter 3, in the remaining chapters of this thesis, mathematical and statistical techniques are used in corroboration with experimental datasets focussed on specific receptor initiated processes. In Chapter 4, deterministic mathematical models are used to simulate the first steps in the JAK/STAT signalling pathway, which is initiated by the interaction between cytokine receptors, and their respective ligands, IL-6 and IL-27. In this situation a deterministic approach is reasonable given that there are large numbers of each receptor and ligand type per cell. Hypothesis testing is firstly carried out in this chapter, using Bayesian methodology, in order to decide between two possible ways in which receptor molecules could be internalised into the cell. The models are parametrised using Bayesian parameter inference, specifically ABC-SMC, along with experimental

data provided by Dr. Stephan Wilmes and Dr. Ignacio Moraga from the University of Dundee, in order to elucidate differences between the two pathways in terms of the underlying reactions and reaction rate constants. This parametrisation allowed to identify specific differences in the affinities of two types of STAT molecule to the different receptor types, providing an explanation for patterns seen in the experimental data. The parametrised model was validated using further datasets and could then be used to make predictions about the signalling responses in different biological regimes, specifically those seen in diseases such as Crohn's disease.

Chapter 5 is also centred around experimental data, here provided by Dr. Chi-Chuan Lin from the group of cellular and molecular biology at the University of Leeds. The data reveals an interesting assembly of proteins into liquid-liquid phase separated droplets, potentially comprised of ternary complexes, never before reported in the literature for the molecules used in these experiments. Three proteins, FGFR2, a receptor tyrosine kinase, Shp2 and Plcγ1 are seen to be co-expressed in the same areas of the cell and hence it was hypothesised that they form a ternary complex with one another. Phase separated droplet formation within a cell could lead to high density regions of signalling proteins, which may provide a mechanism for increased cell signalling. A mathematical model of the system was developed in order to explore this hypothesis, whereby four different ternary complexes were allowed to form, with different orderings of the proteins and different states of protein phosphorylation. By examination of the steady states of the system under different parameter values, it was found that, of the four ternary complexes which *could* form in the model, only the experimentally hypothesised ternary complex existed in the late time dynamics. Given the reasonably large number of variables in the model and underlying reactions, the explicit steady state solutions for each variable were very complex and difficult to analyse analytically, hence numerical techniques were employed. The stability of the steady state reached when using the experimental initial conditions and dissociation constants was assessed for varying parameter values, and it was found to be stable for many biologically relevant parameter regimes. Additionally, the stability of the steady state reached when the initial conditions and dissociation constants were allowed to vary slightly from the experimental values was also

assessed, again for varying parameter values, and in the majority of cases the steady state was found to be stable. This chapter provides an example of how mathematical modelling can be used to test biological hypotheses which would be difficult, or even impossible, to test experimentally. As an extension to this work, if quantitative data were to be produced, the model could be calibrated using the data to determine more accurate parameter values. With these parameter values in hand, the model could be extended to include aggregation of the ternary complexes to form droplets, as is observed experimentally.

Finally, Chapter 6 is based around an experimental dataset from AstraZeneca, in which the abundance of different proteins in the MAPK signalling pathway is measured upon treatment of the cells with various EGFR inhibitors. Given the large size of the dataset, the aim here was to use statistical techniques to identify differences in the data, for example which inhibitors were having the greatest effect on which proteins, and whether these effects were dependent on the cell line (WT or mutant EGFR). A thorough analysis of the dataset was carried out using two-way ANOVA, Tukey's honest significant difference test, and principal component analysis. Some differences between inhibitor types were identified, such as the inhibitor D3 having the greatest effect at the lowest concentration. The original aim for this project was to identify statistically significantly different groupings of data using the aforementioned statistical methods and then to use a mechanistic mathematical model to explain these differences in terms of reactions and rate constants with the use of Bayesian inference, as in Chapter 4. However, given some unexplained trends in the control data and the general lack of significant trends in the inhibitor treated data, the mathematical modelling was not conducted. Instead, a review of the current literature surrounding mathematical modelling of the interactions between EGFR and inhibitors is given, focussing on two clear themes in the literature: enzyme kinetics and combination therapies. The work in this chapter could be extended if further datasets were available, which showed greater and more consistent changes from baseline when the cells are treated with the inhibitors, and if the control data could be mechanistically explained.

Understanding signalling pathways at the level of the individual reactions can be crucial in appreciating the role of specific molecules in different healthy

and disease scenarios. Mathematical and statistical methods, such as those used and developed in this thesis, can allow for insights into cell signalling and its dysregulation, which may be difficult or costly to find experimentally. In this thesis, general receptor-ligand interactions have been modelled, as well as more specific parts of signalling pathways initiated by either a receptor tyrosine kinase, or a cytokine receptor, where the general theme linking all data driven chapters is that these pathways have some relevance in human diseases and disorders. The analysis carried out for the biological systems in this thesis has confirmed experimental hypotheses, elucidated important reactions and rates and allowed for predictions to be made relating to specific cell signalling pathways. As well as this, new methodologies have been developed to analyse a general stochastic model of receptor competition for a common ligand type.

Appendix A

Identifiability analysis for the HypIL-6 mathematical model

In this appendix, the working for the structural identifiability analysis for the HypIL-6 mathematical model under hypothesis 1 is presented based on the method by Castro & de Boer (2020), described in Section 4.3.2. Firstly, Equations (4.1) - (4.22), defining the HypIL-6 mathematical model are rewritten in simpler notation, given in Table A.1, as in Equations (A.1) - (A.22), to aid readability in this section. Each functionally independent term in each of the ODEs is then equated to its scaled form where, for example,

$$r^+_{1,6}x_1x_2 = \frac{1}{u_{x_1}}u_{r^+_{1,6}}r^+_{1,6}u_{x_1}x_1u_{x_2}x_2,$$

from Equation (A.1). Given that the initial concentration of ligand, $[L_6] \equiv x_2$, is known and fixed in the experiments, it is given by the method that $u_{x_2} = 1$. All other scaling constants u must be derived by solving the system of identifiability equations, which is demonstrated here.

$$\frac{dx_1}{dt} = -r^+_{1,6}x_1x_2 + r^-_{1,6}x_3 - \beta_6x_1 \tag{A.1}$$

$$\frac{dx_2}{dt} = -r^+_{1,6}x_1x_2 + r^-_{1,6}x_3 \tag{A.2}$$

A. IDENTIFIABILITY ANALYSIS FOR THE HYPIL-6 MATHEMATICAL MODEL

Notation		Notation	
Original	Simplified	Original	Simplified
$[R_1]$	x_1	$[S_3 \cdot D_6 \cdot S_3]$	x_{12}
$[L_6]$	x_2	$[pS_1 \cdot D_6 \cdot S_1]$	x_{13}
$[C_1]$	x_3	$[pS_3 \cdot D_6 \cdot S_3]$	x_{14}
$[D_6]$	x_4	$[pS_1 \cdot D_6 \cdot pS_1]$	x_{15}
$[S_1]$	x_5	$[pS_3 \cdot D_6 \cdot pS_3]$	x_{16}
$[S_3]$	x_6	$[S_1 \cdot D_6 \cdot S_3]$	x_{17}
$[D_6 \cdot S_1]$	x_7	$[pS_1 \cdot D_6 \cdot S_3]$	x_{18}
$[D_6 \cdot S_3]$	x_8	$[S_1 \cdot D_6 \cdot pS_3]$	x_{19}
$[D_6 \cdot pS_1]$	x_9	$[pS_3 \cdot D_6 \cdot pS_3]$	x_{20}
$[D_6 \cdot pS_3]$	x_{10}	$[pS_1]$	x_{21}
$[S_1 \cdot D_6 \cdot S_1]$	x_{11}	$[pS_3]$	x_{22}

Table A.1: Simplified notation for the variables of the HypIL-6 mathematical model under hypothesis 1, to be used in the structural identifiability analysis.

$$\frac{dx_3}{dt} = r^+_{1,6}x_1x_2 - 2r^+_{2,6}x_3^2 + 2r^-_{2,6}x_4 - (\beta_6 + r^-_{1,6})x_3 \tag{A.3}$$

$$\frac{dx_4}{dt} = r^+_{2,6}x_3^2 - 2k^+_{1a}x_4x_5 + k^-_{1a}(x_7 + x_9) - 2k^+_{3a}x_4x_6 + k^-_{3a}(x_8 + x_{10}) - (\beta_6 + r^-_{2,6})x_4 \tag{A.4}$$

$$\frac{dx_5}{dt} = -k^+_{1a}x_5(2x_4 + x_7 + x_8 + x_9 + x_{10}) + k^-_{1a}(x_7 + 2x_{11} + x_{17} + x_{13} + x_{19}) + d_1x_{21} \tag{A.5}$$

$$\frac{dx_6}{dt} = -k^+_{3a}x_6(2x_4 + x_8 + x_7 + x_{10} + x_9) + k^-_{3a}(x_8 + 2x_{12} + x_{17} + x_{14} + x_{18}) + d_3x_{22} \tag{A.6}$$

$$\frac{dx_7}{dt} = 2k^+_{1a}x_5x_4 - k^+_{1a}x_7x_5 + 2k^-_{1a}x_{11} - k^+_{3a}x_7x_6 + k^-_{3a}x_{17} + k^-_{1a}x_{13} + k^-_{3a}x_{19} - (\beta_6 + k^-_{1a} + q)x_7 \tag{A.7}$$

$$\frac{dx_8}{dt} = 2k^+_{3a}x_6x_4 - k^+_{3a}x_8x_6 + 2k^-_{3a}x_{12} - k^+_{1a}x_8x_5 + k^-_{1a}x_{17} + k^-_{1a}x_{18} + k^-_{3a}x_{14}$$

$$- (\beta_6 + k_{3a}^- + q)x_8 \tag{A.8}$$

$$\frac{dx_9}{dt} = -k_{1a}^+ x_5 x_9 + k_{1a}^- x_{13} - k_{3a}^+ x_6 x_9 + k_{3a}^- x_{18} + q x_7 + 2k_{1a}^- x_{15} + k_{3a}^- x_{20} - (\beta_6 + k_{1a}^-)x_9 \tag{A.9}$$

$$\frac{dx_{10}}{dt} = -k_{3a}^+ x_6 x_{10} + k_{3a}^- x_{14} - k_{1a}^+ x_5 x_{10} + k_{1a}^- x_{19} + q x_8 + 2k_{3a}^- x_{16} + k_{1a}^- x_{20} - (\beta_6 + k_{3a}^-)x_{10} \tag{A.10}$$

$$\frac{dx_{11}}{dt} = k_{1a}^+ x_5 x_7 - (2k_{1a}^- + 2q + \beta_6)x_{11} \tag{A.11}$$

$$\frac{dx_{12}}{dt} = k_{3a}^+ x_6 x_8 - (2k_{3a}^- + 2q + \beta_6)x_{12} \tag{A.12}$$

$$\frac{dx_{13}}{dt} = k_{1a}^+ x_9 x_5 + 2q x_{11} - (q + \beta_6 + 2k_{1a}^-)x_{13} \tag{A.13}$$

$$\frac{dx_{14}}{dt} = k_{3a}^+ x_{10} x_6 + 2q x_{12} - (q + \beta_6 + 2k_{3a}^-)x_{14} \tag{A.14}$$

$$\frac{dx_{15}}{dt} = q x_{13} - (2k_{1a}^- + \beta_6)x_{15} \tag{A.15}$$

$$\frac{dx_{16}}{dt} = q x_{14} - (2k_{3a}^- + \beta_6)x_{16} \tag{A.16}$$

$$\frac{dx_{17}}{dt} = k_{1a}^+ x_5 x_8 + k_{3a}^+ x_7 x_6 - (k_{3a}^- + 2q + \beta_6 + k_{1a}^-)x_{17} \tag{A.17}$$

$$\frac{dx_{18}}{dt} = q x_{17} + k_{3a}^+ x_9 x_6 - (k_{3a}^- + q + k_{1a}^- + \beta_6)x_{18} \tag{A.18}$$

$$\frac{dx_{19}}{dt} = q x_{17} + k_{1a}^+ x_5 x_{10} - (k_{1a}^- + q + k_{3a}^- + \beta_6)x_{19} \tag{A.19}$$

$$\frac{dx_{20}}{dt} = q(x_{19} + x_{18}) - (k_{1a}^- + k_{3a}^- + \beta_6)x_{20} \tag{A.20}$$

$$\frac{dx_{21}}{dt} = k_{1a}^-(x_9 + x_{13} + x_{18} + x_{20} + 2x_{15}) - d_1 x_{21} \tag{A.21}$$

$$\frac{dx_{22}}{dt} = k_{3a}^-(x_{10} + x_{14} + x_{19} + x_{20} + 2x_{16}) - d_3 x_{22} \tag{A.22}$$

A. IDENTIFIABILITY ANALYSIS FOR THE HYPIL-6 MATHEMATICAL MODEL

Firstly, from Equation (A.1),

$$\beta_6 x_1 = \frac{1}{u_{x_1}} u_{x_1} x_1 u_{\beta_6} \beta_6 \implies u_{\beta_6} = 1,$$

since the other terms cancel. Similar linear terms in Equations (A.21) and (A.22) suffice to show that $u_{d_1} = u_{d_3} = 1$. Given that β_6 is identifiable, the identifiability equation

$$(\beta_6 + r^-_{1,6}) x_3 = \frac{1}{u_{x_3}} (u_{\beta_6} \beta_6 + u_{r^-_{1,6}} r^-_{1,6}) u_{x_3} x_3,$$

derived from Equation (A.3) implies that $u_{r^-_{1,6}} = 1$ and a similar equation derived from Equation (A.4) implies that $u_{r^-_{2,6}} = 1$. In a similar fashion, the terms $(\beta_6 + k^-_{1a})x_9$ and $(\beta_6 + k^-_{3a})x_{10}$ from Equations (A.9) and (A.10) respectively, result in identifiability equations which can be simplified to show that $u_{k^-_{1a}} = u_{k^-_{3a}} = 1$. Now that it is known that $u_{\beta_6} = u_{k^-_{1a}} = 1$, the identifiability equation

$$(2k^-_{1a} + 2q + \beta_6) x_{11} = \frac{1}{u_{x_{11}}} (2u_{k^-_{1a}} k^-_{1a} + 2u_q q + u_{\beta_6} \beta_6) u_{x_{11}} x_{11},$$

where the term is taken from Equation (A.11), can be used to show that $u_q = 1$. Then, from Equation (A.1), given that $u_{x_2} = 1$, the term $r^+_{1,6} x_1 x_2$ allows to prove that $u_{r^+_{1,6}} = 1$ and with this, the identifiability equation derived from the same term but taken from Equation (A.2) can be used to show that $u_{x_1} = 1$. It can be found by equating the terms $r^-_{1,6} x_3$ and $2r^-_{2,6} x_4$ from Equations (A.1) and (A.3) respectively, to their scaled counterparts, that $u_{x_3} = u_{x_4} = 1$. The rate of dimer formation, $r^+_{2,6}$, is found to be identifiable when considering the identifiability equation for the term $r^+_{2,6} x_3^2$ from Equation (A.4). The remaining two rate constants, k^+_{1a} and k^+_{3a} are also both identifiable as can be seen by examining the terms $2k^+_{1a} x_4 x_5$ and $2k^+_{3a} x_4 x_6$ from Equations (A.5) and (A.6).

It then remains to determine whether the remaining variables are observable. From Equation (A.4), the identifiability equation for the term $2k^+_{1a} x_4 x_5$ can be used to show that $u_{x_5} = 1$ and the term $2k^+_{3a} x_4 x_6$, can be used to show that $u_{x_6} = 1$. Five functionally independent terms are given by $k^-_{1a}(x_7 + 2x_{11} + x_{17} + x_{13} + x_{19})$ in Equation (A.5), and, given that $u_{x_5} = 1$ and $u_{k^-_{1a}} = 1$, it can be seen that

$u_{x_7} = u_{x_{11}} = u_{x_{13}} = u_{x_{17}} = u_{x_{19}} = 1$. The last term in the same equation can be used to prove that $u_{x_{21}} = 1$. Equation (A.6) contains very similar terms involving the coefficients relating to STAT3 instead of STAT1, and hence from this equation one can find that $u_{x_8} = u_{x_{12}} = u_{x_{14}} = u_{x_{18}} = u_{x_{22}} = 1$. The variable x_{20} is found to be observable through solving the identifiability equation relating to the term qx_{19} in Equation (A.20). The variables x_9 and x_{10} are also observable which can be derived from Equation (A.5), specifically looking at the terms $k^+_{1a}x_5x_9$ and $k^+_{1a}x_5x_{10}$. Finally, the two remaining variables x_{15} and x_{16} appear in the terms $2k^-_{1a}x_{15}$ and $2k^-_{3a}x_{16}$ in Equations (A.21) and (A.22) respectively, and thus again, these variables are observable.

The HypIL-6 mathematical model under hypothesis 1, with known initial concentration $[L_6](0)$ is therefore fully structurally identifiable. Although not presented here, the same is true for the HypIL-6 mathematical model under hypothesis 2 as well as both of the IL-27 mathematical models. It is therefore appropriate to attempt to estimate each parameter in the mathematical models individually as is carried out in Section 4.3.5.

