Internal Report CS Bioinformatics Track 17-03 September 2017

Leiden University

Computer Science

Bioinformatics Track

Computational modeling of drug response and pathway activity in colorectal cancer with missing data reconstruction and subtype analysis

Tom Seinen

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Contents

Summary **2**

1 Background **3**
1.1 Introduction . . . 3
1.2 Previous work . . . 4
1.2.1 Computational models . . . 4
1.3 Contributions . . . 6
1.4 Materials, Method, and Models . . . 7
1.4.1 Data . . . 7
1.4.2 Modeling Method . . . 8
1.4.3 Pathway models . . . 10

2 Missing data replacement **13**
2.1 Phosphorylation and transcription targets . . . 13
2.2 Replacement data . . . 14
2.3 Experimental setup . . . 16
2.4 Results . . . 17

3 Colorectal cancer drug response model **24**
3.1 Experimental setup . . . 24
3.1.1 Elastic net regression . . . 24
3.2 Results . . . 26

4 Colorectal cancer subtype analysis **36**
4.1 Experimental setup . . . 36
4.2 Results . . . 37
4.2.1 Untreated model . . . 37
4.2.2 Treated model . . . 39

5 Discussion **42**
5.1 Acknowledgements . . . 45

Bibliography **52**

Supplementary Figures and Tables **53**

Summary

Colorectal cancer is a major public health problem worldwide and its burden is expected to increase due to the growth and aging of the global population. Colorectal cancer can be treated using targeted drugs that specifically targets gene and protein alterations that influence the proliferation of cancer cells, by inhibiting affected signaling pathways in the cell. Knowledge on the molecular basis of the disease and on targeted inhibitors is increasing, but there is still a variety of problems that have to be solved. The main problem for targeted drugs is that the response to these targeted anticancer therapies are extremely variable and understanding the mechanisms causing this variability is a big challenge in cancer research. Many of these mechanisms have been discovered, but it is still unclear how they interact to induce cell line sensitivity or resistance to a specific drug.

Multiple computational approaches have tried to shed light on this problem, from linear and machine learning models, that try to find associations between genetic features and the drug response, to specific mechanistic modeling, that simulate signal transduction and regulatory networks to predict drug sensitivity. However the biological mechanisms are not taken into account in case of linear and machine learning methods and the number of cell lines that are studied is limited in mechanistic approaches or they only use a single data type.

We use a novel modeling approach developed at the Netherlands Cancer Institute by Thijssen, to explore whether the drug sensitivity in 46 colorectal cancer cell lines can be explained with computational signaling pathway models, combined with extensive multi type data measurements and Bayesian statistics. We will use this approach for the first time with protein mass spectrometry data and we present a method for replacing the missing protein phosphorylation data that are needed as protein activation measurement in the model. This is done by reconstructing the protein activation in different ways based on its phosphorylation or transcription targets. Also we will use the ability of the model to estimated the activation of the signaling pathways in the model to distinct different colorectal cancer subtypes and their response to targeted therapy.

We find that our replacement protocol for protein phosphorylation data is successful in reconstructing a activation signal that can be explained by the model. Thought, not every signaling node benefits from the same replacement approach. The drug response model, using the replacement activation data, can explain the variability in the drug sensitivity between the cell lines to a certain extent for a set of 4 targeted inhibitors and one drug combination. Moreover it captures the genetic associations and dynamic signaling mechanisms underlying the drug response, that are validated against literature and compared to Elastic net regression. The model is also able to roughly distinguish a set of colorectal cancer subtypes based on the estimated activation of the pathways and proliferation, although the effect is not strong.

This research shows that using this integrative modeling approach provides better understanding of the underlying mechanism and complexity of targeted drug sensitivity in colorectal cancer and is a step in the way to precision and personalized medicine.

Chapter 1

Background

1.1 Introduction

Colorectal cancer (CRC) is one of the most common cancers worldwide and a major public health problem[1]. Although it is the third most commonly diagnosed cancer and the fourth cause of cancer death in the world[1], the CRC prevalence and mortality rates have been declining the last few decades in the US and Europe, mainly because of a decrease in risk factors (Smoking, red meat consumption, use of aspirin) and the improvement of treatment and prevention by screening tests [2]. Because of this early detection and treatment combined with a growing and aging population the number of CRC survivors is still increasing[3]. Even if CRC continues to be a disease of the developed world, its incidence rates have been increasing in developing countries[1]. Also the worldwide burden is expected to increase due to the growth and aging of the global population and because of the adoption of western behavior and lifestyle. Therefore extensive research is needed to increase the prevention and treatment of colorectal cancer.

The treatment of CRC consists of local treatments(surgery, and radiotherapy), and full body treatments(chemotherapy, targeted therapy, and Immunotherapy)[3]. Currently the treatment of CRC differs between the different tumor stages. Patients with stage I or II (local tumors) CRC undergo only partial or total colectomy(surgical removal of a part of the colon)[3], two-thirds of patients with stage III tumors (metastasis to lymph nodes) will receive additional chemotherapy, and for stage IV colorectal cancers (distant metastases) chemotherapy is the main treatment, including regular chemotherapy and targeted drugs, and in some limited advanced cases immunotherapy can be used.

In this thesis we will focus on the treatment of metastatic CRC with targeted drugs. These drugs are the product of extensive research on gene and protein alterations that influence the mechanisms of cancer, and by specifically targeting these changes they aim to inhibit the proliferation and growth of the tumor. Many cancer genome research projects have as goal to characterize these genetic alterations in cancer and try to discover the therapeutic targets through large scale genomic profiling and mapping of the cancer genome[4], accordingly increasing our knowledge on the molecular basis of the disease. As a result the number of available targeted drugs for treating metastatic CRC is growing and many protein targeting drugs are in clinical use or development[5].

However, there is a variety of problems concerning targeted therapies that still have to be solved. The main dilemma is that patient and tumor response to targeted anticancer therapies is extremely variable, most drugs work in only a subset of patients, and understanding this variability is a big challenge in cancer research[6]. Known and unknown mutations in oncogenes or tumor suppressors changes the properties of dynamic signaling pathways in the cell, that orchestrate proliferation and apoptosis, causing the sensitivity to a specific targeted drug to vary widely between cells and tumors with different genetic alterations. It is common that the response of patients to targeted therapies is being associated and stratified with biomarkers, usually consisting of either copy number alteration or mutation of specific genes. Before treatment, the tumors are screened for these biomarkers to see if the patient is eligible for a certain therapy. However for many drugs no efficient biomarker exists, and for those where markers exist, they have relatively low predictive power for drug response and their true clinical significance for precision treatment is debatable[7]. A lot of effort is being spent on patient stratification based on these genetic alterations to achieve optimal therapeutic outcome, but the understanding of the underlying mechanisms of the drivers

of cancer is found to be just as important[8]. Particularly when using pathway-targeted therapies, where not all patients with the targeted mutation respond. Therefore identifying good biomarkers, or other factors responsible for variation in drug response, and understanding how these dynamic signaling pathways behave in the cell under the influence of different genetic alterations, is essential for the response prediction and effective use of targeted therapeutics and precision medicine[5].

Another common problem of targeted therapeutics is the problem of drug resistance. When patients receive targeted therapy it is possible that a period of tumor regression is followed by recurrent progression[9]. Although the knowledge about the mechanisms underlying this drug resistance is still scarce, numerous research groups have identified several unique mechanisms. A major cause is the intratumor heterogeneity between the tumor cells[10]. Although all cells in tumors usually originate from a single cell, the evolution of the tumor is chaotic, with genetically different tumor clones coexisting within tumors for long periods of time. They all may have a different response to the therapy and thus only a part of the tumor cells are killed with the targeted therapy, while other cells appear to be resistant. Furthermore, single cells can become resistance to targeted drugs over a period of time, by accumulating constitutive mutations that change the dynamic mechanisms in the cell, making the drug lose its effectiveness or completely obsolete[6]. Further insights in the mechanisms of dynamic signaling networks and the resistance to current therapeutics are therefore needed to tackle the key factors contributing to the lethal outcome of cancer and therapeutic failure and to improve clinical outcome[11][12].