A. IDENTIFIABILITY ANALYSIS FOR THE HYPIL-6 MATHEMATICAL MODEL

Appendix B

BioNetGen code for the HypIL-6 SOCS3 mathematical model

In this appendix, an example BioNetGen code is given for the HypIL-6 mathematical model with the additional reactions describing negative feedback by SOCS3. The parameter values are a random set sampled from the Th-1 cell posterior distributions.

```
begin parameters
   r16p 0.00363718138359308 #GP130 IL-6 binding
   r16m 0.006528167482425672 #GP130 IL-6 unbinding
   r26p 2.4057717991321548 #GP130 dimerisation
   r26m 0.02633952268026585 #GP130 dimer dissociation
   k1ap 0.020010330131956595 #STAT1 binding to GP130
   k1am 0.07470043883835913 #STAT1 unbinding GP130
   k3ap 0.12041002616902625 #STAT3 binding GP130
   k3am 0.16601839306976587 #STAT3 unbinding GP130
   q 0.004642309832977627 #STAT phosphorylation
   d1 0.0011089072518100501 #STAT1 dephosphorylation
   d3 0.0004716010068422507 #STAT3 dephosphorylation
   R10 2.8382589199305848 #Initial concentration of GP130
   S10 589.1642864948094 #Initial concentration of STAT1
   S30 90.66846449691016 #Initial concentration of STAT3
   L0 10 #Initial concentration of IL-6
   X3a0 20 #Initial concentration of SOCS3 dummy 1
   delta1 0.0005 #Rate of SOCS3 dummy 1 to dummy 2
```

```
    delta2 0.0005 #Rate of SOCS3 dummy 2 to dummy 3
    delta3 0.0005 #Rate of SOCS3 dummy 3 to SOCS3
    alpha 0.005 #Rate of SOCS3 binding to receptors
end parameters

begin molecule types
   R(l,r,s,x,T~U~P) #GP130
   L(r) #IL-6
   S1(r,T~U~P) #STAT1
   S3(r,T~U~P) #STAT3
   X3a() #SOCS3 dummy 1
   X3b() #SOCS3 dummy 2
   X3c() #SOCS3 dummy 3
   X3(r) #SOCS3
end molecule types

begin seed species
   R(l,r,s,x,T~U) R10 #GP130
   L(r) L0 #IL-6
   S1(r,T~U) S10 #STAT1
   S3(r,T~U) S30 #STAT3
   X3a() X3a0 #SOCS3 dummy
end seed species

begin reaction rules
   #GP130 IL-6 binding and unbinding
   R(l,r,s,x,T~U) + L(r) <-> R(l!1,r,s,x,T~U).L(r!1) r16p, r16m
   #Dimerisation and dissociation of the dimer
   R(l!+,r,s,x,T~U) + R(l!+,r,s,x,T~U) <->
      R(l!+,r!1,s,x,T~P).R(l!+,r!1,s,x,T~P) 2*r26p, r26m
   #STAT1 binding the GP130 dimer
   S1(r,T~U) + R(l!+,r!1,s,x!?,T~P).R(l!+,r!1,s,x!?) ->
      S1(r!2,T~U).R(l!+,r!1,s!2,x!?,T~P).R(l!+,r!1,s,x!?) k1ap
   S1(r,T~U) + S1(r!2).R(l!+,r!1,s!2,x!?).R(l!+,r!1,s,x!?,T~P) ->
      S1(r!2).R(l!+,r!1,s!2,x!?).R(l!+,r!1,s!3,x!?,T~P).S1(r!3,T~U)
       k1ap
   S1(r,T~U) + S3(r!2).R(l!+,r!1,s!2,x!?).R(l!+,r!1,s,x!?,T~P) ->
      S3(r!2).R(l!+,r!1,s!2,x!?).R(l!+,r!1,s!3,x!?,T~P).S1(r!3,T~U)
       k1ap
   #STAT3 binding the GP130 dimer
   S3(r,T~U) + R(l!+,r!1,s,x!?,T~P).R(l!+,r!1,s,x!?) ->
```

```
      S3(r!2,T~U).R(l!+,r!1,s!2,x!?,T~P).R(l!+,r!1,s,x!?) k3ap
   S3(r,T~U) + S1(r!2).R(l!+,r!1,s!2,x!?).R(l!+,r!1,s,x!?,T~P) ->
      S1(r!2).R(l!+,r!1,s!2,x!?).R(l!+,r!1,s!3,x!?,T~P).S3(r!3,T~U)
       k3ap
   S3(r,T~U) + S3(r!2).R(l!+,r!1,s!2,x!?).R(l!+,r!1,s,x!?,T~P) ->
      S3(r!2).R(l!+,r!1,s!2,x!?).R(l!+,r!1,s!3,x!?,T~P).S3(r!3,T~U)
       k3ap
   #STAT1 phosphorylation on the dimer
   S1(r!2,T~U).R(l!+,r!1,s!2,x!?,T~P).R(l!+,r!1,s!?,x!?) ->
      S1(r!2,T~P).R(l!+,r!1,s!2,x!?,T~P).R(l!+,r!1,s!?,x!?) q
   #STAT3 phosphorylation on the dimer
   S3(r!2,T~U).R(l!+,r!1,s!2,x!?,T~P).R(l!+,r!1,s!?,x!?) ->
      S3(r!2,T~P).R(l!+,r!1,s!2,x!?,T~P).R(l!+,r!1,s!?,x!?) q
   #(p)STAT1 dissociation from the dimer
   S1(r!2).R(l!+,r!1,s!2,x!?).R(l!+,r!1,s!?,x!?) ->
      S1(r) + R(l!+,r!1,s,x!?).R(l!+,r!1,s!?,x!?) k1am
   #(p)STAT3 dissociation from the dimer (8)
   S3(r!2).R(l!+,r!1,s!2,x!?).R(l!+,r!1,s!?,x!?) ->
      S3(r) + R(l!+,r!1,s,x!?).R(l!+,r!1,s!?,x!?) k3am
   #pSTAT1 dephosphorylation in the cytoplasm
   S1(r,T~P) -> S1(r,T~U) d1
   #pSTAT3 dephosphorylation in the cytoplasm
   S3(r,T~P) -> S3(r,T~U) d3
   #SOCS3 dummy 1 becoming SOCS3 dummy 2
   X3a() -> X3b() delta1
   #SOCS3 dummy 2 becoming SOCS3 dummy 3
   X3b() -> X3c() delta2
   #SOCS3 dummy 3 becoming active SOCS3
   X3c() -> X3(r) delta3
   #SOCS3 deactivation of receptor
   X3(r) + R(l!+,r!+,s!?,x,T~P) ->
      X3(r!1).R(l!+,r!+,s!?,x!1,T~U) alpha
end reaction rules

begin observables
   #pSTAT1
   Molecules pS1 S1(r!?,T~P)
   #pSTAT3
   Molecules pS3 S3(r!?,T~P)
end observables
```

B. BIONETGEN CODE FOR THE HYPIL-6 SOCS3 MATHEMATICAL MODEL

```
generate_network();
simulate_ode({t_end=>10800,n_steps=>10800});
```

Appendix C

Partial derivatives within the FGFR2 model Jacobian

In this appendix, the partial derivatives comprising the Jacobian matrix for the FGFR2 model, defined in Section 5.3.3, are given. Where a partial derivative of a function f_i for $i = 1, \ldots, 11$ with respect to a model variable is not given, it can be assumed to be equal to 0.

Partial derivatives of f_1:

$$\begin{aligned}\frac{\partial f_1}{\partial [pF]} = &-k_1 - k_{+2}([S]^T - [pF \cdot S] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] \\ &- [pF \cdot pP \cdot S]) - k_{+3}([P]^T - [pF \cdot P] - [pF \cdot pP] - [pP] - [S \cdot pP] \\ &- [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])\end{aligned}$$

$$\frac{\partial f_1}{\partial [pF \cdot S]} = -k_1 + k_{-2} + k_{+2}[pF]$$

$$\frac{\partial f_1}{\partial [pF \cdot P]} = -k_1 + k_{-3} + k_{+3}[pF]$$

$$\frac{\partial f_1}{\partial [pF \cdot pP]} = -k_1 + k_{+3}[pF] + k_5$$

$$\frac{\partial f_1}{\partial [pP]} = \frac{\partial f_1}{\partial [S \cdot pP]} = k_{+3}[pF]$$

$$\frac{\partial f_1}{\partial [pF \cdot S \cdot P]} = \frac{\partial f_1}{\partial [pF \cdot S \cdot pP]} = -k_1 + k_{-2} + k_{+2}[pF] + k_{+3}[pF]$$

C. PARTIAL DERIVATIVES WITHIN THE FGFR2 MODEL JACOBIAN

$$\frac{\partial f_1}{\partial [pF \cdot P \cdot S]} = -k_1 + k_{+2}[pF] + k_{-3} + k_{+3}[pF]$$

$$\frac{\partial f_1}{\partial [pF \cdot pP \cdot S]} = -k_1 + k_{+2}[pF] + k_{+3}[pF] + k_5$$

Partial derivatives of f_2:

$$\begin{aligned}\frac{\partial f_2}{\partial [pF]} &= k_{+2}([S]^T - [pF \cdot S] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] \\ &\quad - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])\end{aligned}$$

$$\frac{\partial f_2}{\partial [pF \cdot S]} = -k_{+2}[pF] - k_{-2} - k_{+7}[pP]$$

$$\frac{\partial f_2}{\partial [pP]} = -k_{+7}[pF \cdot S]$$

$$\frac{\partial f_2}{\partial [S \cdot P]} = \frac{\partial f_2}{\partial [S \cdot pP]} = \frac{\partial f_2}{\partial [pF \cdot S \cdot P]} = \frac{\partial f_2}{\partial [pF \cdot P \cdot S]} = \frac{\partial f_2}{\partial [pF \cdot pP \cdot S]} = -k_{+2}[pF]$$

$$\frac{\partial f_2}{\partial [pF \cdot S \cdot pP]} = -k_{+2}[pF] + k_{-7}$$

Partial derivatives of f_3:

$$\begin{aligned}\frac{\partial f_3}{\partial [pF]} &= k_{+3}([P]^T - [pF \cdot P] - [pF \cdot pP] - [pP] - [S \cdot P] - [S \cdot pP] \\ &\quad - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])\end{aligned}$$

$$\frac{\partial f_3}{\partial [pF \cdot P]} = -k_{+3}[pF] - k_{-3} - k_4$$

$$\begin{aligned}\frac{\partial f_3}{\partial [pF \cdot pP]} &= \frac{\partial f_3}{\partial [pP]} = \frac{\partial f_3}{\partial [S \cdot P]} = \frac{\partial f_3}{\partial [S \cdot pP]} = \frac{\partial f_3}{\partial [pF \cdot S \cdot P]} = \frac{\partial f_3}{\partial [pF \cdot P \cdot S]} \\ &= \frac{\partial f_3}{\partial [pF \cdot S \cdot pP]} = \frac{\partial f_3}{\partial [pF \cdot pP \cdot S]} = -k_{+3}[pF]\end{aligned}$$

Partial derivatives of f_4:

$$\frac{\partial f_4}{\partial [pF \cdot P]} = k_4$$

$$\frac{\partial f_4}{\partial [pF \cdot pP]} = -k_5$$

Partial derivatives of f_5:

$$\frac{\partial f_5}{\partial [pF]} = 0$$

$$\frac{\partial f_5}{\partial [pF \cdot pP]} = k_5$$

$$\begin{aligned}\frac{\partial f_5}{\partial [pP]} &= -k_{+7}([S]^T - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] \\ &\quad - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])\end{aligned}$$

$$\frac{\partial f_5}{\partial [S \cdot P]} = \frac{\partial f_5}{\partial [pF \cdot S \cdot P]} = \frac{\partial f_5}{\partial [pF \cdot P \cdot S]} = \frac{\partial f_5}{\partial [pF \cdot pP \cdot S]} = k_{+7}[pP]$$

$$\frac{\partial f_5}{\partial [S \cdot pP]} = \frac{\partial f_5}{\partial [pF \cdot S \cdot pP]} = k_{+7}[pP] + k_{-7}$$

Partial derivatives of f_6:

$$\frac{\partial f_6}{\partial [pF]} = -k_{+2}[S \cdot P] - k_{+3}[S \cdot P]$$

$$\begin{aligned}\frac{\partial f_6}{\partial [pF \cdot S]} &= -k_{+6}([P]^T - [pF \cdot P] - [pF \cdot pP] - [pP] - [S \cdot P] - [S \cdot pP] \\ &\quad - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])\end{aligned}$$

$$\begin{aligned}\frac{\partial f_6}{\partial [pF \cdot P]} = \frac{\partial f_6}{\partial [pF \cdot pP]} = \frac{\partial f_6}{\partial [pP]} &= -k_{+6}([S]^T - [pF \cdot S] - [S \cdot P] - [S \cdot pP] \\ &\quad - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])\end{aligned}$$

$$\begin{aligned}\frac{\partial f_6}{\partial [S \cdot P]} &= -k_{-6} - k_{+2}[pF] - k_{+3}[pF] - k_{+6}([P]^T - [pF \cdot P] - [pF \cdot pP] \\ &\quad - [pP] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] \\ &\quad - [pF \cdot pP \cdot S]) - k_{+6}([S]^T - [pF \cdot S] - [S \cdot P] - [S \cdot pP] \\ &\quad - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])\end{aligned}$$

$$\begin{aligned}\frac{\partial f_6}{\partial [S \cdot pP]} = \frac{\partial f_6}{\partial [pF \cdot S \cdot pP]} = \frac{\partial f_6}{\partial [pF \cdot pP \cdot S]} &= -k_{+6}([P]^T - [pF \cdot P] - [pF \cdot pP] \\ &\quad - [pP] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] \\ &\quad - [pF \cdot pP \cdot S]) - k_{+6}([S]^T - [pF \cdot S] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] \\ &\quad - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])\end{aligned}$$

C. PARTIAL DERIVATIVES WITHIN THE FGFR2 MODEL JACOBIAN

$$\begin{aligned}\frac{\partial f_6}{\partial [pF \cdot S \cdot P]} &= k_{-2} - k_{+6}([P]^T - [pF \cdot P] - [pF \cdot pP] - [pP] - [S \cdot P] - [S \cdot pP] \\ &\quad - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) \\ &\quad - k_{+6}([S]^T - [pF \cdot S] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] \\ &\quad - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])\end{aligned}$$

$$\begin{aligned}\frac{\partial f_6}{\partial [pF \cdot P \cdot S]} &= k_{-3} - k_{+6}([P]^T - [pF \cdot P] - [pF \cdot pP] - [pP] - [S \cdot P] - [S \cdot pP] \\ &\quad - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) \\ &\quad - k_{+6}([S]^T - [pF \cdot S] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] \\ &\quad - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])\end{aligned}$$

Partial derivatives of f_7:

$$\frac{\partial f_7}{\partial [pF]} = -k_{+2}[S \cdot pP]$$

$$\frac{\partial f_7}{\partial [pF \cdot S]} = \frac{\partial f_7}{\partial [S \cdot P]} = \frac{\partial f_7}{\partial [pF \cdot S \cdot P]} = \frac{\partial f_7}{\partial [pF \cdot P \cdot S]} = -k_{+7}[pP]$$

$$\begin{aligned}\frac{\partial f_7}{\partial [pP]} &= k_{+7}([S]^T - [pF \cdot S] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] \\ &\quad - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])\end{aligned}$$

$$\frac{\partial f_7}{\partial [S \cdot pP]} = -k_{+7}[pP] - k_{-7} - k_{+2}[pF]$$

$$\frac{\partial f_7}{\partial [pF \cdot S \cdot pP]} = -k_{+7}[pP] + k_{-2}$$

$$\frac{\partial f_7}{\partial [pF \cdot pP \cdot S]} = -k_{+7}[pP] + k_5$$

Partial derivatives of f_8:

$$\frac{\partial f_8}{\partial [pF]} = k_{+2}[S \cdot P]$$

$$\frac{\partial f_8}{\partial [S \cdot P]} = k_{+2}[pF]$$

$$\frac{\partial f_8}{\partial [pF \cdot S \cdot P]} = -k_{-2}$$

Partial derivatives of f_9:

$$\frac{\partial f_9}{\partial [pF]} = k_{+3}[S \cdot P]$$

$$\frac{\partial f_9}{\partial [S \cdot P]} = k_{+3}[pF]$$

$$\frac{\partial f_9}{\partial [pF \cdot P \cdot S]} = -k_{-3} - k_4$$

Partial derivatives of f_{10}:

$$\frac{\partial f_{10}}{\partial [pF]} = k_{+2}[S \cdot pP]$$

$$\frac{\partial f_{10}}{\partial [pF \cdot S]} = k_{+7}[pP]$$

$$\frac{\partial f_{10}}{\partial [pP]} = k_{+7}[pF \cdot S]$$

$$\frac{\partial f_{10}}{\partial [S \cdot pP]} = k_{+2}[pF]$$

$$\frac{\partial f_{10}}{\partial [pF \cdot S \cdot pP]} = -k_{-7} - k_{-2}$$

Partial derivatives of f_{11}:

$$\frac{\partial f_{11}}{\partial [pF \cdot P \cdot S]} = k_4$$

$$\frac{\partial f_{11}}{\partial [pF \cdot pP \cdot S]} = -k_5$$

C. PARTIAL DERIVATIVES WITHIN THE FGFR2 MODEL JACOBIAN

Appendix D

One parameter rescaled FGFR2 mathematical model

In this appendix, the FGFR2 mathematical model defined by Equations (5.29) - (5.39) is rescaled in terms of only the parameter $\frac{k_p}{k_m}$ and the known K_d values where, $k_{-2} = k_{-3} = k_{-6} = k_{-7} = k_5 = k_m$, $k_1 = k_4 = k_p$ and $k_{+i} = \frac{k_m}{K_{d,i}}$ for $i \in \{2, 3, 6, 7\}$. Dividing by k_m throughout results in the Equations (D.1) - (D.11) with $\tau = k_m t$.

$$\begin{aligned}\frac{d[pF]}{d\tau} = \frac{k_p}{k_m}([F]^T &- [pF] - [pF \cdot S] - [pF \cdot P] - [pF \cdot pP] - [pF \cdot S \cdot P] \\ &- [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) + ([pF \cdot S] \\ &+ [pF \cdot S \cdot P] + [pF \cdot S \cdot pP]) - \frac{1}{K_{d,2}}[pF]([S]^T - [pF \cdot S] - [pF \cdot S \cdot P] \\ &- [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) + ([pF \cdot P] \\ &+ [pF \cdot P \cdot S]) - \frac{1}{K_{d,3}}[pF]([P]^T - [pF \cdot P] - [pF \cdot pP] - [pP] - [S \cdot pP] \\ &- [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) \\ &+ ([pF \cdot pP] + [pF \cdot pP \cdot S]) \end{aligned} \tag{D.1}$$

$$\begin{aligned}\frac{d[pF \cdot S]}{d\tau} = \frac{1}{K_{d,2}}[pF]([S]^T &- [pF \cdot S] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] \\ &- [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])\end{aligned}$$

$$- [pF \cdot S] - \frac{1}{K_{d,7}}[pF \cdot S][pP] + [pF \cdot S \cdot pP] \tag{D.2}$$

$$\begin{aligned}\frac{d[pF \cdot P]}{d\tau} &= \frac{1}{K_{d,3}}[pF]([P]^T - [pF \cdot P] - [pF \cdot pP] - [pP] - [S \cdot P] - [S \cdot pP] \\ &- [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) \\ &- [pF \cdot P] - \frac{k_p}{k_m}[pF \cdot P]\end{aligned} \tag{D.3}$$

$$\frac{d[pF \cdot pP]}{d\tau} = \frac{k_p}{k_m}[pF \cdot P] - [pF \cdot pP] \tag{D.4}$$

$$\begin{aligned}\frac{d[pP]}{d\tau} &= [pF \cdot pP] - \frac{1}{K_{d,7}}[pP]([S]^T - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] \\ &- [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) + ([S \cdot pP] \\ &+ [pF \cdot S \cdot pP])\end{aligned} \tag{D.5}$$

$$\begin{aligned}\frac{d[S \cdot P]}{d\tau} &= \frac{1}{K_{d,6}}([S]^T - [pF \cdot S] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] \\ &- [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S])([P]^T - [pF \cdot P] - [pF \cdot pP] - [pP] \\ &- [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] - [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] \\ &- [pF \cdot pP \cdot S]) - [S \cdot P] - \frac{1}{K_{d,2}}[pF][S \cdot P] + [pF \cdot S \cdot P] \\ &- \frac{1}{K_{d,3}}[pF][S \cdot P] + [pF \cdot P \cdot S]\end{aligned} \tag{D.6}$$

$$\begin{aligned}\frac{d[S \cdot pP]}{d\tau} &= \frac{1}{K_{d,7}}[pP]([S]^T - [pF \cdot S] - [S \cdot P] - [S \cdot pP] - [pF \cdot S \cdot P] \\ &- [pF \cdot P \cdot S] - [pF \cdot S \cdot pP] - [pF \cdot pP \cdot S]) - [S \cdot pP] \\ &- \frac{1}{K_{d,2}}[pF][S \cdot pP] + [pF \cdot S \cdot pP] + [pF \cdot pP \cdot S]\end{aligned} \tag{D.7}$$

$$\frac{d[pF \cdot S \cdot P]}{d\tau} = \frac{1}{K_{d,2}}[pF][S \cdot P] - [pF \cdot S \cdot P] \tag{D.8}$$

$$\frac{d[pF \cdot P \cdot S]}{d\tau} = \frac{1}{K_{d,3}}[pF][S \cdot P] - [pF \cdot P \cdot S] - \frac{k_p}{k_m}[pF \cdot P \cdot S] \tag{D.9}$$

$$\frac{d[pF \cdot S \cdot pP]}{d\tau} = \frac{1}{K_{d,7}}[pF \cdot S][pP] - [pF \cdot S \cdot pP] + \frac{1}{K_{d,2}}[pF][S \cdot pP]$$

$$- [pF \cdot S \cdot pP] \tag{D.10}$$

$$\frac{d[pF \cdot pP \cdot S]}{d\tau} = -[pF \cdot pP \cdot S] + \frac{k_p}{k_m}[pF \cdot P \cdot S] \tag{D.11}$$

D. ONE PARAMETER RESCALED FGFR2 MATHEMATICAL MODEL

Appendix E

EGFR inhibition data

Figures of the data used in Chapter 6 are presented in this appendix, for each cell type and inhibitor type. Figures E.1 to E.8 show the data in the WT cell line for the inhibitors D1-D8 and figures E.9 to E.16 show the data in the SVD cell line. Each subplot of a figure shows the MFI data for a specific protein in a specific cellular compartment, where "PM" stands for "plasma membrane".

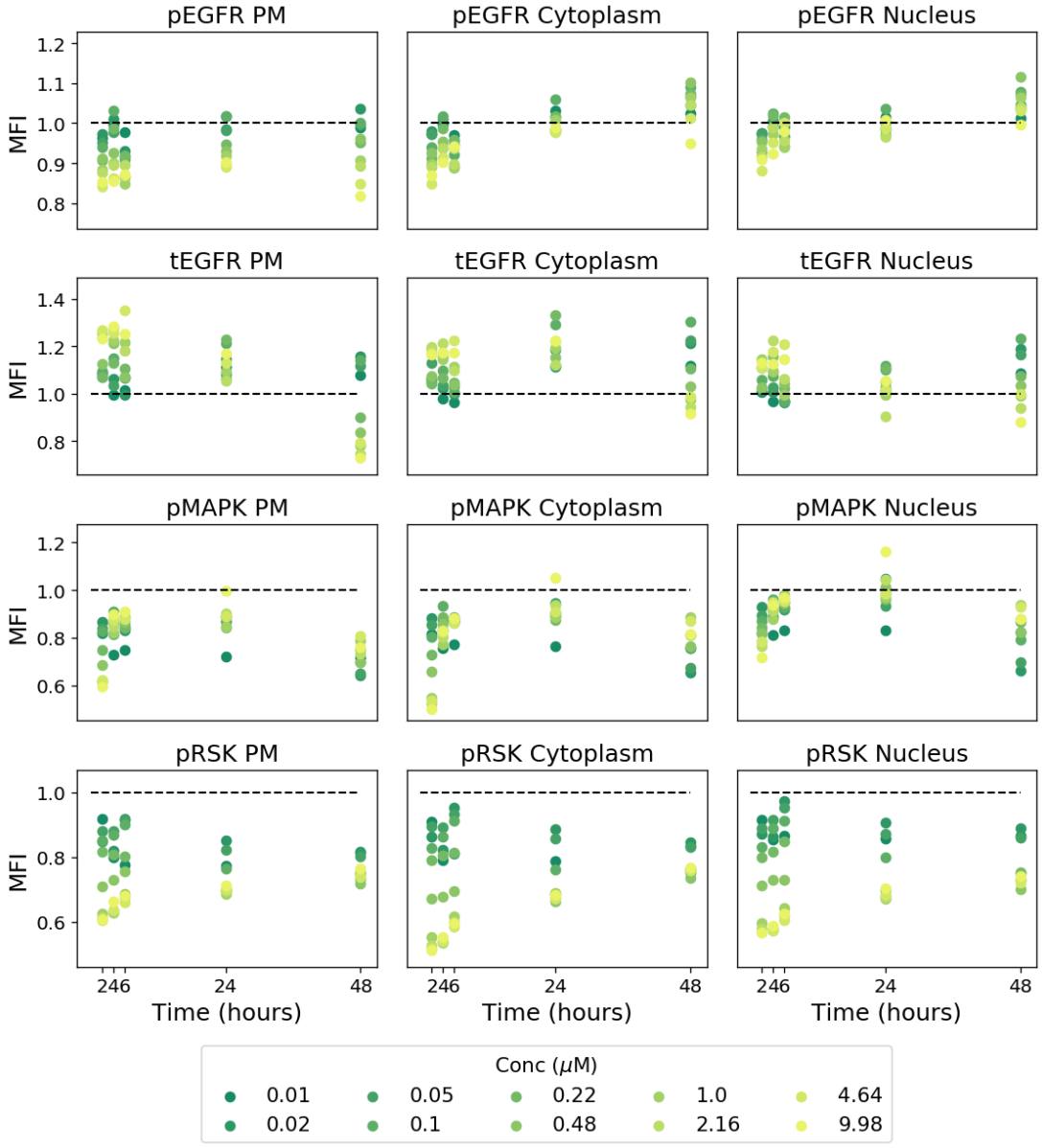

Figure E.1: Figure of the MFI data in the WT cell line under inhibition with inhibitor type D1. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

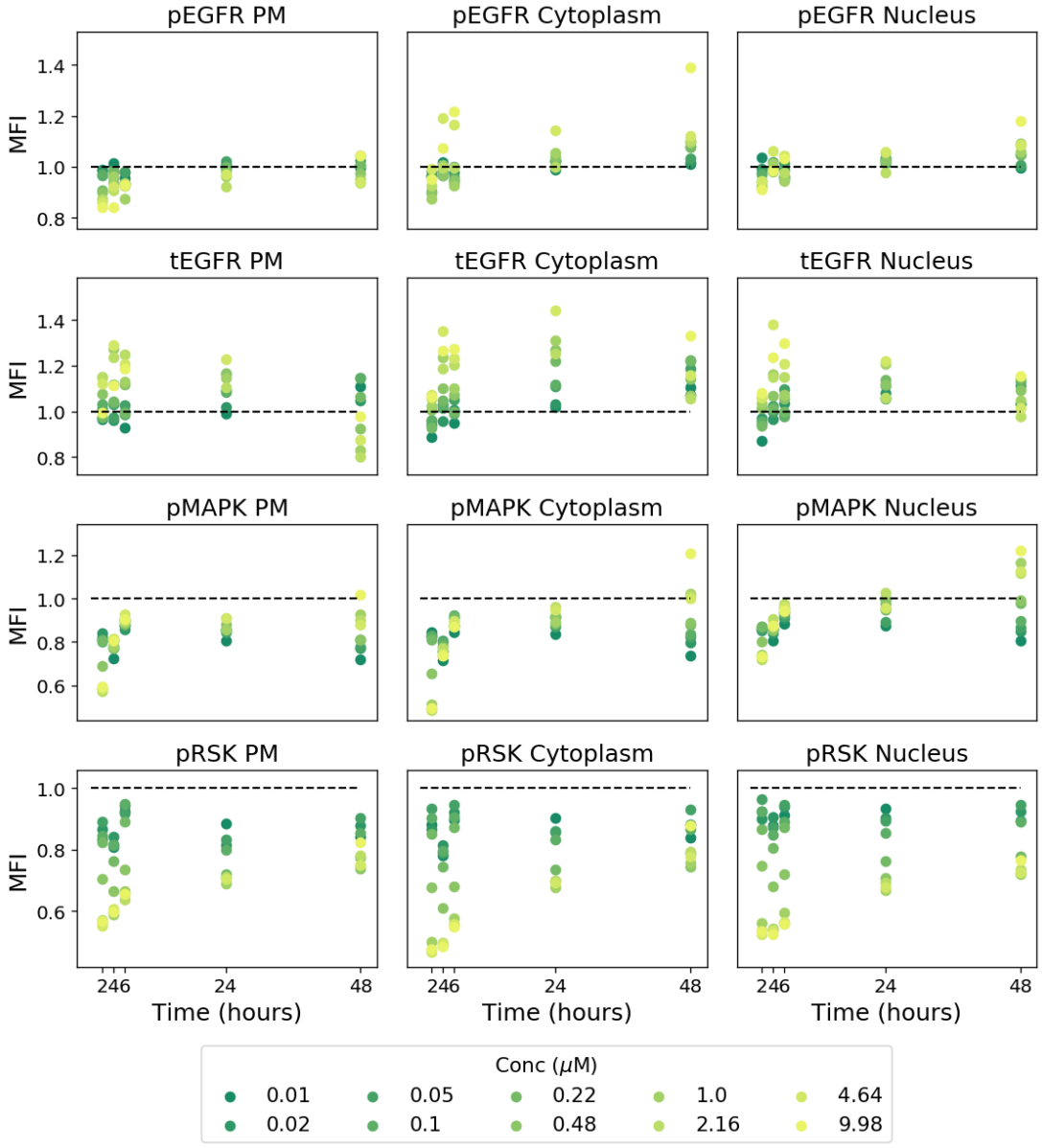

Figure E.2: Figure of the MFI data in the WT cell line under inhibition with inhibitor type D2. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

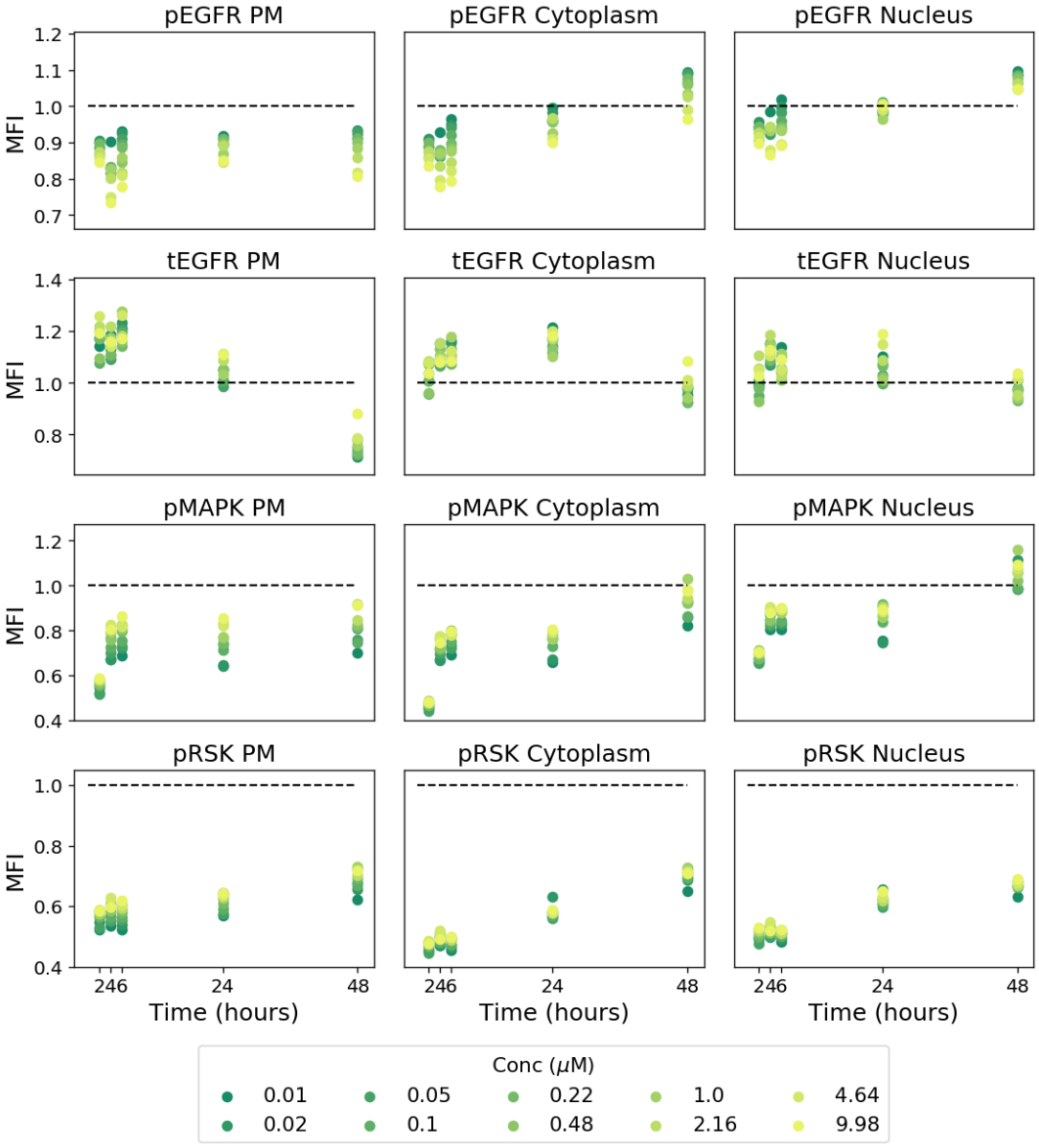

Figure E.3: Figure of the MFI data in the WT cell line under inhibition with inhibitor type D3. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

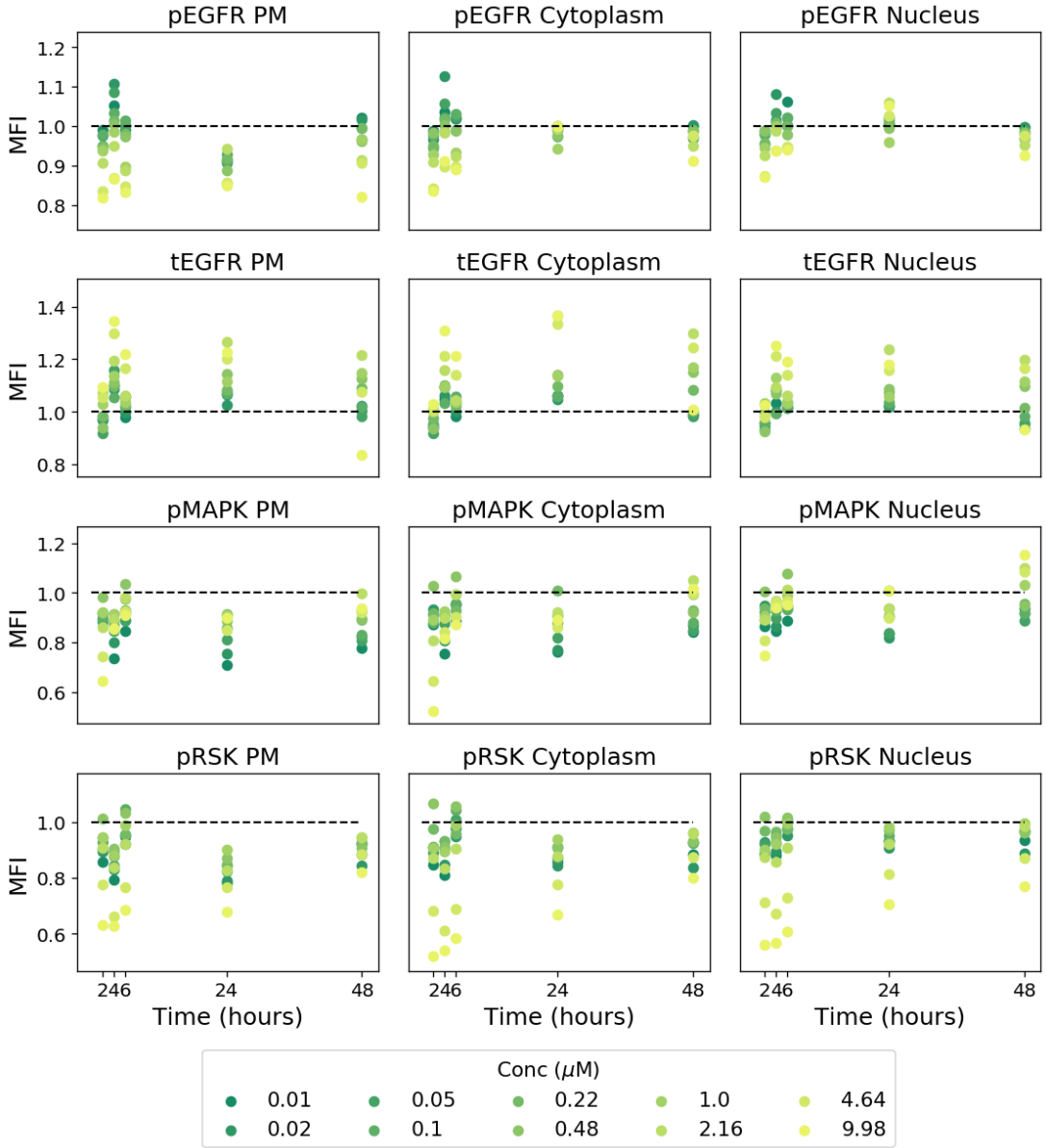

Figure E.4: Figure of the MFI data in the WT cell line under inhibition with inhibitor type D4. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