Understanding the molecular effect of single targeted drugs on the dynamic signaling network of a cell is crucial to explain the drug sensitivity and drug resistance of tumor cells. The analysis of the activity of signaling pathways will support the development of predictive and response biomarkers that can be linked to cellular mechanisms[13][6]. Because the dynamic mechanisms in cancer are very complex, multiple computational approaches have been designed to model the signaling pathways in a cell and to shed light on the behavior of cancer cells treated with targeted drugs.

1.2 Previous work

1.2.1 Computational models

Computational models of drug sensitivity, resistance, and cancer cell signaling have a lot of potential to overcome the limitations of the reductionist approach, where individual genes and proteins are studied without information from other elements in the system in which they interact and function[14][15]. These studies have been very effective in finding specific characteristics of particular biological processes, but biomolecules depend on interactions with many other biomolecules, therefore other, more integrative, methods are needed for new scientific discovery. Modeling, and especially computational modeling, has become a powerful tool in this field[15].

Linear and Machine learning models

To untangle the complexity of drug sensitivity and resistance, a variety of statistical and machine learning approaches have been developed over the last decades, making use of the advances in high-throughput drug screening technologies that have enabled the testing of thousands of drug candidates on large panels of cancer cell lines[16].

Studies on large scale datasets from TCGA[1], GDSC[2], and NCI-60[3] have been conducted using a variety of supervised learning approaches such as commonly used elastic net regression and MANOVA(Garnett[17], Barretina[18], Iorio[19]), random forest(Daemen[20], Riddick[16]), Support vector machines(Daemen[20], Dong[[21]), Naive bayes classification(Barritina[18]) and other classifiers(Lee[22], Stanton[23]) and also some unsupervised clustering methods(Seashore[24]).

As Deep Learning is increasingly applied in various fields of computer science it is also applied here. Vougas et al.[25] state that cancer datasets are too multidimensional to be effectively managed by classical Machine Learning algorithms and have therefore developed a deep learning neural network and use association rule mining to analyze the complex biological data. All these statistical approaches have provided a lot of knowledge and many specific targeted drug biomarkers and genetic associations, however, because no intrinsic knowledge and details about the biological

[1] https://cancergenome.nih.gov/
[2] http://www.cancerrxgene.org/
[3] https://dtp.cancer.gov/discovery_development/nci-60/

systems in the cell is incorporated they do not explain or merely guess the underlying mechanisms of the strongest associations found. More specific and smaller associations or individual cell characteristics, such as the activity of certain biological pathways, and sensitivities cannot directly be revealed.

Mechanistic approaches

More specific mechanistic modeling approaches have been proposed to analyze and simulate signal transduction and regulatory networks to predict drug sensitivity and the underlying mechanisms[26]. They integrate molecular and phenotype data with pathway knowledge to gain a better understanding of the genetic and signaling alterations that determine the response to targeted drugs[14]. This enables models to determine the intracellular signaling activity under the influence of different mutations and different therapeutic treatments, resulting in different cell phenotypes, such as apoptosis, cell differentiation, and proliferation[14]. These mechanistic modeling approaches use different levels of abstraction, ranging from very detailed kinetic models between signaling molecules to abstract topological boolean models of multiple pathways.

Because signaling pathways are real biological systems where chemicals react according to chemical and physical laws, it seems appropriate to model the system using a set of differential equations. However due unknown kinetic parameters, missing data, and incomplete mechanistic details of many biological systems, these models are only feasible on smaller, known, and well-studied systems[27]. Studies that use such methods include Fey et. al.[13], using a ODE model to describe the behavior of a select number of molecules in the JNK signaling pathway and use the pathway dynamics as a biomarker for therapy, Bidkhori et al.[28] who built similar mathematical models representing EGFR signaling, and Sulaimanov et. al.[29] modeling the MTOR signaling pathway based on multiple ODE models.

On the other hand, the qualitative topology models provide coarse-grained descriptions of the underlying biological systems, using less complex logic or boolean operators and require less data for parameterization[30][[27]. These models are commonly used to study abstract biological systems where many smaller subsystems and reactions are clustered together, but is also used on more complex detailed systems to acquire qualitative knowledge of the network. Logic models are larger, incorporating more pathways and elements than ODE models, which is possible because the model mechanics are less complex. Examples are Calzone et. al.[31], using a logic model of multiple signaling pathways to predict the survival or apoptosis of cells on the activation of death receptors, Zhu et. al.[32] used a logic model of signaling pathways in cancer to identify potential drug targets in breast cancer, and Bonzanni et. al.[33] shed light on the behavior of blood stem cells by creating a dynamic regulatory network logic model.

In between these two approaches there are many hybrid models that capture signaling systems at different levels of abstraction and use a variation of mathematical formalisms and modeling techniques[34]. Hybrid modeling approaches are becoming more important[34], in which qualitative and quantitative representations are combined, because of the growing number of biological mechanisms to model and the increasing availability of experimental measurements. Four notable studies can be mentioned, one from Ryll et. al.[35] who coupled logical models of signaling and gene-regulatory networks with kinetic models of metabolic processes. Klinger et. al.[30] build a framework, MRA, to calculate the response of an ODE model to a set of perturbations to analyze the MAPK and PI3K pathway with measured specific node stimulation and inhibition perturbations. A study by Kirouac et. al.[36] introduced a method of quantitative logic, and Eduanti et. al.[7], used a similar method as klinger, with perturbed multitype data, to build cell line–specific dynamic logic models to investigate the signaling pathway dynamics and drug response. These mechanistic approaches of signaling pathway models have been successful and have already been used to propose single or combined targeted therapies to block one or multiple signaling pathways[7] and to analyze the signaling pathway dynamics besides finding novel genetic or proteomic associations.

However, it must be noted that all approaches are tailored to a specific problem, a chosen level of abstraction, and rely on different data types, such as perturbed data or measurements of time. This means that the model approaches are not always easily adaptable or interchangeable for one another.

Parameter estimation

Methods used in above mentioned modeling approaches all rely on parameters that have to be

estimated from the data. There are some differences in the parameter estimation and optimization methods of these models. The ODE modeling methods usually make use of numerical analysis simulations to solve the differential equations[27][28] or use adaptive simulated annealing and Monte Carlo-based approaches[13], by repeatedly changing the parameters and refitting the model. The latter are also used in logic and hybrid approaches. Other parameter solving methods include, particle swarm optimization, using a population of candidate solutions that move over a search-space [36], and modular response analysis, for the analysis of parameter perturbations [30]. Distinction must be made between methods that provide single point parameter values and stochastic methods estimating the parameter as random variable with a probability density distribution. When models are complex and have a high level of uncertainty, the use of parameters as random variable can provide more information on the true parameter value, as it models its uncertainty. However estimating probability density distributions is much more computationally intense. There is also a difference between parameter optimization methods that specifically optimize or maximize a certain likelihood based on the data and parameter estimation methods that generalize the parameter values from the data without necessarily maximizing the likelihood, such as Bayesian parameter inference techniques[37].