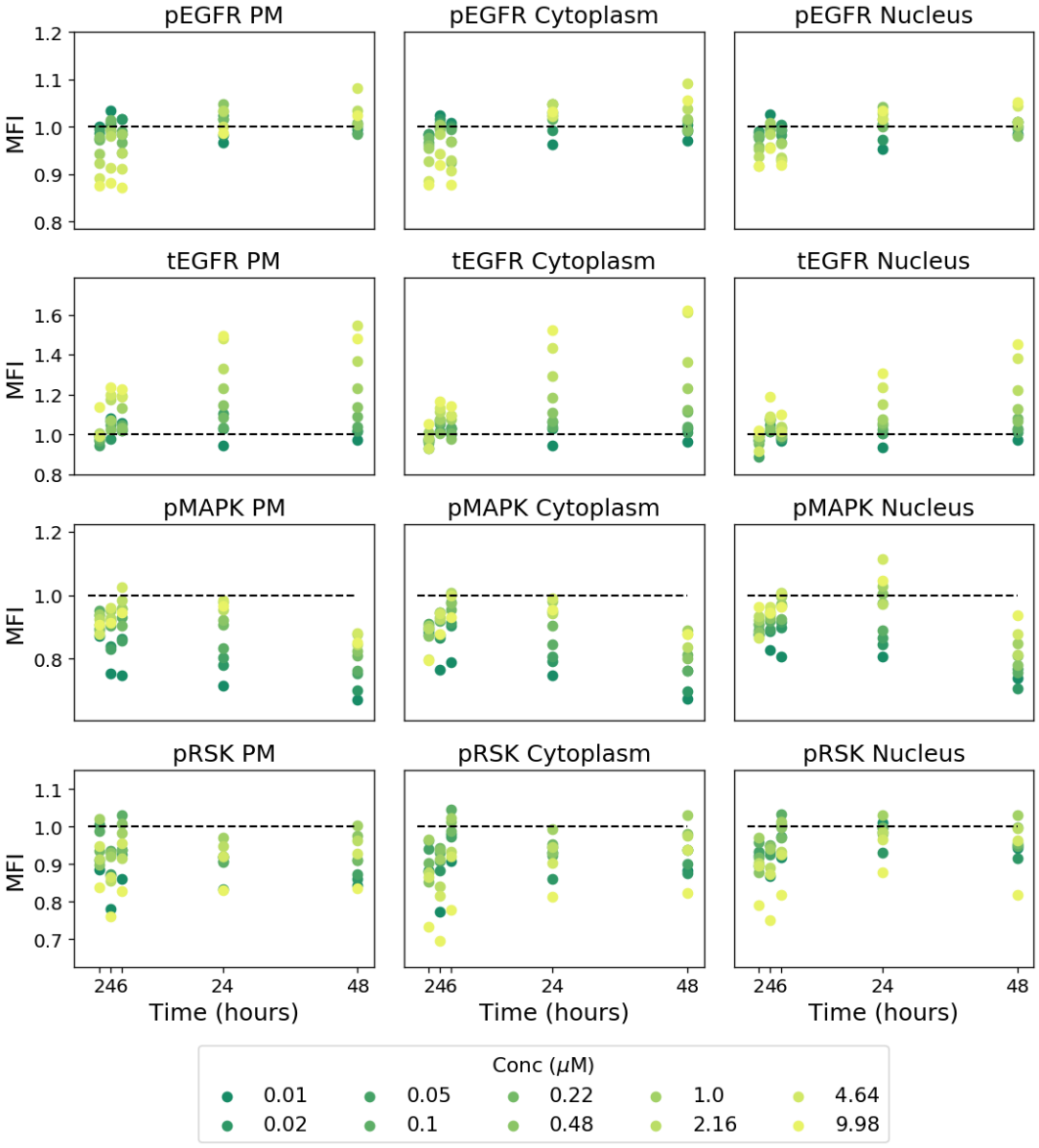

Figure E.5: Figure of the MFI data in the WT cell line under inhibition with inhibitor type D5. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

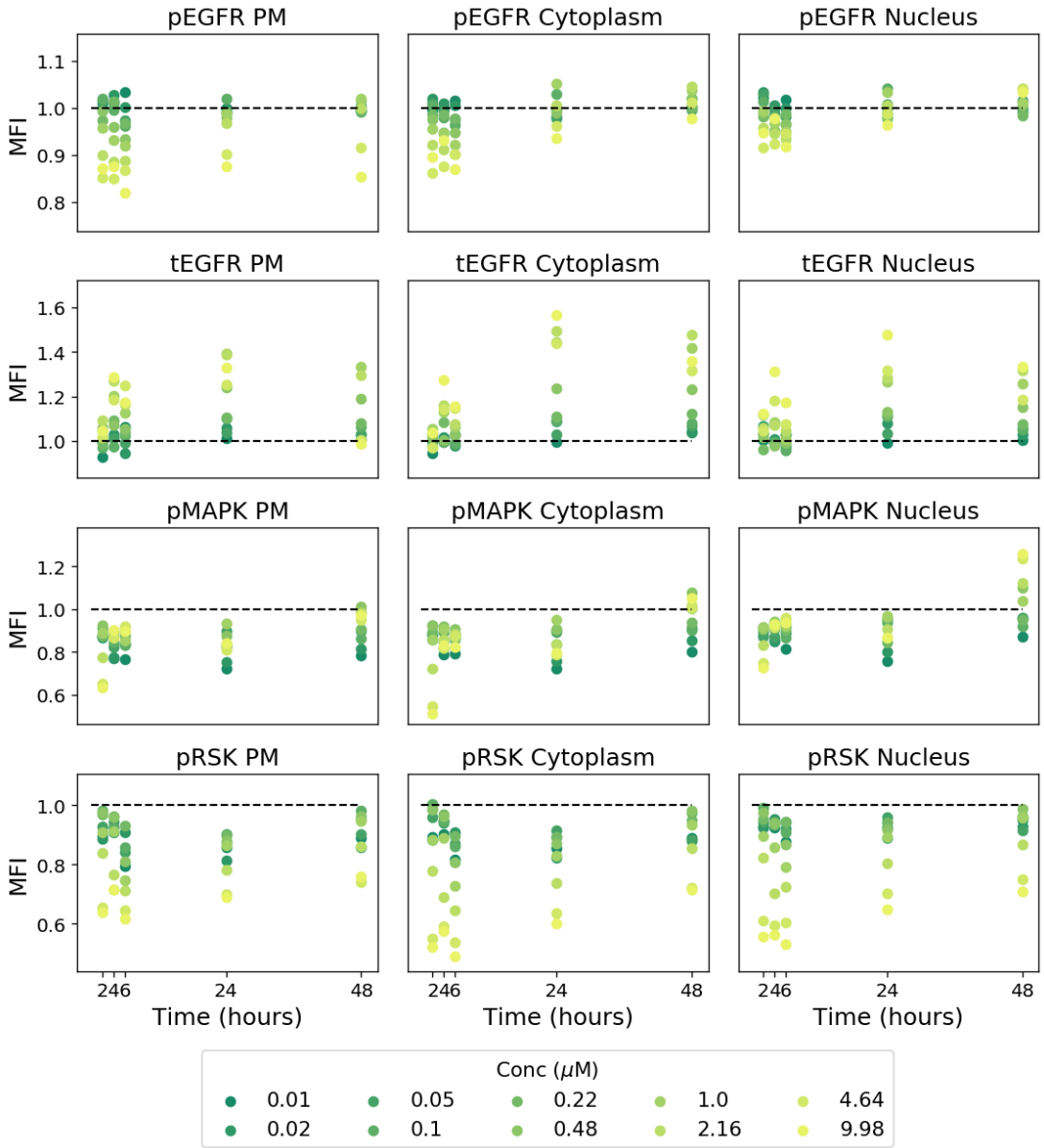

Figure E.6: Figure of the MFI data in the WT cell line under inhibition with inhibitor type D6. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

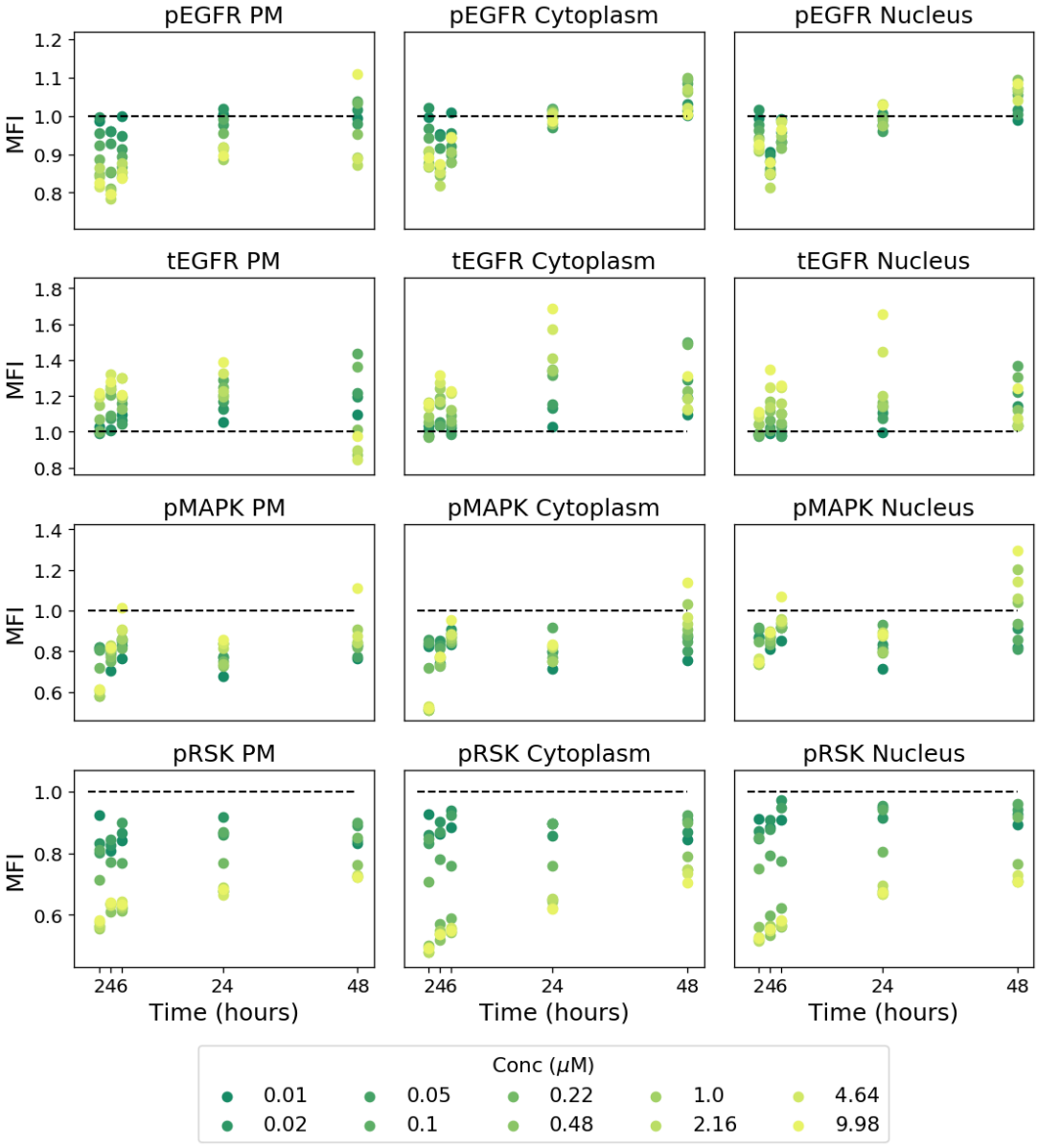

Figure E.7: Figure of the MFI data in the WT cell line under inhibition with inhibitor type D7. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

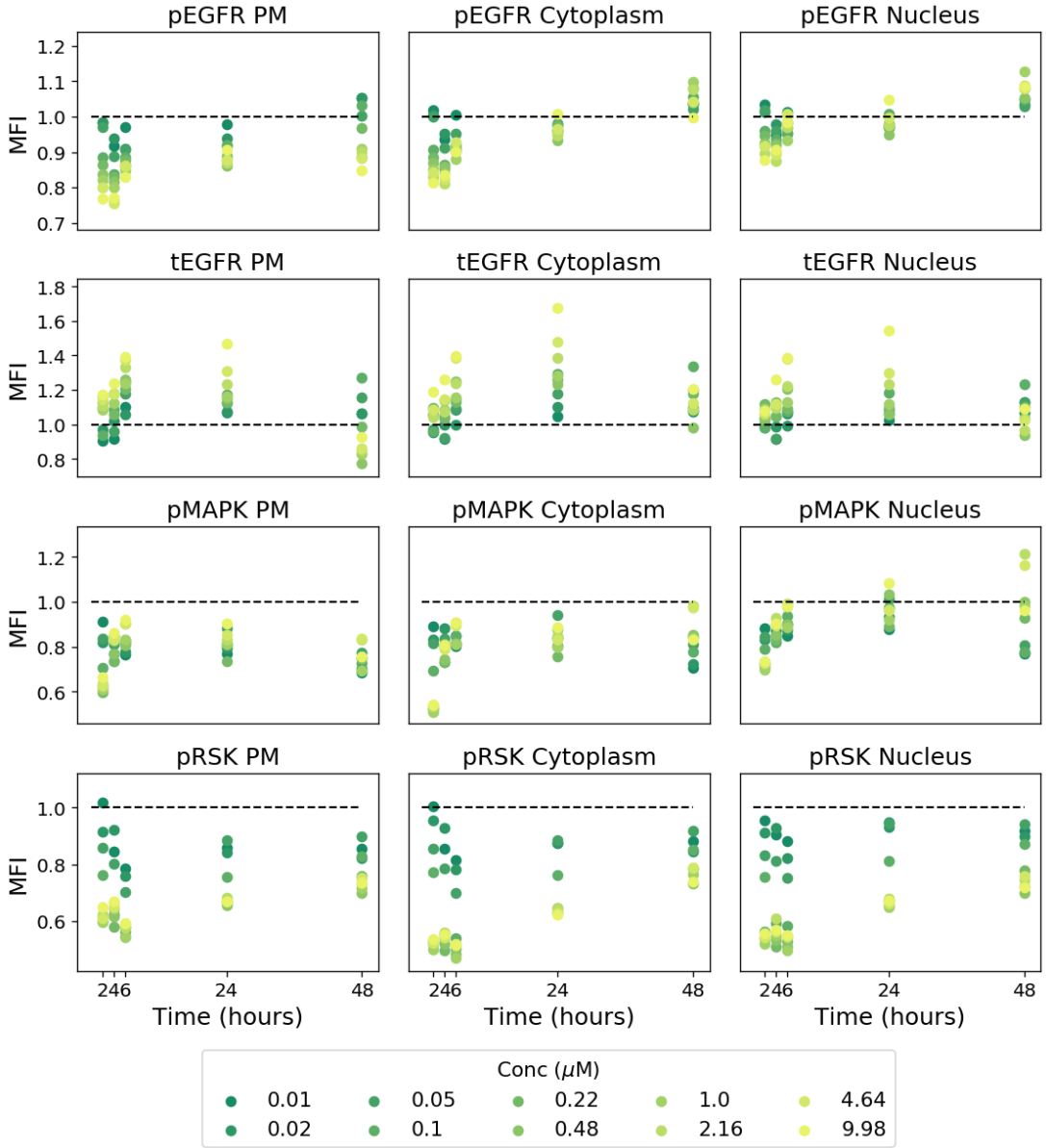

Figure E.8: Figure of the MFI data in the WT cell line under inhibition with inhibitor type D8. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

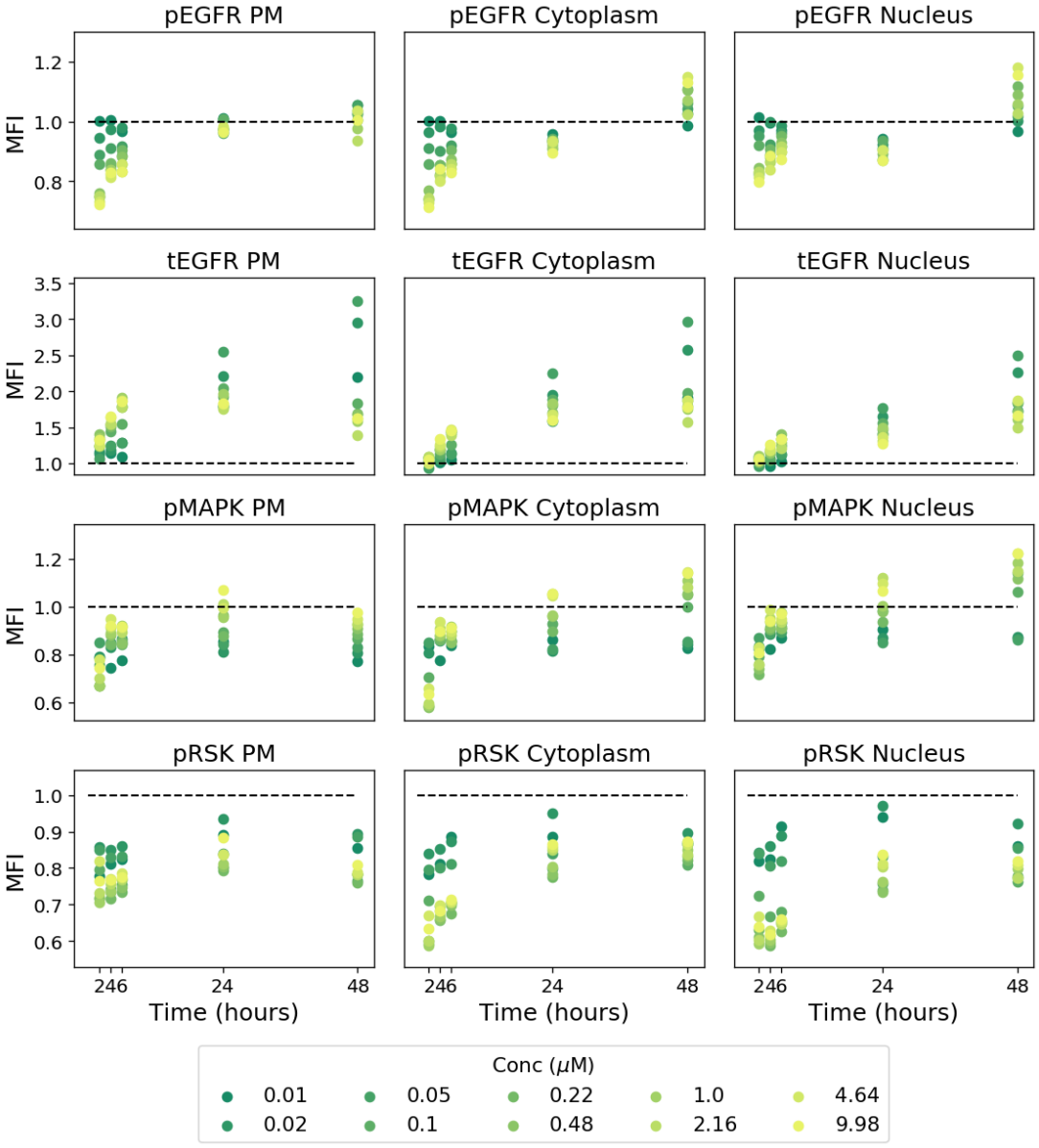

Figure E.9: Figure of the MFI data in the SVD cell line under inhibition with inhibitor type D1. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

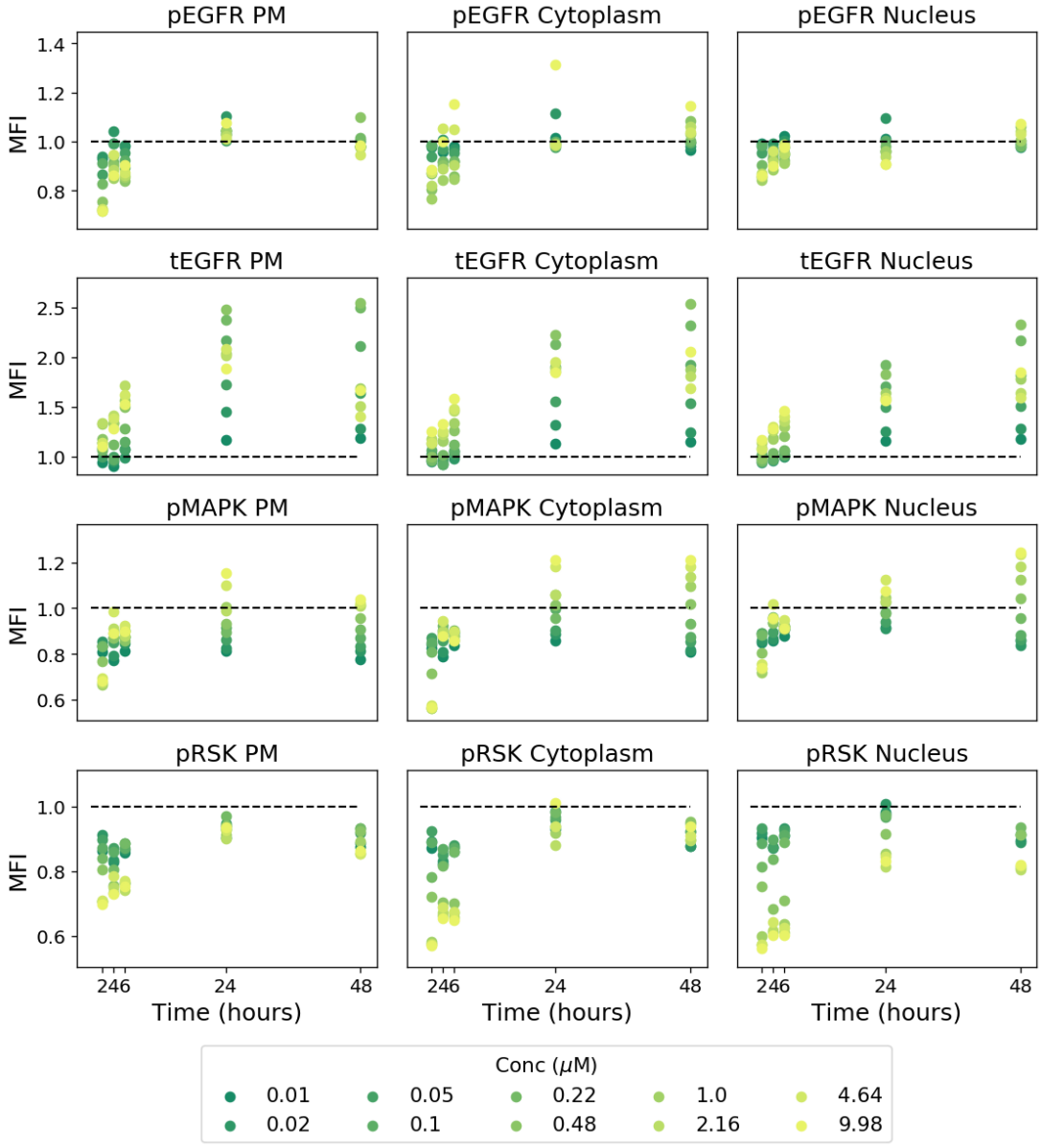

Figure E.10: Figure of the MFI data in the SVD cell line under inhibition with inhibitor type D2. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

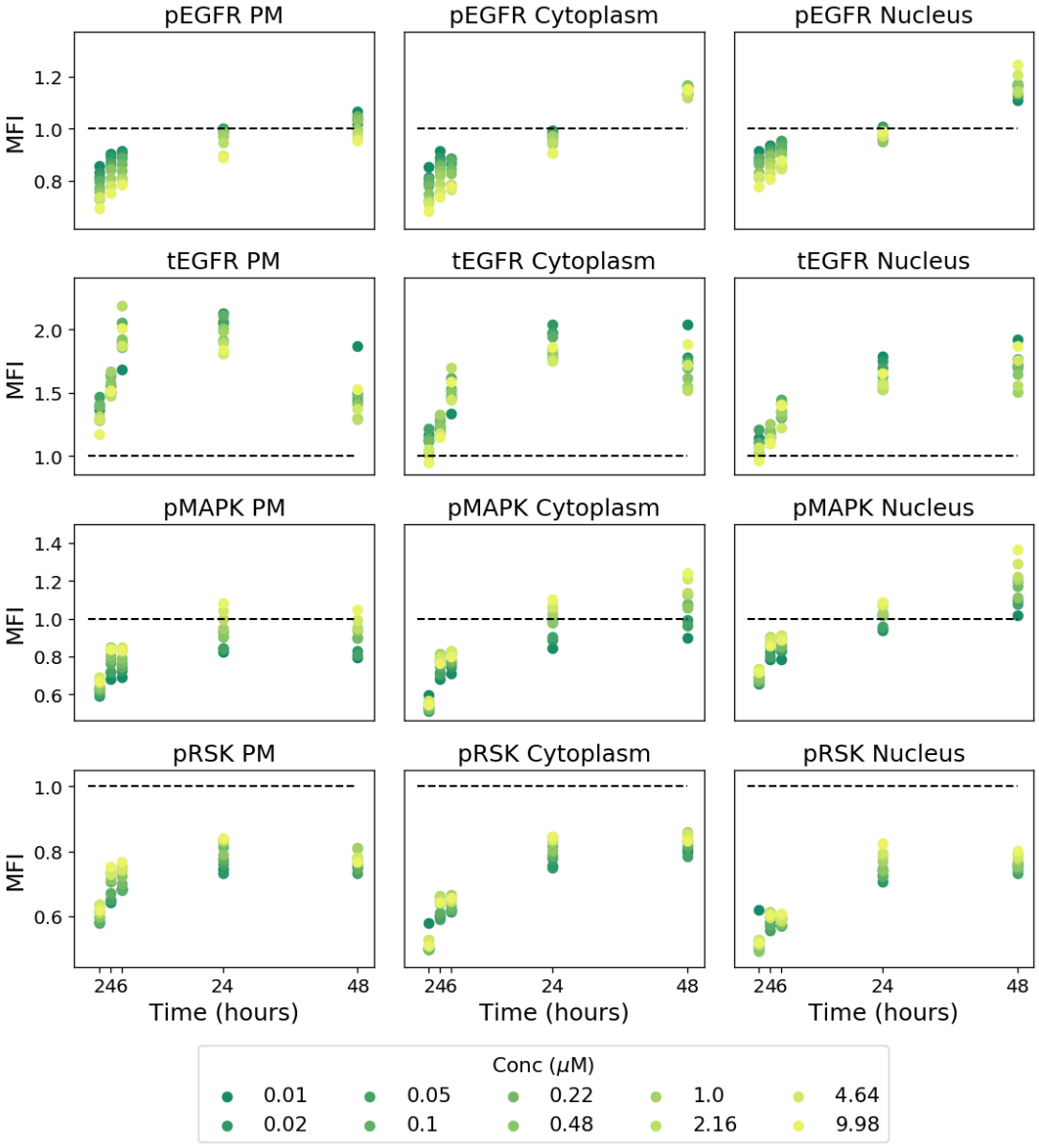

Figure E.11: Figure of the MFI data in the SVD cell line under inhibition with inhibitor type D3. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

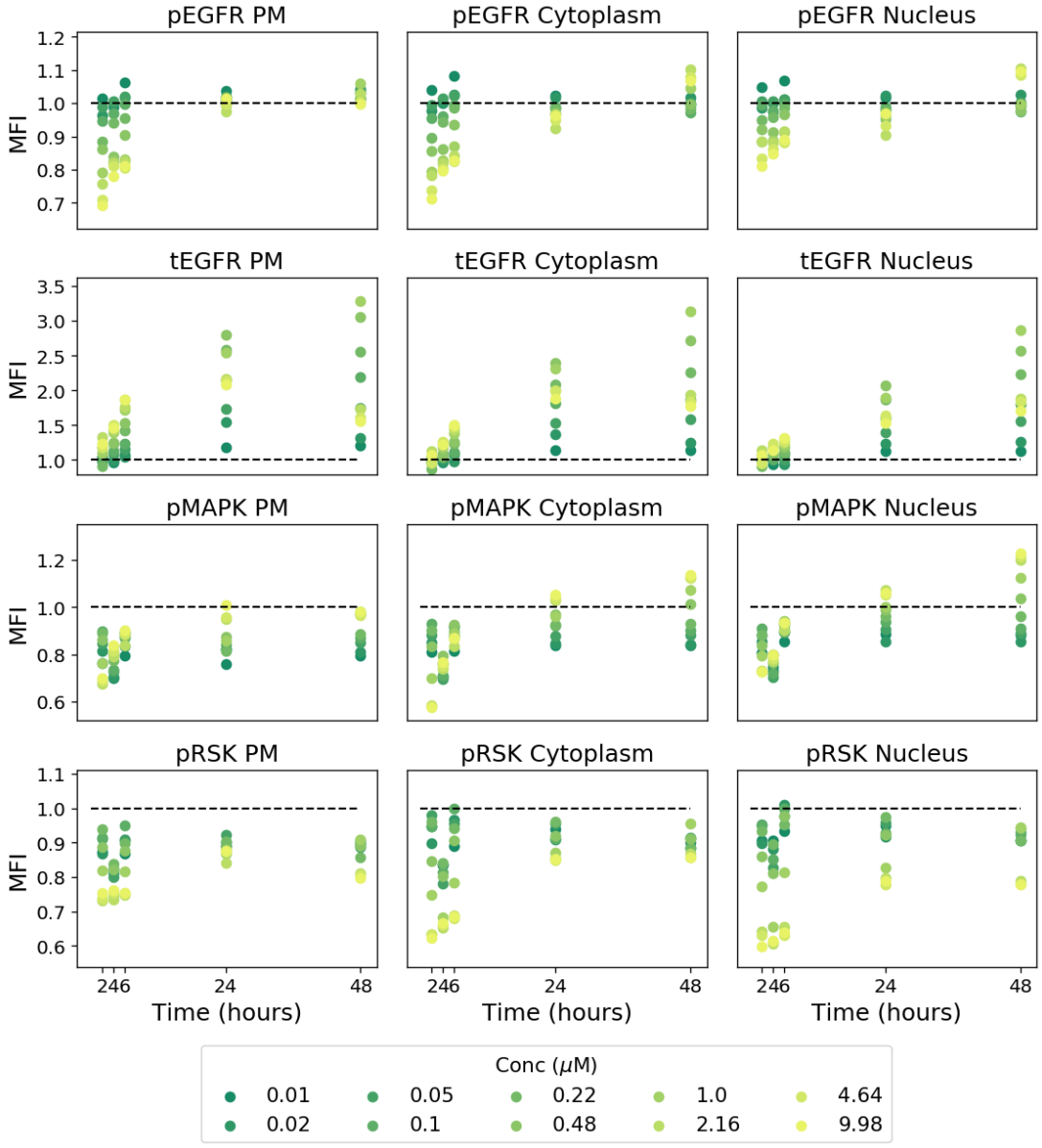

Figure E.12: Figure of the MFI data in the SVD cell line under inhibition with inhibitor type D4. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

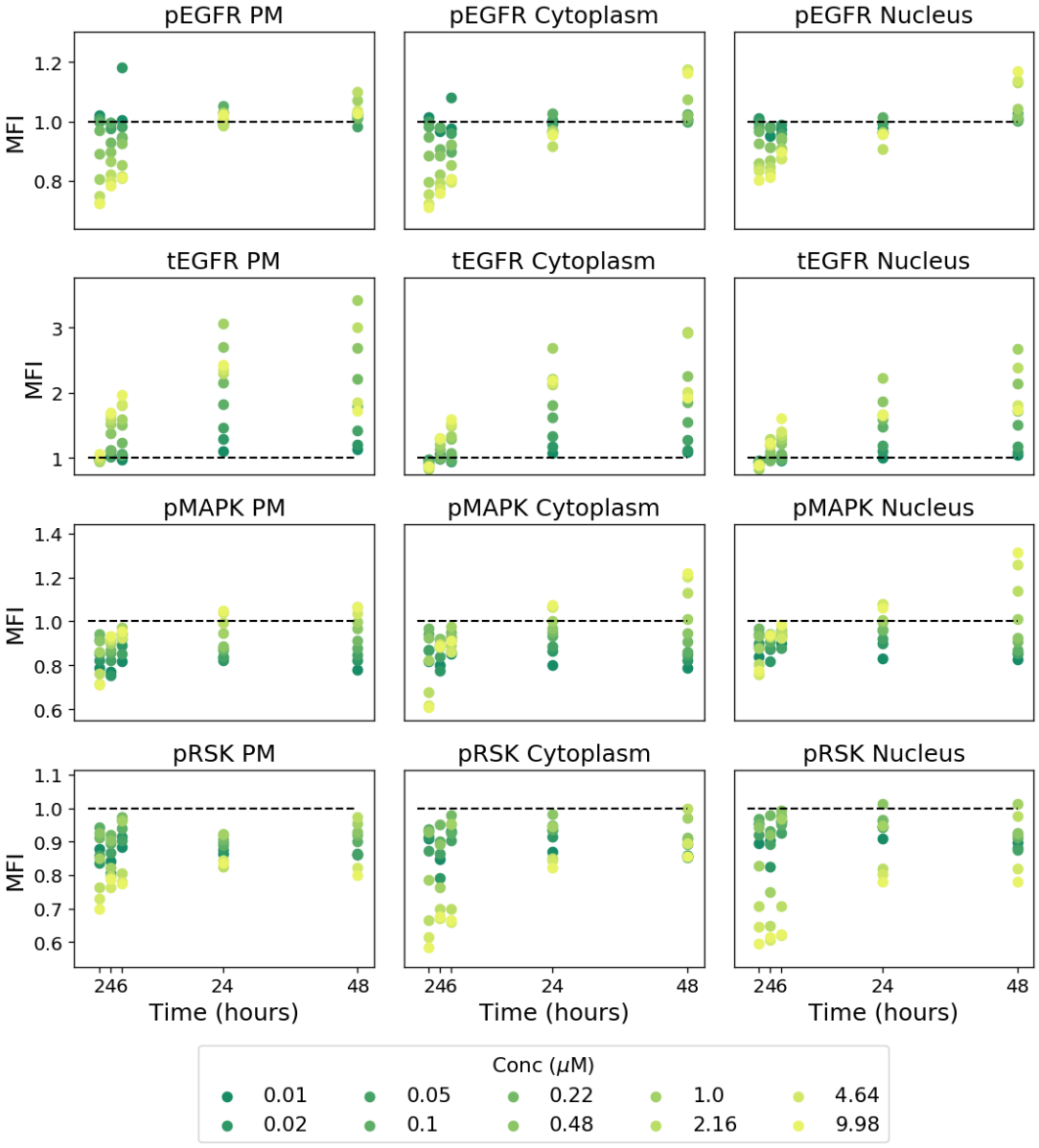

Figure E.13: Figure of the MFI data in the SVD cell line under inhibition with inhibitor type D5. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