1.3 Contributions

Modeling CRC pathways
As mentioned in the previous paragraph there exist multiple ways of modeling signaling pathways and targeted drug sensitivity, however these approaches lack the available knowledge of biological mechanisms in the models in case of linear and machine learning methods, the number of cell lines that are studied is limited in mechanistic approaches, or they only use a single data type. Therefore in this work we will use a signaling network model approach, developed by Thijssen et al.[38], which uses multiple data types and Bayesian inference to analyze the drug response and pathway activation. We developed a model of the two important signaling pathways of CRC and cancer in general, the MAPK, and PI3k pathway, and integrate it with the molecular measurements of a panel of 46 CRC cell lines. A description of the modeling method can be found in the next section 1.4.2. In chapter 3 we will describe the modeling for four targeted drugs and one combination drug set and validate our findings by comparing them to current knowledge from literature, to see if we can recapitulate the known and maybe find unknown associations and mechanisms. We also compare the results to the results of a elastic net regression model, that is performed on the same dataset and is commonly used for finding genetic associations with drug response[17].

PMS instead of RPPA
The model developed by Thijssen has so far only proven to work with reverse phase protein array(RPPA) data. RPPA is a powerful tool but limited to the few hundred antibodies of good quality that are currently available on the market[32]. The use of data from other sources, such as protein mass spectrometry has not yet been tested with this modeling technique. Protein mass spectrometry (PMS) has rapidly advanced in recent years[39], but it is still less accurate than RPPA and has a hard time quantifying proteins in low abundance[40]. However PMS can measure much more different proteins than RPPA. As it does not depend on specific antibodies, PMS can be used to explore a very wide range of proteins, which we will use in this work to our advantage to select phosphorylation targets. This capability to identify thousands of proteins in complex biological samples, without specification of the protein that must be measured, makes PMS increasingly important in clinical proteomics[32]. It is therefore very interesting to see whether this modeling approach is able to distinguish the variation in the less accurate mass spectrometry data as good as in until now used RPPA data.

Missing data replacement
Another major disadvantage of the protein mass spectrometry measurements is that expression data of important phosphorylated proteins are missing, allegedly due to the low abundance of these proteins and peptides in the cell. This is a problem because the activation data of the pathway components is essential to be able to correctly model the signaling pathways. Also because the use of PMS in research and databases is increasing[41][42], it is important to be able to also use it for signaling node activation modeling.

To our knowledge there is not yet a solution for dealing with these missing activation data other than leaving it out of the model. Therefore we propose a solution to this problem in the next chapter, in the form of a method to replace or reconstruct this missing data. Using the signaling network based on current knowledge, PMS data, gene expression data, and knowledge of phosphorylation targets and transcribed genes of proteins from literature and databases, we reconstruct the activation expression of the missing node data. We based our proposal on the idea that the variance of the activation of a pathway component is transferred to its targets downstream, to each in a different way and strength. When combining the variances of all or a selection of these targets the variance of the parent component should be able to be retrieved. This idea is also used in studies concerning signal processing and 2d and 3d image and surface reconstruction, which show that a relatively small number of random projections or samples of a signal can contain most of its information from which a very accurate reconstruction can be created[43][44][45][46].

We will validate our replacement data protocol by comparing it to the known node activation, and to how well the model can explain the variance in the data. With this research we will try to answer the question to what extend the activation of a signaling node can be reconstructed using its phosphorylation targets or transcription genes and if the model can explain the variance. Using the results of the experiments on this target replacement data protocol we select the best new target data to our signaling pathway model to use in our models with the drug response data in chapter 3.

CRC subtypes differentiation
In the last chapter we describe how we use the model's ability to estimate the activation of specific pathways to distinguish different CRC subtypes, as these subtypes are characterized by specific mechanisms of cell signaling[47][48]. We will do this by creating a model of five signaling pathways commonly involved in CRC, namely the MAPK, PI3K, WNT, P53, and TGFR pathways. Incorporating this number of pathways will increase the complexity of the model, therefore we increase the level of abstraction by representing the pathways with less nodes, to enable proper convergence. Comparing our findings with literature will validate our results, giving insight into whether the modeling method can differentiate between CRC subtypes and to what extent the known characteristics can be recapitulated.

1.4 Materials, Method, and Models

1.4.1 Data

In this study we use a panel of 46 CRC COSMIC cell lines [4]. For all these cell lines genetic aberration data is publicly available, including mutations, losses, and amplifications, together with gene expression data. Besides this public available data, we were also able to work with additional dataset provided by the Wellcome Trust Sanger Institute(WTSI), containing mass spectrometry measurements of protein and phosphorylated protein expression in the individual cell lines and drug dose response measurements of the cell lines treated with a set of targeted inhibitors.

A set of 93 SNPs were profiled for all cell lines, with as a result a binarized matrix indicating if a mutation is present in the cell line. Besides SNPs also copy number variation was analyzed for both 425 pancancer and 66 colorectal cancer specific altered genes and chromosomal segments, the data was also binarized for use. All this genetic data is been studies in the COSMIC cell line project and the Genomics of Drug Sensitivity in Cancer (GDSC) project[49][5].

We were provided with protein mass spectrometry dataset by the WTSI, which covered the abundance measurements of 9473 proteins and 11188 phosphosites located on 3974 different proteins (as a protein can have multiple phosphosites) for each cell line. All measurements were normalized between 0 and 1 and similar sequences incorporating the same phosphosites were combined. The phosphorylation expression also was normalized by subtracting its protein abundance, to obtain the relative phosphorylation measurements per protein.

Additionally, we received from the WTSI a drug dose response screen using 27 targeted inhibitors and one combination treatment, with targets including RTKs, MAPK pathway and PI3K pathway signaling components. For every drug-cell line combination 7 dose concentrations were measured, some with multiple plate replicates. The concentration doses are decreasing 2-fold with

[4] http://cancer.sanger.ac.uk/cell_lines
[5] http://www.cancerrxgene.org/

a drug specific maximum concentration. Before analysis the drug responses were normalized on blanks and controls and a selection of 9 drugs was made that had at least a 50 percent reduction of relative proliferation.

1.4.2 Modeling Method

Signaling Model

The signaling pathway modeling method developed by Thijssen at the Netherlands Cancer Institute is based on estimating the activation of the signaling nodes, the strength of the activation or inhibition between the nodes, and their effect on the proliferation of the cell under treatment of targeted drugs, Figure 1.1.

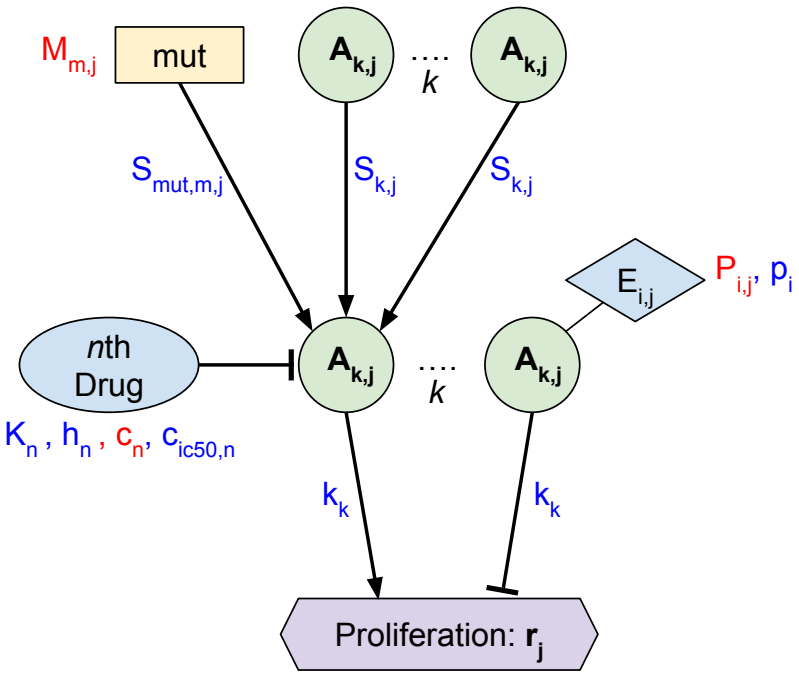


Figure 1.1: Overview of the signaling model and its functions; In blue are depicted the function parameters and in red the data variables.

The activation ($A_{i,j}$) of the ith signaling node in the jth cell line depends on the following variables; the basal activity (b_i) of the ith signaling node, the activation ($A_{k,j}$) and strength ($s_{k,j}$) of the activation of the (k) parent signaling nodes, the presence and strength of (m) mutations affecting the ith node, and optionally the measured abundance ($E_{i,j}$) of the node's protein. All together the activation is defined as:

$$A_{i,j} = E_{i,j} \cdot (b_i + \sum_{kparent\_nodes} s_{k,i} \cdot A_{k,j} + \sum_{menode\_mutations} s_{mut,m,i} \cdot M_{m,j}) \tag{1.1}$$

Where b_i, $s_{k,i}$, $s_{mut,m,i}$, are the modeled parameters, $M_{m,j}$ is the mutation and cnv data in binary matrix form, and $E_{i,j}$ is the expression by the formula:

$$E_{i,j} = (p_i \cdot P_{i,j}) + (1 - p_i) \tag{1.2}$$

The expression coefficient parameter (p_i) controls whether the abundance of the protein i is limiting the signal transduction and $P_{i,j}$ is the protein expression data. The node activation is then clamped between 0 and 1 using a sigmoid function.

The drug effect is modeled by multiplying the node's activation by the effect of the model depending on the maximum inhibition (K_n) of the drug, the log steepness (h_n) of dose response effects,

the half-maximal inhibition concentration ($IC50_n$), and the log concentration of the drug(C_n), all for the nth drug. The inhibited activation becomes:

$$A_{i,j,inhibited} = A_{i,j} \cdot (K_n + \frac{1 - K_n}{10^{h_n \cdot (c_n - c_{IC50,n})} + 1}) \tag{1.3}$$

Containing the estimated parameters K_n, h_n, and $IC50_n$, and data variable C_n.

The proliferation (r_j) of the jth cell line, depends on the activation of k nodes signaling to proliferation and their strength(parameter k_k) on the proliferation signal. If the effect of the drugs is not incorporated in to the model the untreated proliferation becomes,

$$r_j = \sum_{k \epsilon parent\_nodes} k_k \cdot A_{k,j} \tag{1.4}$$

, and including the inhibition of the nodes by the targeted drugs will pass along the effect in the activation functions to the indirectly inhibited proliferation.

$$r_{j,inhibited} = \sum_{k \epsilon parent\_nodes} k_k \cdot A_{k,j,inhibited} \tag{1.5}$$

Using these two functions we can calculate the cell line drug response(D_j) by normalizing the proliferation under drug treatment at a particular concentration to the untreated proliferation.

$$D_j = \frac{x_{0,j} \cdot e^{r_{j,inhibited} \cdot t_{treatment}}}{x_{0,j} \cdot e^{r_j \cdot t_{treatment}}} = e^{t_{treatment}(r_{j,inhibited} - r_j)} \tag{1.6}$$

Here the $x_{0,j}$ is the seeding density of the cell line, which is the starting number of cells, and $t_{treatment}$ is the treatment duration.

Bayesian inference

The model likelihood is equal to the probability density of the observed measured outcomes given the parameter values. Using the activation function and drug response function we can create a likelihood function for each activation and drug response measured data point. Including each data type the likelihood function becomes:

$$L(\theta|y) = \prod_{i \epsilon observed\_variables} \prod_{j \epsilon cell\_lines} \prod_{k \epsilon replicates} P(y_{i,j,k}|\theta) \tag{1.7}$$

Variable y depicts the measured data and θ the parameter vector. As individual likelihood function for every variable a student's t-distribution is used. The number of degrees of freedom is set to three, because the t-distribution only serves as a means of robust inference. The likelihood function per data point is therefore as follows:

$$P(y_{i,j,k}|\theta) = t(y_{i,j,k}|\mu = x_{i,j}, \sigma = \sigma_i, \nu = 3) \tag{1.8}$$

$$x_{i,j} = \begin{cases} g_i + (1 - g_i)A_{i,j}(\Theta) & \text{for observed activation data} \\ D_j(\Theta) & \text{for observed drug response data} \end{cases} \tag{1.9}$$

The variance of the measured variable σ_i is a parameter and the mean is a modeled variable that depends in case of signaling node activation measurements on the activation function and a background signal parameter and for the drug response measurements the drug response function is used.

The aim is to use Bayesian inference to calculate the posterior density distribution of the parameters given the data $P(\theta|y)$. According to the Bayesian theorem,

$$P(\Theta|y) = \frac{P(y|\Theta)P(\Theta)}{P(y)} \tag{1.10}$$

the posterior can be obtained by dividing the product of the likelihood and the prior distribution of the parameter over the marginal likelihood $P(y)$. The likelihood is as given as in Equation 1.7 and the manually set prior distributions for the parameters can be found in Supplementary Table

5.1. The denominator cannot be computed analytically, therefore we have to sample from the posterior density distribution in a high dimensional parameter space. The sampling method we use is PTMCMC, a description can be found in the next paragraph. From sampling the posterior we obtain the marginal probability densities for each of the parameters and an estimation of the posterior distribution, which we can use to calculate the marginal likelihood of the model.

Using the approximated Monte Carlo samples we are able to calculate the posterior predictive distribution. This posterior predictive is a probability distribution of a new modeled dataset, after we have seen the observed data set , and can be compared to the observed data set to see what extend the model can explain the data. It is calculated by integrating the likelihood times the posterior over all the parameters given model M:

$$P(y^{pred}|y, M) = \int P(y^{pred}|\theta, M) \cdot P(\theta|y, M)d\theta \tag{1.11}$$

Which can be approximated by taking the average over all obtained samples N of the likelihoods of the new data points given a parameter vector (θ_i) and the model.

$$P(y^{pred}|y, M) \approx \frac{1}{N} \sum_{i=1}^{N} P(y^{pred}|\theta_i, M) \tag{1.12}$$

Posterior sampling

To estimate the posterior distribution from the likelihood and prior distributions we use parallel tempered Markov Chain Monte Carlo sampling to generate the samples from the posterior distribution. To aid the MCMC we use additional approaches, consisting of adaptive temperatures, parameter blocking, and adaptive proposals, to improve and speed up the convergence. This sampling approach is implemented in a software package developed by Thijsen et. al.[38] and is available on the NKI's Computational Cancer Biology website[6].

Markov Chain Monte Carlo sampling can generate samples from a high dimensional search space by using a Markov Chain that proposes and accepts or rejects smart jumps in this space[50]. The sampling starts with a burn-in period, to let the chain find the sampling space, these samples are then discarded, and continuous with the sampling period, where the sampling distribution is generated.

A variation on MCMC is parallel tempered MCMC, which runs multiple chains at different temperatures that influence the posterior. At high temperatures the sampling distribution will become the prior distribution making it easier to explore the parameter space while in low temperatures the chains can explore the peaks of the likelihood. After a predefined number of Monte Carlo samples a sequence of swaps, the exchange of chains at neighboring temperatures, is suggested and accepted with a certain probability[51]. This method alone improves the convergence effectively.

We want the chains to visit both high and low temperatures as many times as possible to obtain independent samples, so the number of round-trips between the lowest and highest temperature must be maximized. Therefore we make use of a variance of this PTMCMC, by Katzgraber et al.[51], that adapts the set of temperatures to optimize the number of round trips of the chains, by minimizing the up and downwards trip times.

Additionally, we use parameter blocking, by Turek et al.[52], which jointly updates multiple parameters simultaneously, as a parameter block, if they are correlated with each other. This increases the sampling performance compared to updating each parameter independently.