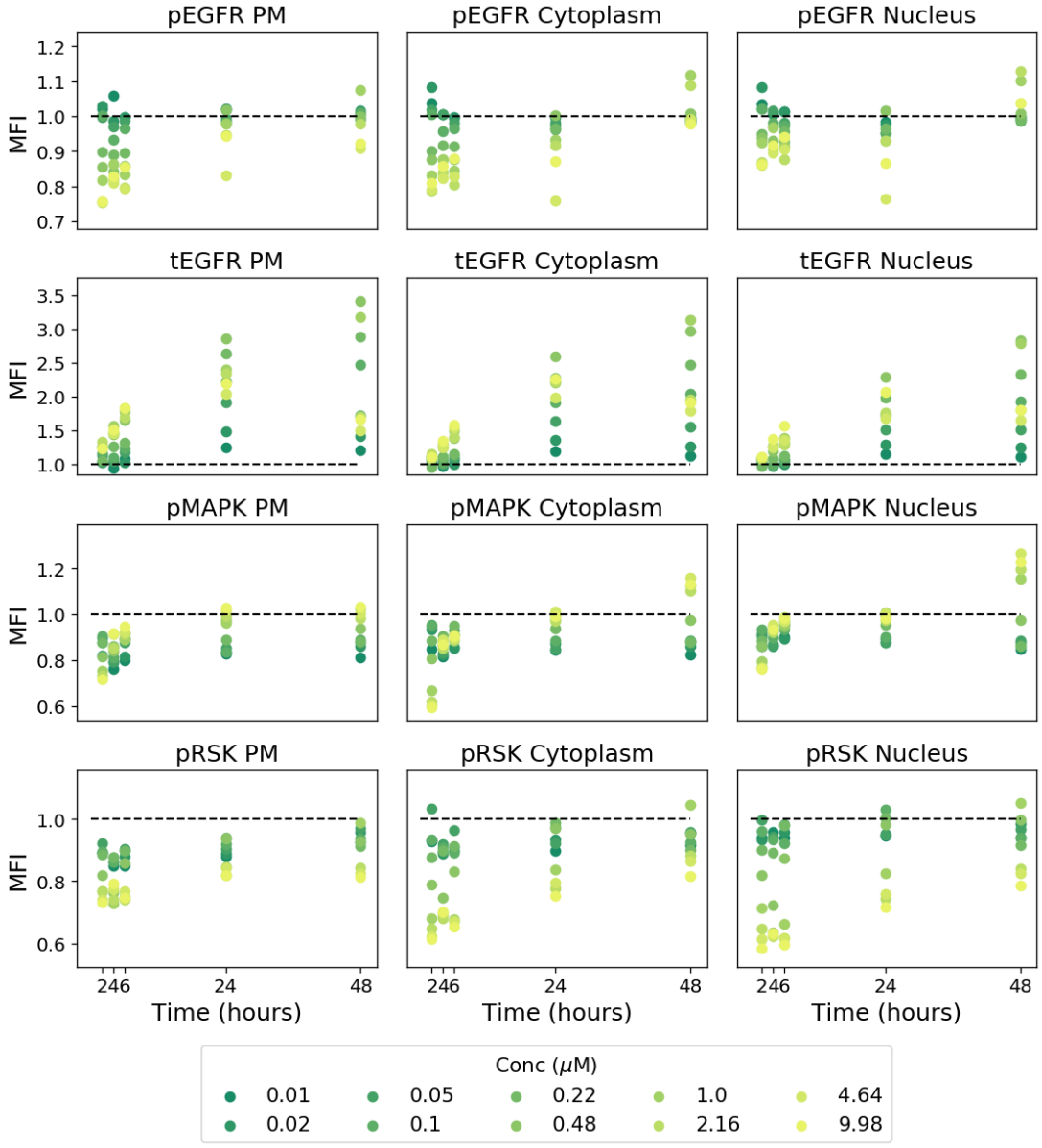

Figure E.14: Figure of the MFI data in the SVD cell line under inhibition with inhibitor type D6. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

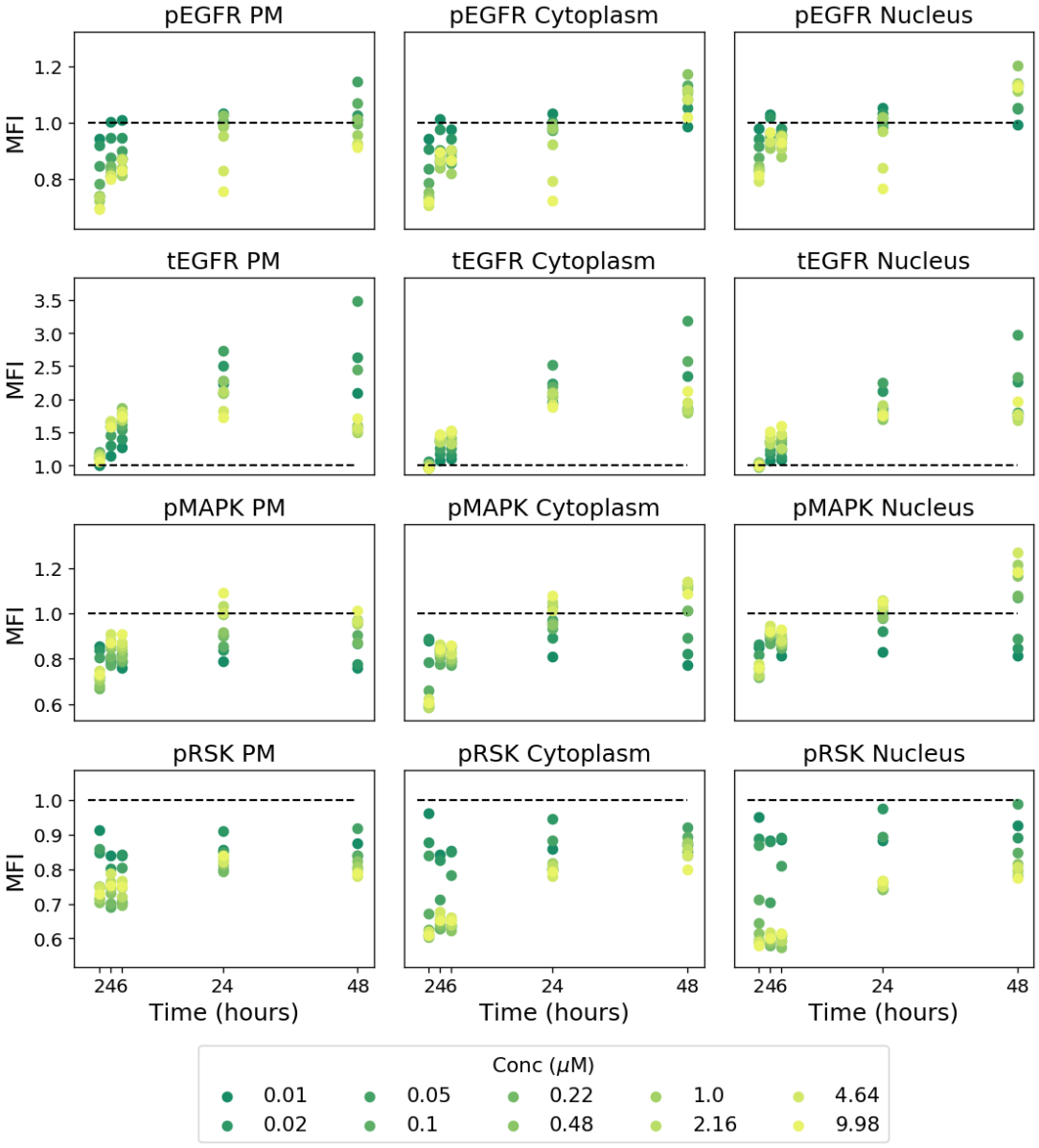

Figure E.15: Figure of the MFI data in the SVD cell line under inhibition with inhibitor type D7. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

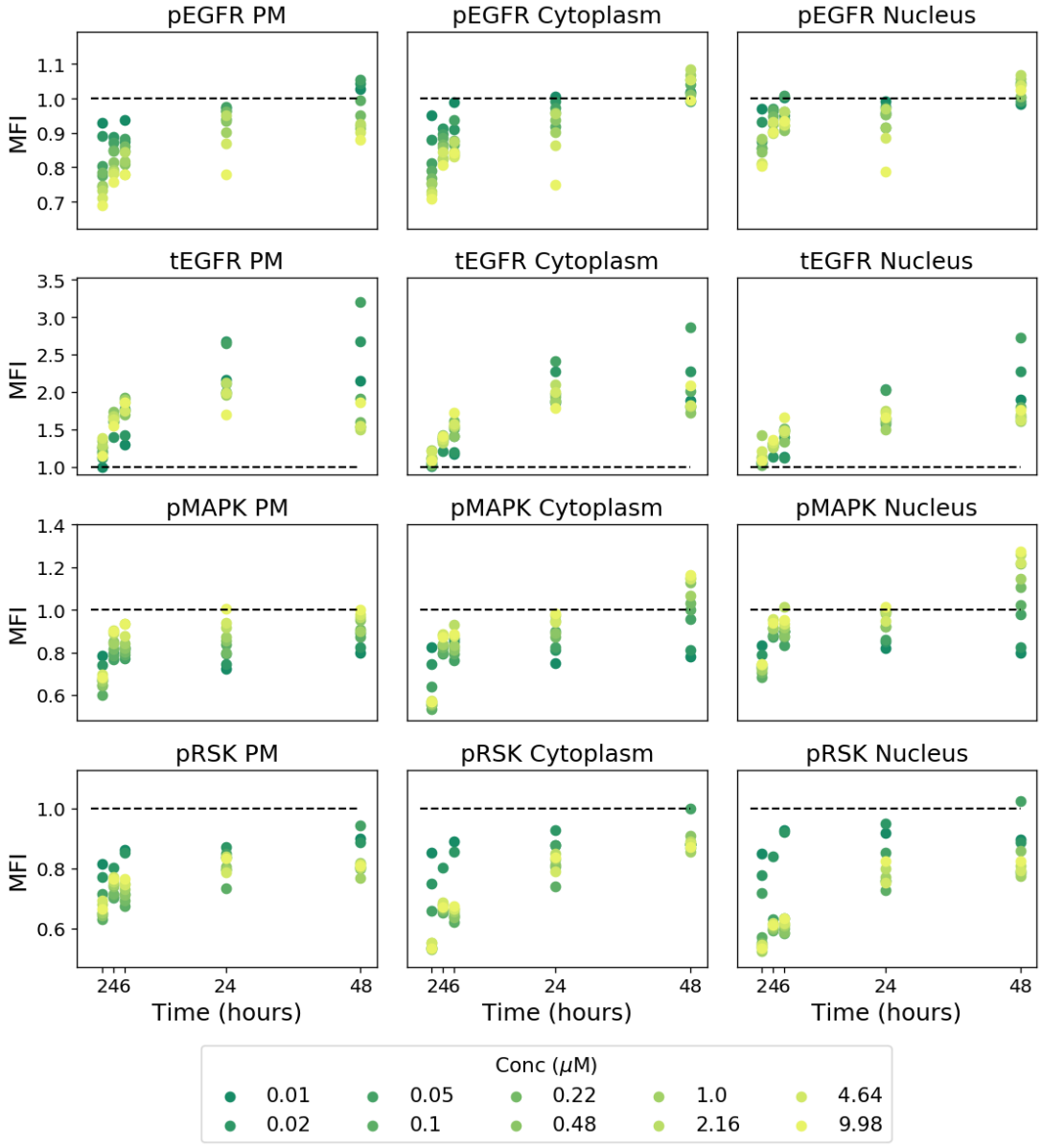

Figure E.16: Figure of the MFI data in the SVD cell line under inhibition with inhibitor type D8. The dashed line on each subplot represents the normalisation to the DMSO control and the colour of a point refers to the concentration of inhibitor with units μM as given in the legend.

References

Abdi, H. & Williams, L.J. (2010). Tukey's honestly significant difference (HSD) test. *Encyclopedia of research design*, **3**, 1–5. 44

Akdis, M., Burgler, S., Crameri, R., Eiwegger, T., Fujita, H., Gomez, E., Klunker, S., Meyer, N., O'Mahony, L., Palomares, O. *et al.* (2011). Interleukins, from 1 to 37, and interferon-γ: receptors, functions, and roles in diseases. *Journal of allergy and clinical immunology*, **127**, 701–721. 124

Al-Quteimat, O.M. & Amer, A.M. (2020). A review of Osimertinib in NSCLC and pharmacist role in NSCLC patient care. *Journal of Oncology Pharmacy Practice*, **26**, 1452–1460. 277

Alarcón, T. & Page, K.M. (2006). Stochastic models of receptor oligomerization by bivalent ligand. *Journal of The Royal Society Interface*, **3**, 545–559. 96

Allen, L.J. (2007). *Introduction to mathematical biology*. Pearson/Prentice Hall. 7, 25, 210

Allen, L.J. (2010). *An introduction to stochastic processes with applications to biology*. Chapman and Hall/CRC. 7, 11, 16, 18, 59, 63, 86, 87

Altan-Bonnet, G. & Mukherjee, R. (2019). Cytokine-mediated communication: a quantitative appraisal of immune complexity. *Nature reviews Immunology*, **19**, 205–217. 2

Ammar, A., Magnin, M., Roux, O., Cueto, E. & Chinesta, F. (2016). Chemical master equation empirical moment closure. 20

Araujo, R., Petricoin, E. & Liotta, L. (2005). A mathematical model of combination therapy using the EGFR signaling network. *Biosystems*, **80**, 57–69. 282

Arbouzova, N.I. & Zeidler, M.P. (2006). JAK/STAT signalling in Drosophila: insights into conserved regulatory and cellular functions. 127

Asthagiri, A.R. & Lauffenburger, D.A. (2001). A computational study of feedback effects on signal dynamics in a mitogen-activated protein kinase (MAPK) pathway model. *Biotechnology progress*, **17**, 227–239. 238

Babon, J.J., Kershaw, N.J., Murphy, J.M., Varghese, L.N., Laktyushin, A., Young, S.N., Lucet, I.S., Norton, R.S. & Nicola, N.A. (2012). Suppression of cytokine signaling by socs3: characterization of the mode of inhibition and the basis of its specificity. *Immunity*, **36**, 239–250. 183

Bade, B.C. & Cruz, C.S.D. (2020). Lung cancer 2020: epidemiology, etiology, and prevention. *Clinics in chest medicine*, **41**, 1–24. 238

Baraibar, I., Mezquita, L., Gil-Bazo, I. & Planchard, D. (2020). Novel drugs targeting EGFR and HER2 exon 20 mutations in metastatic NSCLC. *Critical Reviews in Oncology/Hematology*, **148**, 102906. 238

Barnette, J.J. & McLean, J.E. (2005). Type i error of four pairwise mean comparison procedures conducted as protected and unprotected tests. *Journal of Modern Applied Statistical Methods*, **4**, 10. 251

Bates, D. (2012). Bertini user's manual. *University of Notre Dame, Notre Dame, Ind, USA*. 213

Bates, D., Brake, D. & Niemerg, M. (2018). Paramotopy: Parameter homotopies in parallel. In *International Congress on Mathematical Software*, 28–35, Springer. xxii, 213, 216, 217

Bates, D.J., Sommese, A.J., Hauenstein, J.D. & Wampler, C.W. (2013). *Numerically solving polynomial systems with Bertini*. SIAM. xxii, 213, 215, 216

Beaumont, M.A., Zhang, W. & Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035. 33

Bethune, G., Bethune, D., Ridgway, N. & Xu, Z. (2010). Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. *Journal of thoracic disease*, **2**, 48. 6

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M.C., Tassani, S., Piva, F. *et al.* (2013). An estimation of the number of cells in the human body. *Annals of human biology*, **40**, 463–471. 1

Blätke, M.A., Dittrich, A., Rohr, C., Heiner, M., Schaper, F. & Marwan, W. (2013). Jak/stat signalling–an executable model assembled from molecule-centred modules demonstrating a module-oriented database concept for systems and synthetic biology. *Molecular BioSystems*, **9**, 1290–1307. 183

Blinov, M.L., Faeder, J.R., Goldstein, B. & Hlavacek, W.S. (2006). A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. *Biosystems*, **83**, 136–151. 238

Blitzstein, J.K. & Hwang, J. (2019). *Introduction to probability*. Crc Press. 32

Boulanger, M.J., Chow, D.c., Brevnova, E.E. & Garcia, K.C. (2003). Hexameric structure and assembly of the interleukin-6/IL-6 α-receptor/gp130 complex. *Science*, **300**, 2101–2104. 3

Brender, C., Tannahill, G.M., Jenkins, B.J., Fletcher, J., Columbus, R., Saris, C.J., Ernst, M., Nicola, N.A., Hilton, D.J., Alexander, W.S. *et al.* (2007). Suppressor of cytokine signaling 3 regulates CD8 T-cell proliferation by inhibition of interleukins 6 and 27. *Blood, The Journal of the American Society of Hematology*, **110**, 2528–2536. 127

Butcher, J.C. & Goodwin, N. (2008). *Numerical methods for ordinary differential equations*, vol. 2. Wiley Online Library. 215