Finally, adaptive proposals, by Haario et al.[53], are used to adapts the proposal distribution continuously to the target distribution based on all the previous states. This adaptation makes sure that the search is more effective at an early stage of the simulation.

The convergence of the PTMCMC can be determined by monitoring the autocorrelation of the parameters, the traces of the variables, and the number of temperature round trips.

1.4.3 Pathway models

For this thesis we created two signaling pathway models that we will use to model the pathway activation and drug response of CRC cell lines. We created these pathway models based on

[6] http://ccb.nki.nl/software/bcm/

literature, pathway databases, e.g. Kegg database, and available data. To select the signaling pathways of interest we looked at the most important and frequent mutations in CRC and which pathways they influence. Also the availability of a targeted drug that inhibits the pathway is of important for obvious reasons. The MAPK and PI3K signaling pathways harbor many of these important mutations, such as the BRAF, KRAS, EGFR, PTEN, and PI3K mutations[54], and are therefore the first pathway models we created. There are more mutations that play a significant role in CRC, including APC, β Catenin, SMAD, and P53 mutations, that influence multiple pathways[54]. Based on these mutations we created a second pathway model consisting of the MAPK, PI3K, P53, TGFBR, and WNT signaling pathways.

Both models will be used in chapter 2 for testing the missing data replacement, the first model will be used in chapter 3 for modeling the cell line drug response, and the second model for the colorectal cancer subtype analysis. We designed both models in CellDesigner[7], a SBML modeling tool of biochemical networks, and described them in more detail below.

MAPK and PI3K model

The MAPK and PI3K pathway model is based on the mutations that are most frequent in colorectal cancer and cancer in general. Important are the KRAS and BRAF mutations, which are involved in respectively 25–60% and 13% of colorectal cancers[54][55], both mutations directly influence the MAPK pathway. One third of colorectal cancers have activating somatic mutations in PI3KCA[54], together with PTEN mutations they deregulate the PI3K pathway.

Both pathways are activated by the same growth factor receptors, including EGFR, FGFR, IGFR, and ERBB2(HER2). These receptors are also often subject to mutations, genetic amplifications, and losses[54].

We modeled the MAPK pathway by the cascade including RAS, BRAF, MEK, and ERK signaling to proliferation, see figure 1.2. The PI3K pathway is depicted by its signaling components PI3K, AKT, mTOR, S6K, and 4EBP1, where AKT is inhibited by PTEN. Both RAS and PI3K are activated by the growth factor receptors, and the pathways interact with each other via RAS activating PI3K[56] and ERK activating MTOR[6]. We added to each node in the model its available mutations and genetic loss or gain. Five targeted drugs are used in this thesis, all directly targeting the MAPK and PI3K pathway. Drugs targeting the MAPK pathway are Trametinib(MEK), SCH772984(ERK), Afatinib(EGFR), and a combination of Afatinib and Trametinib. MK2206(AKT) targets the PI3K pathway together with Afatinib and the combination drug.

Multi pathway model

The second model is an abstract or less complex version of the MAPK and PI3K model with the additional TGFB, WNT, and P53 pathways. We had to reduce the number of nodes per pathway because a model with all nodes would be very hard to converge. We reduced the MAPK pathway to one node, activated by the growth factor rectors, and the PI3K pathway was reduced to only AKT and two small proteins S6K 4EBP1, to confer the proliferation inhibition and activation, figure 1.3.

TP53 is a tumor suppressor gene and mutations in the TP53 gene occur in almost half of all metastatic colorectal cancers, causing the activation of target genes[55]. P53 is regulated by the protein MDM2, which is inhibited by AKT. We modeled this inhibitory effect directly on P53.

Mutations in SMAD4 and SMAD2 result in disruptions of TGF-β signaling pathway[57][55]. We modeled the TGF-β pathway with two nodes, the TGF-β receptor and SMAD, representing the SMAD proteins, inhibiting proliferation.

The last pathway we included was the WNT signaling pathway. This pathway, that regulates transcription of a number of critical cell proliferation genes via β catenin, harbors mutations that often occur in CRC. Mutations in the APC gene are one of the initiating events of CRC, around 85% of CRC tumors have these mutations[54]. APC regulates together with GSK3β the degradation of β catenin. β catenin itself is mutated in up to 10% of all sporadic CRCs[54]. Furthermore mutations occur in GSK3β and AXIN2, a regulator of GSK3β[55]. GSK3β is regulated by the WNT receptor, which is affected by the RNF43 mutation[58]. Moreover, GSK3β is inhibited by AKT and activated the growth factor receptors[59]. We use the three nodes, WNTR, GSK3β, and β catenin, to model the pathway. The same five targeted drugs can be applied to this model,

[7] http://www.celldesigner.org

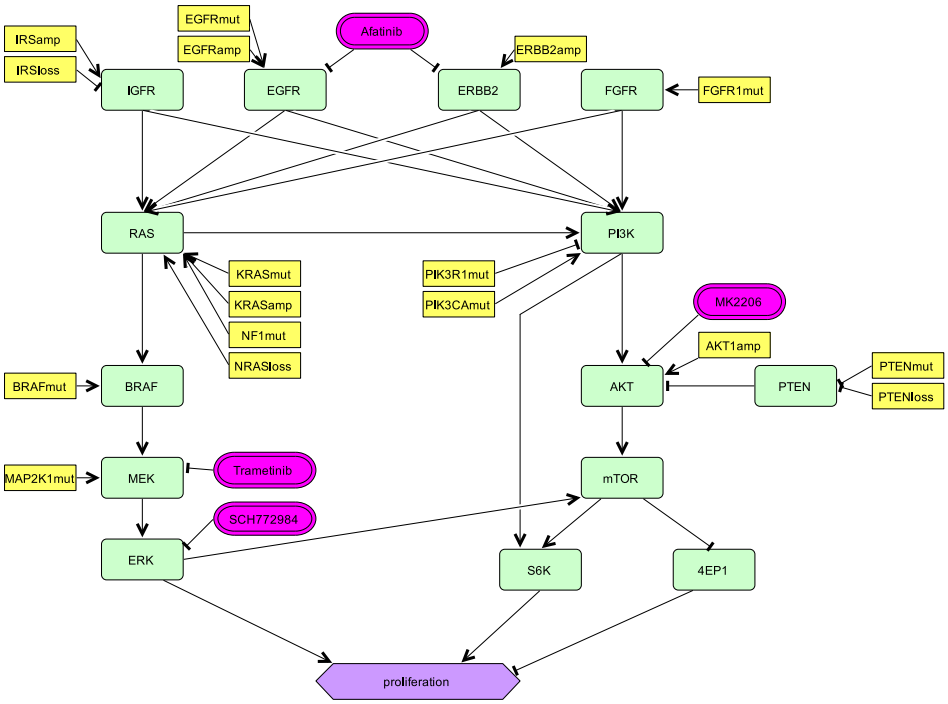


Figure 1.2: MAPK and PI3K pathway model

although they do not directly interfere with all pathways they do have an indirect effect. We modeled the AKT inhibitor MK2206, MEK inhibitor Trametinib, and ERK inhibitor SCH772984.