Cameron, M.J. & Kelvin, D.J. (2013). Cytokines, chemokines and their receptors. In *Madame Curie Bioscience Database [Internet]*, Landes Bioscience. 2

Carpenter, G. & Cohen, S. (1976). 125I-labeled human epidermal growth factor. binding, internalization, and degradation in human fibroblasts. *The Journal of cell biology*, **71**, 159–171. 235

Casadei, C., Dizman, N., Schepisi, G., Cursano, M.C., Basso, U., Santini, D., Pal, S.K. & De Giorgi, U. (2019). Targeted therapies for advanced bladder cancer: new strategies with FGFR inhibitors. *Therapeutic advances in medical oncology*, **11**, 1758835919890285. 7

Case, L.B., Zhang, X., Ditlev, J.A. & Rosen, M.K. (2019). Stoichiometry controls activity of phase-separated clusters of actin signaling proteins. *Science*, **363**, 1093–1097. 197

Casella, G. & Berger, R.L. (2002). *Statistical inference*, vol. 2. Duxbury Pacific Grove, CA. 11

Castro, M. & de Boer, R.J. (2020). Testing structural identifiability by a simple scaling method. *PLOS Computational Biology*, **16**, e1008248. 156, 157, 297

Castro, M., van Santen, H.M., Férez, M., Alarcón, B., Lythe, G. & Molina-París, C. (2014). Receptor pre-clustering and T cell responses: insights into molecular mechanisms. *Frontiers in immunology*, **5**, 132. 95, 96

Cataldo, V.D., Gibbons, D.L., Pérez-Soler, R. & Quintás-Cardama, A. (2011). Treatment of non–small-cell lung cancer with erlotinib or gefitinib. *New England Journal of Medicine*, **364**, 947–955. 238

Cebecauer, M., Spitaler, M., Sergé, A. & Magee, A.I. (2010). Signalling complexes and clusters: functional advantages and methodological hurdles. *Journal of cell science*, **123**, 309–320. 196

Chatfield, C. & Collins, A. (1981). *Introduction to multivariate analysis*, vol. 1. CRC Press. 45, 266

Chen, S., Guo, X., Imarenezor, O. & Imoukhuede, P. (2015). Quantification of VEGFRs, NRP1, and PDGFRs on endothelial cells and fibroblasts reveals serum, intra-family ligand, and cross-family ligand regulation. *Cellular and Molecular Bioengineering*, **8**, 383–403. 47, 48

Chen, Y., Munteanu, A.C., Huang, Y.F., Phillips, J., Zhu, Z., Mavros, M. & Tan, W. (2009). Mapping receptor density on live cells by using fluorescence correlation spectroscopy. *Chemistry–A European Journal*, **15**, 5327–5336. 8

Claas, A.M., Atta, L., Gordonov, S., Meyer, A.S. & Lauffenburger, D.A. (2018a). Systems modeling identifies divergent receptor tyrosine kinase reprogramming to MAPK pathway inhibition. *Cellular and molecular bioengineering*, **11**, 451–469. 238

Claas, A.M., Atta, L., Gordonov, S., Meyer, A.S. & Lauffenburger, D.A. (2018b). Systems modeling identifies divergent receptor tyrosine kinase reprogramming to MAPK pathway inhibition. *Cellular and molecular bioengineering*, **11**, 451–469. 276

Croker, B.A., Krebs, D.L., Zhang, J.G., Wormald, S., Willson, T.A., Stanley, E.G., Robb, L., Greenhalgh, C.J., Förster, I., Clausen, B.E. *et al.* (2003). SOCS3 negatively regulates IL-6 signaling in vivo. *Nature immunology*, **4**, 540–545. 127

Cross, M.J., Dixelius, J., Matsumoto, T. & Claesson-Welsh, L. (2003). VEGF-receptor signal transduction. *Trends in biochemical sciences*, **28**, 488–494. 48, 49

Currie, J., Castro, M., Lythe, G., Palmer, E. & Molina-París, C. (2012). A stochastic T cell response criterion. *Journal of The Royal Society Interface*, **9**, 2856–2870. 95, 96

da Cunha Santos, G., Shepherd, F.A. & Tsao, M.S. (2011). EGFR mutations and lung cancer. *Annual Review of Pathology: Mechanisms of Disease*, **6**, 49–69. 237

Danos, V., Feret, J., Fontana, W., Harmer, R. & Krivine, J. (2007). Rule-based modelling of cellular signalling. In *International conference on concurrency theory*, 17–41, Springer. 186

Dikov, M.M., Ohm, J.E., Ray, N., Tchekneva, E.E., Burlison, J., Moghanaki, D., Nadaf, S. & Carbone, D.P. (2005). Differential roles of vascular endothelial growth factor receptors 1 and 2 in dendritic cell differentiation. *The Journal of Immunology*, **174**, 215–222. 48

Dinarello, C.A. (2007). Historical insights into cytokines. *European journal of immunology*, **37**, S34–S45. 2

Dittrich, A., Quaiser, T., Khouri, C., Görtz, D., Mönnigmann, M. & Schaper, F. (2012). Model-driven experimental analysis of the function of shp-2 in il-6-induced jak/stat signaling. *Molecular BioSystems*, **8**, 2119–2134. 184, 188

Dodington, D.W., Desai, H.R. & Woo, M. (2018). JAK/STAT–emerging players in metabolism. *Trends in Endocrinology & Metabolism*, **29**, 55–65. 125

Drexler, F.J. (1977). Eine methode zur berechnung sämtlicher lösungen von polynomgleichungssystemen. *Numerische Mathematik*, **29**, 45–58. 214

Elf, J. & Ehrenberg, M. (2003a). Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome research*, **13**, 2475–2484. 20

Elf, J. & Ehrenberg, M. (2003b). Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome research*, **13**, 2475–2484. 53

Emmanouilidi, A., Lattanzio, R., Sala, G., Piantelli, M. & Falasca, M. (2017). The role of phospholipase Cγ1 in breast cancer and its clinical significance. *Future Oncology*, **13**, 1991–1997. 196

Ernst, M. & Jenkins, B.J. (2004). Acquiring signalling specificity from the cytokine receptor gp130. *Trends in genetics*, **20**, 23–32. 47

Faeder, J.R., Blinov, M.L. & Hlavacek, W.S. (2009). Rule-based modeling of biochemical systems with bionetgen. In *Systems biology*, 113–167, Springer. 186

Fang, J.Y. & Richardson, B.C. (2005). The MAPK signalling pathways and colorectal cancer. *The lancet oncology*, **6**, 322–327. 5

Feinberg, M. (2019). *Foundations of chemical reaction network theory*. Springer. 25

Feinerman, O., Veiga, J., Dorfman, J.R., Germain, R.N. & Altan-Bonnet, G. (2008). Variability and robustness in T cell activation from regulated heterogeneity in protein levels. *Science*, **321**, 1081–1084. 8

Ferner, R.E. & Aronson, J.K. (2016). Cato Guldberg and Peter Waage, the history of the Law of Mass Action, and its relevance to clinical pharmacology. *British journal of clinical pharmacology*, **81**, 52–55. 25

Filippi, S., Barnes, C.P., Cornebise, J. & Stumpf, M.P. (2013). On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Statistical applications in genetics and molecular biology*, **12**, 87–107. 33, 35, 158

Flynn, C.M., Kespohl, B., Daunke, T., Garbers, Y., Düsterhöft, S., Rose-John, S., Haybaeck, J., Lokau, J., Aparicio-Siegmund, S. & Garbers, C. (2021). Interleukin-6 controls recycling and degradation, but not internalization of its receptors. *Journal of Biological Chemistry*, **296**. 180

Fritsche-Guenther, R., Witzel, F., Sieber, A., Herr, R., Schmidt, N., Braun, S., Brummer, T., Sers, C. & Blüthgen, N. (2011). Strong negative feedback from Erk to Raf confers robustness to MAPK signalling. *Molecular systems biology*, **7**, 489. 8

Garcia, C. & Zangwill, W.I. (1979). Finding all solutions to polynomial systems and other systems of equations. *Mathematical Programming*, **16**, 159–176. 214

Gaver, D., Jacobs, P. & Latouche, G. (1984). Finite birth-and-death models in randomly changing environments. *Advances in applied probability*, **16**, 715–731. 61, 62

Ghedini, G.C., Ronca, R., Presta, M. & Giacomini, A. (2018). Future applications of FGF/FGFR inhibitors in cancer. *Expert review of anticancer therapy*, **18**, 861–872. 7

Gillespie, D.T. (1976a). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, **22**, 403–434. 23

Gillespie, D.T. (1976b). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, **22**, 403–434. 57

Gillespie, D.T. (1992). A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, **188**, 404–425. 52

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. 122

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. 122

Gómez-Corral, A. & López-García, M. (2015). Lifetime and reproduction of a marked individual in a two-species competition process. *Applied Mathematics and Computation*, **264**, 223–245. 122

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. 62

Gordziel, C., Bratsch, J., Moriggl, R., Knösel, T. & Friedrich, K. (2013). Both STAT1 and STAT3 are favourable prognostic determinants in colorectal carcinoma. *British journal of cancer*, **109**, 138–146. 123

Griffiths, D.F. & Higham, D.J. (2010). Euler's method. In *Numerical Methods for Ordinary Differential Equations*, 19–31, Springer. 26

Grötzinger, J. (2002). Molecular mechanisms of cytokine receptor activation. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, **1592**, 215–223. 1, 2

Günther, M., Juchum, M., Kelter, G., Fiebig, H. & Laufer, S. (2016). Lung cancer: EGFR inhibitors with low nanomolar activity against a therapy-resistant L858R/T790M/C797S mutant. *Angewandte Chemie International Edition*, **55**, 10890–10894. 238

Gurevich, K.G., Agutter, P.S. & Wheatley, D.N. (2003). Stochastic description of the ligand–receptor interaction of biologically active substances at extremely low doses. *Cellular signalling*, **15**, 447–453. 48

Haan, C., Kreis, S., Margue, C. & Behrmann, I. (2006). Jaks and cytokine receptors—an intimate relationship. *Biochemical pharmacology*, **72**, 1538–1546. xi, 1, 2, 3

Hackel, P.O., Zwick, E., Prenzel, N. & Ullrich, A. (1999). Epidermal growth factor receptors: critical mediators of multiple receptor pathways. *Current opinion in cell biology*, **11**, 184–189. 47

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. 24

Hallinan, N., Finn, S., Cuffe, S., Rafee, S., O'Byrne, K. & Gately, K. (2016). Targeting the fibroblast growth factor receptor family in cancer. *Cancer treatment reviews*, **46**, 51–62. 195

Handel, A., La Gruta, N.L. & Thomas, P.G. (2020). Simulation modelling for immunologists. *Nature Reviews Immunology*, **20**, 186–195. 25

Harris, R.C., Chung, E. & Coffey, R.J. (2003). EGF receptor ligands. In *The EGF Receptor Family*, 3–14, Elsevier. 235

Hayot, F. & Jayaprakash, C. (2004). The linear noise approximation for molecular fluctuations within cells. *Physical Biology*, **1**, 205. 20, 53

He, Q.M. (2014). *Fundamentals of matrix-analytic methods*, vol. 365. Springer. 16

Heinrich, P.C., Behrmann, I., Müller-Newen, G., Schaper, F. & Graeve, L. (1998). Interleukin-6-type cytokine signalling through the gp130/jak/STAT pathway. *Biochemical journal*, **334**, 297–314. 123

Herbst, R.S. (2004). Review of epidermal growth factor receptor biology. *International Journal of Radiation Oncology* Biology* Physics*, **59**, S21–S26. 235

Hibbert, L., Pflanz, S., de Waal Malefyt, R. & Kastelein, R.A. (2003). IL-27 and IFN-α signal via Stat1 and Stat3 and induce T-Bet and IL-12R β 2 in naive T cells. *Journal of interferon & cytokine research*, **23**, 513–522. 123

Higham, D.J. (2008). Modeling and simulating chemical reactions. *SIAM review*, **50**, 347–368. 25

Hodge, J.A., Kawabata, T.T., Krishnaswami, S., Clark, J.D., Telliez, J.B., Dowty, M.E., Menon, S., Lamba, M. & Zwillich, S. (2016). The mechanism of action of tofacitinib-an oral janus kinase inhibitor for the treatment of rheumatoid arthritis. *Clin Exp Rheumatol*, **34**, 318–328. 149

Homma, T. & Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, **52**, 1–17. 30, 31

Hornberg, J.J., Binder, B., Bruggeman, F.J., Schoeberl, B., Heinrich, R. & Westerhoff, H.V. (2005). Control of MAPK signalling: from complexity to what really matters. *Oncogene*, **24**, 5533–5542. 283

Huang, L. & Fu, L. (2015). Mechanisms of resistance to EGFR tyrosine kinase inhibitors. *Acta Pharmaceutica Sinica B*, **5**, 390–401. 6

Huang, L., Jiang, Y. & Chen, Y. (2017). Predicting drug combination index and simulating the network-regulation dynamics by mathematical modeling of drug-targeted EGFR-ERK signaling pathway. *Scientific reports*, **7**, 40752. xxviii, 238, 282, 283, 284

Huang, W.Y., Alvarez, S., Kondo, Y., Lee, Y.K., Chung, J.K., Lam, H.Y.M., Biswas, K.H., Kuriyan, J. & Groves, J.T. (2019). A molecular assembly phase transition and kinetic proofreading modulate Ras activation by SOS. *Science*, **363**, 1098–1103. 197

Hulme, E.C. & Trevethick, M.A. (2010). Ligand binding assays at equilibrium: validation and interpretation. *British journal of pharmacology*, **161**, 1219–1237. 66

Hunter, C.A. & Jones, S.A. (2015). Il-6 as a keystone cytokine in health and disease. *Nature immunology*, **16**, 448–457. 2

Iglehart, D.L. *et al.* (1964). Multivariate competition processes. *The Annals of Mathematical Statistics*, **35**, 350–361. 48, 117

Itzhak, D.N., Tyanova, S., Cox, J. & Borner, G.H. (2016). Global, quantitative and dynamic mapping of protein subcellular localization. *elife*, **5**, e16950. 155

Johnson, K.A. & Goody, R.S. (2011). The original Michaelis constant: translation of the 1913 Michaelis–Menten paper. *Biochemistry*, **50**, 8264–8269. 277, 278

Kang, S. & Cheek, J.B. (1972). *Numerical solution of differential equations*. Waterways Experiment Station. 26

Kassambara, A. (2017). *Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra*, vol. 2. Sthda. 45, 266

Katoh, M. & Katoh, M. (2006). FGF signaling network in the gastrointestinal tract. *International journal of oncology*, **29**, 163–168. 47

Kershaw, N.J., Murphy, J.M., Liau, N.P., Varghese, L.N., Laktyushin, A., Whitlock, E.L., Lucet, I.S., Nicola, N.A. & Babon, J.J. (2013). Socs3 binds specific receptor–jak complexes to control cytokine signaling by direct kinase inhibition. *Nature structural & molecular biology*, **20**, 469–476. 183

Kholodenko, B.N., Demin, O.V., Moehren, G. & Hoek, J.B. (1999). Quantification of short term signaling by the epidermal growth factor receptor. *Journal of Biological Chemistry*, **274**, 30169–30181. 276, 282

Klein, P., Mattoon, D., Lemmon, M.A. & Schlessinger, J. (2004). A structure-based model for ligand binding and dimerization of EGF receptors. *Proceedings of the National Academy of Sciences*, **101**, 929–934. 238

Klinger, B., Sieber, A., Fritsche-Guenther, R., Witzel, F., Berry, L., Schumacher, D., Yan, Y., Durek, P., Merchant, M., Schäfer, R. *et al.* (2013). Network quantification of EGFR signaling unveils potential for targeted combination therapy. *Molecular systems biology*, **9**, 673. 285

Kozer, N., Barua, D., Orchard, S., Nice, E.C., Burgess, A.W., Hlavacek, W.S. & Clayton, A.H. (2013a). Exploring higher-order EGFR oligomerisation and phosphorylation—a combined experimental and theoretical approach. *Molecular bioSystems*, **9**, 1849–1863. 148

Kozer, N., Barua, D., Orchard, S., Nice, E.C., Burgess, A.W., Hlavacek, W.S. & Clayton, A.H. (2013b). Exploring higher-order EGFR oligomerisation and phosphorylation—a combined experimental and theoretical approach. *Molecular bioSystems*, **9**, 1849–1863. 238

Kulkarni, V.G. (2016). *Modeling and analysis of stochastic systems*. Chapman and Hall/CRC. 16, 60

Lake, D., Corrêa, S.A. & Müller, J. (2016). Negative feedback regulation of the ERK1/2 MAPK pathway. *Cellular and Molecular Life Sciences*, **73**, 4397–4413. xxiii, 236, 237

Latouche, G., Ramaswami, V. & Kulkarni, V. (1999). Introduction to matrix analytic methods in stochastic modeling. *Journal of Applied Mathematics and Stochastic Analysis*, **12**, 435–436. 16, 49, 59, 118

Lauffenburger, D.A. & Linderman, J.J. (1996). *Receptors: models for binding, trafficking, and signaling*. Oxford University Press on Demand. 181

Lee, S. & Lee, D.K. (2018). What is the proper way to apply the multiple comparison test? *Korean journal of anesthesiology*, **71**, 353. 252

Lemmon, M.A. & Schlessinger, J. (2010). Cell signaling by receptor tyrosine kinases. *Cell*, **141**, 1117–1134. 47

Lemmon, M.A., Schlessinger, J. & Ferguson, K.M. (2014). The EGFR family: not so prototypical receptor tyrosine kinases. *Cold Spring Harbor perspectives in biology*, **6**, a020768. 235

Leon, A.J., Gomez, E., Garrote, J.A., Bernardo, D., Barrera, A., Marcos, J.L., Fernández-Salazar, L., Velayos, B., Blanco-Quirós, A. & Arranz, E. (2009). High levels of proinflammatory cytokines, but not markers of tissue injury, in unaffected intestinal areas from patients with IBD. *Mediators of inflammation*, **2009**. 6