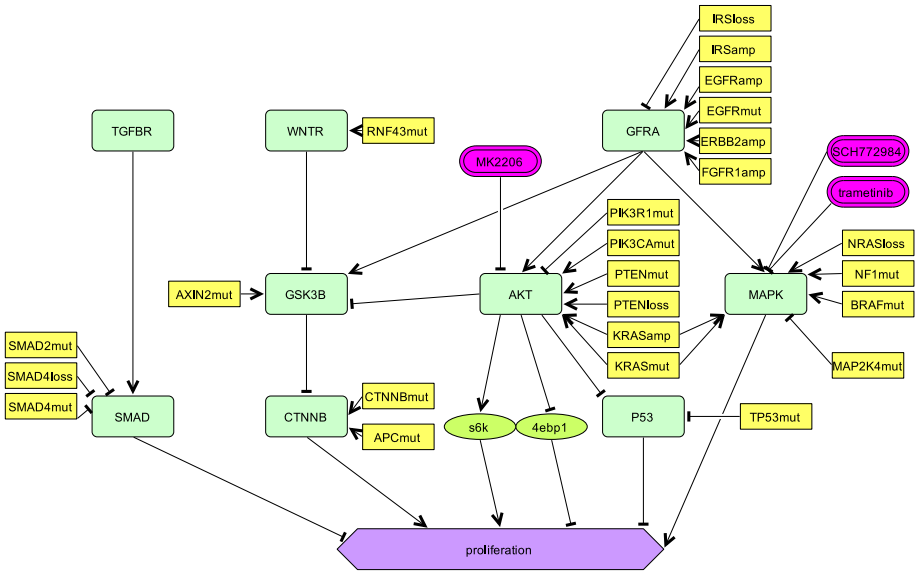


Figure 1.3: multi pathway model

Chapter 2

Missing data replacement

As mentioned in the previous chapter, protein mass spectrometry fails to accurately measure the expression of all phosphorylated proteins in the cell. This causes a problem, in particular, when the signaling nodes, from which you want to model the activity, are missing the data that directly captures their activity. For a number of nodes in the pathways that we chose to model lack this activation data in the protein mass spectrometry dataset. Examples of phosphorylation sites that are missing but are regarded as the true activation measurements of their proteins are phosphorylated S(serine)473 for AKT1[60], S222 and S218 for MAP1K2(MEK1)[61], and S394, T(Threonine)412, and S434 for S6K[62]. Without these data the activation of these proteins cannot be directly computed.

However, there is hope, because the protein mass spectrometry has measured the expression of a broad range of proteins, we do have measurements of the proteins and their phosphorylation in the functional neighborhood of these missing nodes. Meaning that the nodes that lack activation data have phosphorylation targets that do have data, which can be used to reconstruct the activity of the nodes itself.

This reconstruction can take place because the activated protein is phosphorylating its targets and consequently passes along its activity. This transfer, however, is subject to different factors of noise that change the original signal, including other proteins that also phosphorylate the target protein on the exact same site, the dependence of the proteins on their micro environment, and of course measurement noise. By combining or averaging the measured data of the multiple targets, the original signal and most importantly its variance can be reconstructed.

We used this theory to propose a method for finding, selecting, and incorporating the target measurement data in the model, filling the gaps of missing activation data.

2.1 Phosphorylation and transcription targets

To replace missing activation data for signaling components we introduce a protocol for finding and selecting replacement data. As explained above, the main idea is that when the true protein phosphorylation data is missing for a certain node A we look further downstream in the signaling cascade to find its N targets that are phosphorylated by A and thus indirectly convey the activation status of A. We then use the expression of these phosphosite targets as the activation expression of the original node A. However each signaling component is different and can have one, none, or multiple downstream targets. We consider two different downstream target options that can be used for alternative activation measurements, see figure 2.1. The first option are the phosphorylation measurements of the direct phosphorylation targets. For example the protein AKT is activated by phorylation of serine 473, but if the expression data of this phosphosite it missing, targets of AKT can be considered. AKT1 has multiple other phosphorylation targets, such as PRAS40(T246), MTOR(S2448), GSK3B(S9), and FOXO3(S253), and therefore their expression can be used as alternative measurement. If the node that is missing data is a transcription factor or if has a transcription factor directly downstream, the second option is to use gene or protein expression of the transcribed genes as alternative activation data. TP53, for example, is a transcription factor of genes including, CDKN1A, MDM2, RRM2B, SUSD6, whose measured gene or protein expression can be used as replacement data.

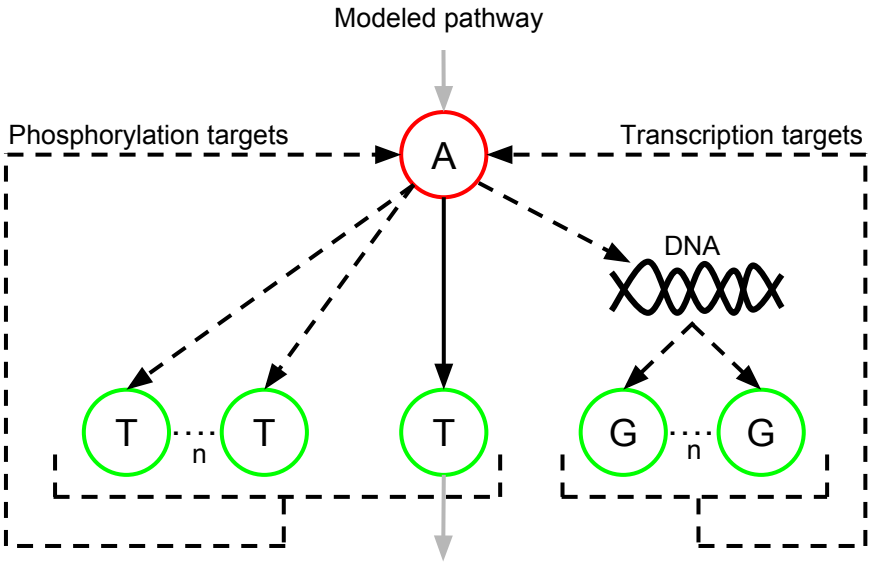


Figure 2.1: Missing activation data reconstruction; The modeled pathway with a node A missing its activation data. Node A has n phosphorylation targets, T, with measured expression or if it is a transcription factor, n target genes, G, with measured gene expression. These targets are then combined to reconstruct the activation of node A.

For this method it is crucial to know what the targets of a node are and if these targets are available in the dataset. Therefore one has to rely on the present knowledge on protein target phosphorylation. The phosphorylation targets of the signaling nodes can be found in literature, curated databases (eg. Phosphosite[1] and Kegg[2]), or databases that predict the possible targets (eg. Networkin[3]). Secondary or indirect activation targets are not always in Kegg, so literature curation is still needed in some cases. The strength or confidence of a phosphorylation relationship between two proteins must also be taken into consideration to be able to select the best suitable targets that represent the activation of the parent node in the most accurate way, as some proteins, eg. AKT1, have more that 100 possible targets. It must also be mentioned that the number of targets per node is very variable, as for some nodes such as AKT1 have many targets while for MEK1 it is hard to find multiple strong targets. This makes the search for targets a laborious and complicated job.

2.2 Replacement data

To see how the effect of targets represent the parent node and how well the model can recapitulate and explain the variance we incorporated the target data in the model in three different ways. First we used the single best target of the node, then we took the average of multiple best nth targets, and also used the multiple n targets as separate replicates to the model. To be able to select the targets we created a target list per node in the following way:

First conduct a search for all phosphorylation targets of the node of interest in databases and literature, also include targets that are already used by other nodes in the model as activation data, because these still hold variance information about the data missing node and should not be disregarded.

Then rank the list of targets on importance and confidence using Networkin[63][64] [4] prediction scores, which represent how likely the phosphorylation interaction is to occur in a probabilistic manner based on their occurrence in literature and databases such as STRING[5] and Kegg. If no such score can be found the targets can be ranked on self curation from literature and databases.

[1] http://www.phosphosite.org/
[2] http://www.genome.jp/kegg/
[3] http://networkin.info/
[4] http://networkin.info/
[5] https://string-db.org/

If a node is a transcription factor or directly influences one, a list of target genes is created and similarly ranked on their known interaction quality score. This score can be obtained from databases such as Transfac[65] or self curated from literature. Using this ranked lists, which can be found in Supplementary Table 5.2, we took the target with the best score or the n multiple best targets for use in the model.

For the single target approach we took the number one target, with the highest confidence of the ranked list. For example for AKT1 we selected FOXO3 S253 with the maximum confidence score of 54.934, only slightly higher than the second target FOXO1 S319. We then use this target's measured expression as measurement data for the activation of node AKT1.

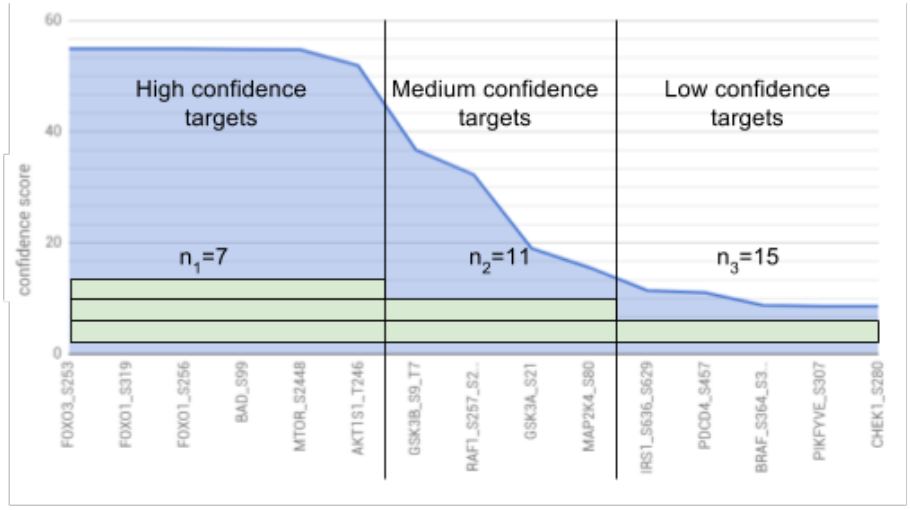


Figure 2.2: Grouping of the ranked targets of signaling node AKT1 based on confidence score

For the use of multiple targets from the ranked list we took the n targets with the highest confidence level. This n is different per node and depends on the trend of the ranked confidence scores and is semi-arbitrary. We created three groups of n targets per node, the first n_1 targets are the targets with the highest score, before a significant drop in the score for the next target in the list, see Figure 2.2. The second n_2 targets are the targets are the best targets until the confidence becomes steady, and the third n_3 targets include also the rest of the targets with less confidence, with a maximum number of 15. For example, see figure 2.2, for AKT1 after target AKT1S1 T246 there is a drop in confidence, and MAP2K4 S80 is the last target before the weaker targets are listed. As mentioned we implemented two ways of incorporating the multiple targets in the model. In the first method we used the n targets as separate replicates for the parent node in the model and in the second method we took the average of targets and used that as replacement data for the parent node. For calculating the average of the target phosphorylated nodes, $\bar{y}_{i,j}$, we took the average of the N targets from node i, all normalized so that their own data mean is equal to 0.5 to make the variances between the targets comparable to each other:

$$\bar{y}_{i,j} = \frac{1}{N_{targets}} \sum_{k \epsilon targets}^{N_{targets}} (y_{i,j,k} - (\frac{1}{N_{celllines}} \sum_{j \epsilon celllines}^{N_{celllines}} y_{i,j,k} - 0.5)) \tag{2.1}$$

Here $y_{i,j,k}$ is the phosphorylation expression of kth target node of node i in the jth cell line, $N_{celllines}$ are the number of cell lines, and $N_{targets}$ are the number of targets of node i. For targets of transcription factors normalization of the gene expression is not needed, the average therefore becomes:

$$\bar{y}_{i,j} = \frac{1}{N_{targets}} \sum_{k \epsilon targets}^{N_{targets}} y_{i,j,k} \tag{2.2}$$

Where $y_{i,j,k}$ is the gene expression of the kth gene transcribed by node ih in the jth cell line.

2.3 Experimental setup

We set up a set of experiments to test our method of replacing a node's activation data and our target selection protocol, Figure 2.3. Starting with a baseline experiment we test how well the variance can be explained of the nodes for which we have available measured true activation data. Next we test how well the model can explain the data from the single best target of each node. And lastly we test the performance of the multiple target sets, with only the most confident targets, the medium confidence targets, and the best 15 targets, see methods above, both in the form of replicates and as their precomputed means. We performed all these experiments on both the MAPK and PI3K pathway model and the multi pathway model. An overview of all the targets and data involved in the experiments can be found in Supplementary Table 5.2.

To compare the models, and thus the use of different targets, to each other, we cannot look at the absolute difference or the sum of error between the posterior predictive points and the observed data. This is due to the fact that the spread of the data points is different for each data used. Especially when using more targets in a set the spread of the point cloud is much narrower, making them closer to the posterior predictive mean. However because we want to see how well the model can explain the variance of the data, we compare the variances of the predictive and observed data points. We calculate therefore the correlation between the modeled posterior predictive results of nodes and their observed measurements. A higher correlation means that the variance of the predictive posterior is close to that of the observed data and thus the the model can explain more of the data variance. A bad or no correlation between the posterior predictive and observed data does not directly mean that the estimation of the node activation is faulty, as it could also be the case that just the current model cannot explain the steady state activation correctly.

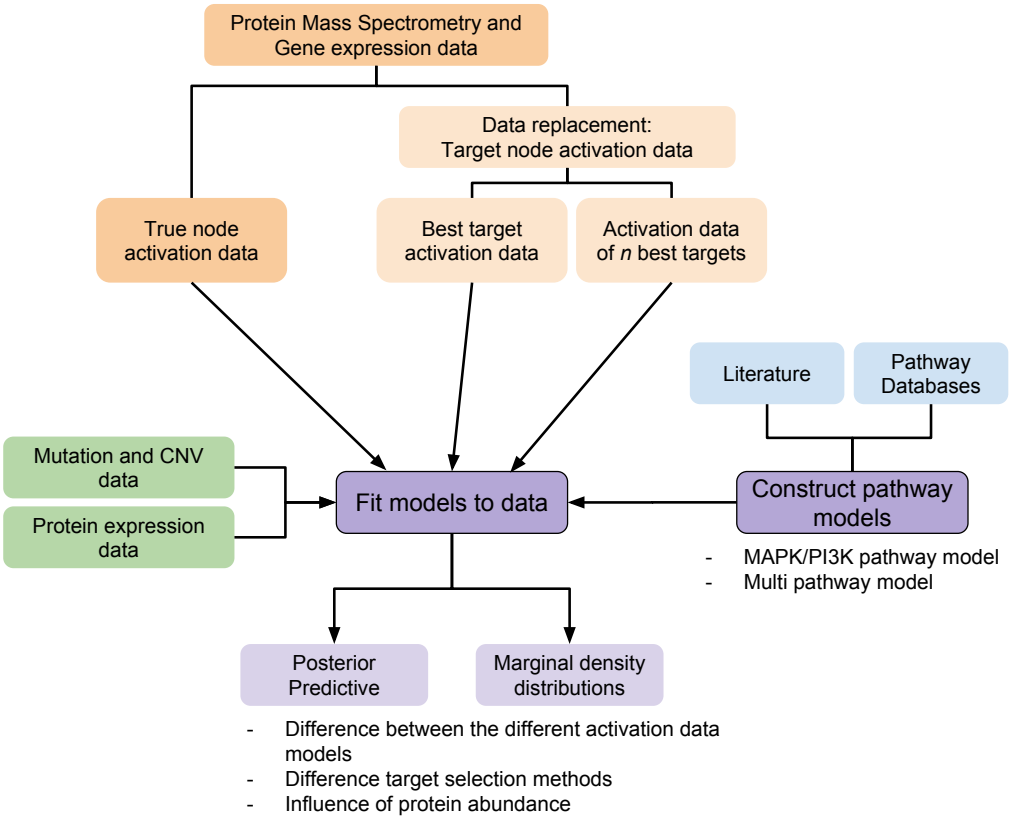


Figure 2.3: Schematic overview of the experimental setup of chapter 2.