Li, P., Banjade, S., Cheng, H.C., Kim, S., Chen, B., Guo, L., Llaguno, M., Hollingsworth, J.V., King, D.S., Banani, S.F. *et al.* (2012). Phase transitions in the assembly of multivalent signalling proteins. *Nature*, **483**, 336–340. 197

Li, T.Y. (1997). Numerical solution of multivariate polynomial systems by homotopy continuation methods. *Acta numerica*, **6**, 399–436. 213

Lin, C.C., Suen, K.M., Jeffrey, P.A., Wieteska, L., Stainthorp, A., Seiler, C., Koss, H., Molina-París, C., Miska, E., Ahmed, Z. *et al.* (2019). Receptor tyrosine kinases regulate signal transduction through a liquid–liquid phase separated state. *bioRxiv*, 783720. xxi, 199, 200

Liu, F., Yang, X., Geng, M. & Huang, M. (2018). Targeting ERK, an Achilles' heel of the MAPK pathway, in cancer therapy. *Acta pharmaceutica sinica B*, **8**, 552–562. xi, 4, 5

López-García, M., Nowicka, M., Bendtsen, C., Lythe, G., Ponnambalam, S. & Molina-Paris, C. (2016). Stochastic models of the binding kinetics of VEGF-A to VEGFR1 and VEGFR2 in endothelial cells. *arXiv preprint arXiv:1606.07269*. 76, 118

López-García, M., Nowicka, M., Bendtsen, C., Lythe, G., Ponnambalam, S. & Molina-París, C. (2018). Quantifying the phosphorylation timescales of receptor–ligand complexes: a markovian matrix-analytic approach. *Open biology*, **8**, 180126. 49, 96

Marijanovic, Z., Ragimbeau, J., Van der Heyden, J., Uzé, G. & Pellegrini, S. (2007). Comparable potency of ifnα2 and ifnβ on immediate jak/-stat activation but differential down-regulation of ifnar2. *Biochemical Journal*, **407**, 141–151. 180

Marques de Sá, J.P. (2003). Applied statistics using SPSS, STATISTICA and MATLAB. 39, 246

McKay, M. & Morrison, D. (2007). Integrating signals from RTKs to ERK/MAPK. *Oncogene*, **26**, 3113–3121. 236

Meier-Schellersheim, M. & Mack, G. (1999). Simmune, a tool for simulating and analyzing immune system behavior. *arXiv preprint cs/9903017*. 186

Mendes, P., Hoops, S., Sahle, S., Gauges, R., Dada, J. & Kummer, U. (2009). Computational modeling of biochemical networks using copasi. In *Systems Biology*, 17–59, Springer. 186

Misale, S., Bozic, I., Tong, J., Peraza-Penton, A., Lallo, A., Baldi, F., Lin, K.H., Truini, M., Trusolino, L., Bertotti, A. *et al.* (2015). Vertical suppression of the EGFR pathway prevents onset of resistance in colorectal cancers. *Nature communications*, **6**, 1–9. 283, 285

Mokhtari, R.B., Homayouni, T.S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B. & Yeger, H. (2017). Combination therapy in combating cancer. *Oncotarget*, **8**, 38022. 281

Montgomery, D.C. (2017). *Design and analysis of experiments*. John wiley & sons. 44

Morris, R., Kershaw, N.J. & Babon, J.J. (2018). The molecular details of cytokine signaling via the JAK/STAT pathway. *Protein Science*, **27**, 1984–2009. 126

Murphy, G.M. (2011). *Ordinary differential equations and their solutions*. Courier Corporation. 26

Murphy, M., Jason-Moller, L. & Bruno, J. (2006). Using biacore to measure the binding kinetics of an antibody-antigen interaction. *Current protocols in protein science*, **45**, 19–14. 147

Nazari, F., Pearson, A.T., Nör, J.E. & Jackson, T.L. (2018). A mathematical model for il-6-mediated, stem cell driven tumor growth and targeted treatment. *PLoS computational biology*, **14**, e1005920. 181

Nicholson, R.I., Gee, J.M.W. & Harper, M.E. (2001). EGFR and cancer prognosis. *European journal of cancer*, **37**, 9–15. 236

Nishimoto, N. & Kishimoto, T. (2004). Inhibition of IL-6 for the treatment of inflammatory diseases. *Current opinion in pharmacology*, **4**, 386–391. 192

Niu, W. & Wang, D. (2008). Algebraic approaches to stability analysis of biological systems. *Mathematics in Computer Science*, **1**, 507–539. 224

Normanno, N., De Luca, A., Bianco, C., Strizzi, L., Mancino, M., Maiello, M.R., Carotenuto, A., De Feo, G., Caponigro, F. & Salomon, D.S. (2006). Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene*, **366**, 2–16. 6, 236

Ohl, K. & Tenbrock, K. (2011). Inflammatory cytokines in systemic lupus erythematosus. *Journal of Biomedicine and Biotechnology*, **2011**. 6

Ornitz, D.M. & Itoh, N. (2015). The fibroblast growth factor signaling pathway. *Wiley Interdisciplinary Reviews: Developmental Biology*, **4**, 215–266. 6, 195

O'Shea, J.J. & Plenge, R. (2012). JAK and STAT signaling molecules in immunoregulation and immune-mediated disease. *Immunity*, **36**, 542–550. 175

Peck, R., Olsen, C. & Devore, J.L. (2015). *Introduction to statistics and data analysis*. Cengage Learning. 165

Petes, C., Mariani, M.K., Yang, Y., Grandvaux, N. & Gee, K. (2018). Interleukin (IL)-6 inhibits IL-27-and IL-30-mediated inflammatory responses in human monocytes. *Frontiers in immunology*, **9**, 256. 123

Pierce, K.L., Premont, R.T. & Lefkowitz, R.J. (2002). Seven-transmembrane receptors. *Nature reviews Molecular cell biology*, **3**, 639–650. 47

Pinsky, M. & Karlin, S. (2010). *An introduction to stochastic modeling*. Academic press. 11, 16, 18, 58, 81

Porta, R., Borea, R., Coelho, A., Khan, S., Araújo, A., Reclusa, P., Franchina, T., Van Der Steen, N., Van Dam, P., Ferri, J. *et al.* (2017). FGFR a promising druggable target in cancer: Molecular biology and new drugs. *Critical reviews in oncology/hematology*, **113**, 256–267. 7

Pratilas, C.A. & Solit, D.B. (2010). Targeting the mitogen-activated protein kinase pathway: physiological feedback and drug response. *Clinical Cancer Research*, **16**, 3329–3334. xxiii, 237

Puck, T.T., Marcus, P.I. & Cieciura, S.J. (1956). Clonal growth of mammalian cells in vitro growth characteristics of colonies from single hela cells with and without a feeder layer. *Journal of Experimental Medicine*, **103**, 273–284. 153

Qu, C.K. (2000). The SHP-2 tyrosine phosphatase: signaling mechanisms and biological functions. *Cell research*, **10**, 279–288. 196

Reeh, H., Rudolph, N., Billing, U., Christen, H., Streif, S., Bullinger, E., Schliemann-Bullinger, M., Findeisen, R., Schaper, F., Huber, H.J. *et al.* (2019). Response to il-6 trans-and il-6 classic signalling is determined by the ratio of the il-6 receptor α to gp130 expression: fusing experimental insights and dynamic modelling. *Cell Communication and Signaling*, **17**, 1–21. 183, 184, 185, 186, 190

Reuter, G. (1961). Competition processes. In *Proc. 4th Berkeley Symp. Math. Statist. Prob*, vol. 2, 421–430. 48, 117

Roberts, P.J. & Der, C.J. (2007). Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene*, **26**, 3291–3310. 237

Rock, K., Brand, S., Moir, J. & Keeling, M.J. (2014). Dynamics of infectious diseases. *Reports on Progress in Physics*, **77**, 026602. 25

Roepstorff, K., Grandal, M.V., Henriksen, L., Knudsen, S.L.J., Lerdrup, M., Grøvdal, L., Willumsen, B.M. & Van Deurs, B. (2009). Differential effects of EGFR ligands on endocytic sorting of the receptor. *Traffic*, **10**, 1115–1127. 121

Rogers, A. & Gibon, Y. (2009). Enzyme kinetics: Theory and practice. *Plant Metabolic Networks*, 71–103. xxviii, 278, 280

Rose-John, S. (2018). Interleukin-6 family cytokines. *Cold Spring Harbor perspectives in biology*, **10**, a028415. 2

Rottenberg, M.E. & Carow, B. (2014). Socs3, a major regulator of infection and inflammation. *Frontiers in immunology*, **5**, 58. 183

Santos, S.D., Verveer, P.J. & Bastiaens, P.I. (2007). Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nature cell biology*, **9**, 324–330. 47

Sarabipour, S. (2017). Parallels and distinctions in FGFR, VEGFR, and EGFR mechanisms of transmembrane signaling. *Biochemistry*, **56**, 3159–3173. 6

Schelp, C. (2018). An alternative way to plot the covariance ellipse. https://carstenschelp.github.io/2018/09/14/Plot_Confidence_Ellipse_001.html, [Online; accessed 18-Aug-2021]. 79

Schlessinger, J. (2002). Ligand-induced, receptor-mediated dimerization and activation of EGF receptor. *Cell*, **110**, 669–672. 235

Schlessinger, J. & Ullrich, A. (1992). Growth factor signaling by receptor tyrosine kinases. *Neuron*, **9**, 383–391. 1

Schoeberl, B., Eichler-Jonsson, C., Gilles, E.D. & Müller, G. (2002). Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nature biotechnology*, **20**, 370–375. 276

Schwartz, P.A., Kuzmic, P., Solowiej, J., Bergqvist, S., Bolanos, B., Almaden, C., Nagata, A., Ryan, K., Feng, J., Dalvie, D. *et al.* (2014). Covalent EGFR inhibitor analysis reveals importance of reversible interactions to potency and mechanisms of drug resistance. *Proceedings of the National Academy of Sciences*, **111**, 173–178. 280, 281

Sender, R. & Milo, R. (2021). The distribution of cellular turnover in the human body. *Nature Medicine*, **27**, 45–48. 1

Shankaran, H., Zhang, Y., Opresko, L. & Resat, H. (2008). Quantifying the effects of co-expressing EGFR and HER2 on HER activation and trafficking. *Biochemical and biophysical research communications*, **371**, 220–224. 276

Sharma, S.V., Bell, D.W., Settleman, J. & Haber, D.A. (2007). Epidermal growth factor receptor mutations in lung cancer. *Nature Reviews Cancer*, **7**, 169–181. 6, 238

Shibuya, M. (2006). Differential roles of vascular endothelial growth factor receptor-1 and receptor-2 in angiogenesis. *BMB Reports*, **39**, 469–478. 48, 49

Shuai, K. & Liu, B. (2003). Regulation of JAK–STAT signalling in the immune system. *Nature Reviews Immunology*, **3**, 900–911. 127

Simola, U., Cisewski-Kehe, J., Gutmann, M.U., Corander, J. *et al.* (2020). Adaptive approximate Bayesian computation tolerance selection. *Bayesian Analysis*. 33

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. 7

Sobol, I. (1993). Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput. Exp*, **1**, 407–414. 30, 31

Sobotta, S., Raue, A., Huang, X., Vanlier, J., Jünger, A., Bohl, S., Albrecht, U., Hahnel, M.J., Wolf, S., Mueller, N.S. *et al.* (2017). Model based targeting of il-6-induced inflammatory responses in cultured primary hepatocytes to improve application of the jak inhibitor ruxolitinib. *Frontiers in physiology*, **8**, 775. 183

Sommese, A.J., Verschelde, J. & Wampler, C.W. (2005). Introduction to numerical algebraic geometry. In *Solving polynomial equations*, 301–337, Springer. 213

Starbuck, C. & Lauffenburger, D.A. (1992). Mathematical model for the effects of epidermal growth factor receptor trafficking dynamics on fibroblast proliferation responses. *Biotechnology progress*, **8**, 132–143. 47

Steinfeld, J.I., Francisco, J.S. & Hase, W.L. (1989). *Chemical kinetics and dynamics*, vol. 3. Prentice Hall Englewood Cliffs (New Jersey). 49

Su, X., Ditlev, J.A., Hui, E., Xing, W., Banjade, S., Okrut, J., King, D.S., Taunton, J., Rosen, M.K. & Vale, R.D. (2016). Phase separation of signaling molecules promotes T cell receptor signal transduction. *Science*, **352**, 595–599. 196

Sunnåker, M., Busetto, A.G., Numminen, E., Corander, J., Foll, M. & Dessimoz, C. (2013). Approximate bayesian computation. *PLoS Comput Biol*, **9**, e1002803. 33

Tanaka, Y., Tanaka, N., Saeki, Y., Tanaka, K., Murakami, M., Hirano, T., Ishii, N. & Sugamura, K. (2008). c-cbl-dependent monoubiquitination and lysosomal degradation of gp130. *Molecular and cellular biology*, **28**, 4805–4818. xx, 180, 181

Tebbutt, N., Pedersen, M.W. & Johns, T.G. (2013). Targeting the ERBB family in cancer: couples therapy. *Nature reviews Cancer*, **13**, 663–673. 4

Tian, T. & Song, J. (2012). Mathematical modelling of the MAP kinase pathway using proteomic datasets. *PloS one*, **7**, e42230. 238

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M.P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6**, 187–202. 33, 34, 35, 36, 37, 38

Trefethen, L.N., Birkisson, A. & Driscoll, T.A. (2017). *Exploring ODEs*, vol. 157. SIAM. 229

Tucker, H.G. (2014). *An introduction to probability and mathematical statistics*. Academic Press. 79

Turner, N. & Grose, R. (2010). Fibroblast growth factor signalling: from development to cancer. *Nature Reviews Cancer*, **10**, 116–129. 5, 7, 195

Uings, I. & Farrow, S. (2000). Cell receptors and cell signalling. *Molecular Pathology*, **53**, 295. 1

Vaillant, A.A.J. & Qurie, A. (2019). Interleukin. *StatPearls [Internet]*. 2

Van Kampen, N.G. (1976). The expansion of the master equation. *Adv. Chem. Phys*, **34**, 245–309. 20, 52, 53

Van Kampen, N.G. (1992). *Stochastic processes in physics and chemistry*, vol. 1. Elsevier. 52

Vera, J., Rateitschak, K., Lange, F., Kossow, C., Wolkenhauer, O. & Jaster, R. (2011). Systems biology of jak-stat signalling in human malignancies. *Progress in biophysics and molecular biology*, **106**, 426–434. 183

Vyse, S. & Huang, P.H. (2019). Targeting EGFR exon 20 insertion mutations in non-small cell lung cancer. *Signal transduction and targeted therapy*, **4**, 1–10. 238

Wahl, M.I., Jones, G., Nishibe, S., Rhee, S.G. & Carpenter, G. (1992). Growth factor stimulation of phospholipase C-gamma 1 activity. Comparative properties of control and activated enzymes. *Journal of Biological Chemistry*, **267**, 10447–10456. 203

Walfish, S. (2006). A review of statistical outlier methods. *Pharmaceutical technology*, **30**, 82. 241

Weddell, J.C. & Imoukhuede, P. (2014). Quantitative characterization of cellular membrane-receptor heterogeneity through statistical and computational modeling. *PloS one*, **9**, e97271. 57, 76

Weddell, J.C. & Imoukhuede, P.I. (2017). Integrative meta-modeling identifies endocytic vesicles, late endosome and the nucleus as the cellular compartments primarily directing RTK signaling. *Integrative Biology*, **9**, 464–484. 48, 57, 76

Wee, P. & Wang, Z. (2017). Epidermal growth factor receptor cell proliferation signaling pathways. *Cancers*, **9**, 52. 4

Wiederkehr-Adam, M., Ernst, P., Müller, K., Bieck, E., Gombert, F.O., Ottl, J., Graff, P., Grossmüller, F. & Heim, M.H. (2003). Characterization of phosphopeptide motifs specific for the src homology 2 domains of signal transducer and activator of transcription 1 (STAT1) and STAT3. *Journal of Biological Chemistry*, **278**, 16117–16128. 148

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. 189

Yates, J.W., Ashton, S., Cross, D., Mellor, M.J., Powell, S.J. & Ballard, P. (2016). Irreversible inhibition of EGFR: modeling the combined pharmacokinetic–pharmacodynamic relationship of osimertinib and its active metabolite AZ5104. *Molecular cancer therapeutics*, **15**, 2378–2387. 280, 281

Yoshida, H. & Hunter, C.A. (2015a). The immunobiology of interleukin-27. *Annual review of immunology*, **33**, 417–443. 2, 3

Yoshida, H. & Hunter, C.A. (2015b). The immunobiology of interleukin-27. *Annual review of immunology*, **33**, 417–443. 175

Yuzhalin, A.E. & Kutikhin, A.G. (2015). Chapter 1 - introduction: Basic concepts. In A.E. Yuzhalin & A.G. Kutikhin, eds., *Interleukins in Cancer Biology*, 1 – 16, Academic Press, Amsterdam. 123

Zhai, X., Ward, R.A., Doig, P. & Argyrou, A. (2020). Insight into the therapeutic selectivity of the irreversible EGFR tyrosine kinase inhibitor osimertinib through enzyme kinetic studies. *Biochemistry*, **59**, 1428–1441. 238, 280

Zhang, G., Wang, Z., Du, Z. & Zhang, H. (2018). mTOR regulates phase separation of PGL granules to modulate their autophagic degradation. *Cell*, **174**, 1492–1506. 197

Zhang, J.M. & An, J. (2007). Cytokines, inflammation and pain. *International anesthesiology clinics*, **45**, 27. 123

Zhang, X., Cao, J. & Carroll, R.J. (2015a). On the selection of ordinary differential equation models with application to predator-prey dynamical models. *Biometrics*, **71**, 131–138. 25

Zhang, X.Y., Trame, M., Lesko, L. & Schmidt, S. (2015b). Sobol sensitivity analysis: a tool to guide the development and evaluation of systems pharmacology models. *CPT: pharmacometrics & systems pharmacology*, **4**, 69–79. 30, 222