2.4 Results

MAPK and PI3K pathway model

We took five nodes from the MAPK and PI3K pathway, described in 1.4.3, to test the target replacement approach, consisting of ERK1 and MTOR, for which true activation data was available, and AKT, S6K, and MEK, that lacked activation data but do have a set of measured phosphorylation targets.

To test how well the our method could reconstruct the true activation of a node, we looked at the correlation between the true activation data of the nodes for which true activation measurements is available and different replacement data approaches, see Table 2.1. The correlation replacement data and the true activation data of the MTOR and GSK3B nodes increases and becomes more significant as more targets are added together. This is however not the case for ERK, CTNNB, and P53, for which none of the correlations was close to significant, meaning that the variation in the target signals is not the same as in the true activation data. This does not mean that the targets are wrong, because they all random samples a common parent, but mainly that they are not influenced by the true activation site or protein abundance (in case of CTNNB and P53). It is also not a predicate that the variance of the true data can be well explained by the model.

Therefore, we employ our pathway model to see if the variation in the true activation data or the target replacement data or both can be explained by the mutations and copy number variation in the signaling pathway model.

Table 2.1: Correlation between the node's true activation data and the four replacement data approaches.
*** p-value<0.01, ** p-value<0.05, '-' p-value>0.5.

	ERK	MTOR	GSK3B	CTNNB	P53
best target	-	0.2753414**	0.2072318	-	-
best n_1	-	0.3377795***	0.2337567	-	-
best n_2	-	0.4406812***	0.3189748**	-	-
best n_3	-	0.522124***	0.4873573***	-	-

We run in total 8 untreated models, one with only the true activation data for the nodes that have that data available, one with alternative data from the nodes' best target, and six models with the three n target groups, modeled using their average and modeled directly as replicates.

All models converged using a burn-in period of 700, sampling period of 1500, sample size of 200, and 1 time temperature adaption of 2000 samples, the computation times ranged from 20 min to an hour, Supplementary Table 5.4. The correlation of the posterior predictive with the observed data points for each of the 8 experimental models using the MAPK and PI3K pathway can be found in table 2.2. For the true data and best target models the phosphorylation site per node is mentioned, for the multiple targets they can be found in Supplementary Table 5.2.

True activation data

In Figure 2.4 A is the posterior predictive of MTOR depicted with blue bars, that describe the modeled posterior predictive sample distribution with 90% confidence, overlayed with the observed data points. For both MTOR and ERK(Supplementary figure 5.1 A), the model was not able to accurately explain the variance of their true PMS activation data. The observed data points do not overlap with the posterior predictive. There are small peaks for the PTEN mutations for MTOR, but the correlation between the predictive and observed data is still small and insignificant, see Table 2.2. The same is true for ERK, where the EGFR mutations create small increases in the activation.

Best target activation data

With taking only the best phosphorylation target of the nodes as activation data, the model can explain the variance of some of the nodes. A part of the variance of the activation of AKT, using FOXO3 S253 as replacement, can be recapitulated by the model, see Figure 2.4 B. The correlation of the predictive and the model is high and significant, looking at the posterior predictive we see that this correlation is the result of correctly modeling the variance of the cell lines with PTEN

Table 2.2: Correlations between posterior predictive and observed data.
*** p-value<0.01, ** p-value<0.05, * p-value<0.1.

MAPK and PI3K model	True data	Best target	Target set:	n_1	n_2	n_3
AKT		0.543***	modeled as replicates:	0.510***	0.263*	0.177
		FOXO3 S253	modeled as average:	0.507***	0.264*	0.178
MTOR	-0.177	-0.001		0.148	0.081	0.055
	MTOR S2448	EIF4EBP1 S65, T68, T70		0.165	0.075	0.063
S6K		-0.312*		-0.018	-0.054	0.017
		RPS6 S236, S240		-0.056	-0.048	0.004
MEK		0.329**		-0.081	-0.075	0.129
		MAPK3 T202, Y204		-0.062	-0.063	0.148
ERK	-0.212	-0.063		0.232	0.195	0.310**
	MAPK3 T202, Y204	STAT5A S780		0.251*	0.207	0.351**

Multi pathway model	True data	Best target	target set:	n_1	n_2	n_3
AKT		0.454***	modeled as replicates:	0.549***	0.323**	0.232
		FOXO3 S253	modeled as average:	0.509***	0.217	0.134
ERK	0.157	0.584***		0.083	0.095	0.155
	MAPK3 T202, Y204	STAT5A S780		0.088	0.083	0.172
GSK3B	0.455***	0.181		0.056	-0.049	-0.165
	GSK3B S9, T7	MYC T58, S64		0.065	-0.123	-0.231
CTNNB	0.250*	-0.083		0.054	-0.101	-0.039
	Protein Expression	CCND1		0.039	-0.115	-0.036
P53	-0.065	0.413***		0.345**	0.474***	0.515***
	Protein Expression	RRM2B		0.244	0.364**	0.406***

mutations, which all have increased AKT activity. The decrease in activation for some cell lines is still unexplained.

We see a similar result for the S6K node, supplementary Figure 5.1 B, with data of RPS6 S236 and S240, and the node MEK1, Figure 2.5 A, using the expression of MAPK3 T202 and Y204. The model is able to correctly explain the direction of the variance for almost all cell lines. Note that this exact phosphorylation site was also used for ERK as its true activation data in the baseline

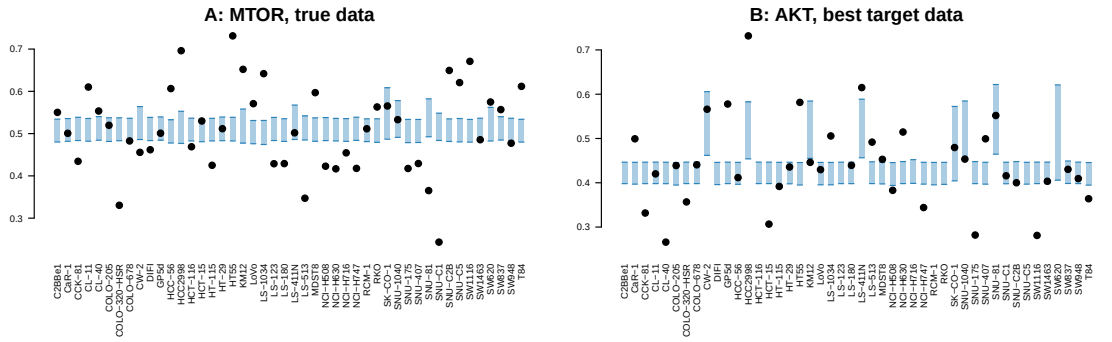


Figure 2.4: Posterior predictive and observed data of A: MTOR using its true data and B: AKT using its best target as activation data. Both for the MAPK and PI3K pathway model.

model, Figure 2.5 B. S6K and MEK both have significant correlations, Table 2.2.

It is very interesting to see that the variation in the expression of MAPK3 T202 and Y204 can be explained by the model when used as activation data for MEK, but not for ERK itself, see Figure 2.5 A and B. An explanation lies in the protein expression($E_{i,j}$) of MEK, as can be seen from the marginal density of the expression parameter p_i in the same figure the abundance of the protein MAP2K1(MEK1) is a limiting factor for the signal transduction.

Using the best target as replacement data does not improve the result for the nodes MTOR and ERK over the use of their true activation data, the correlation is as expected slightly lower.

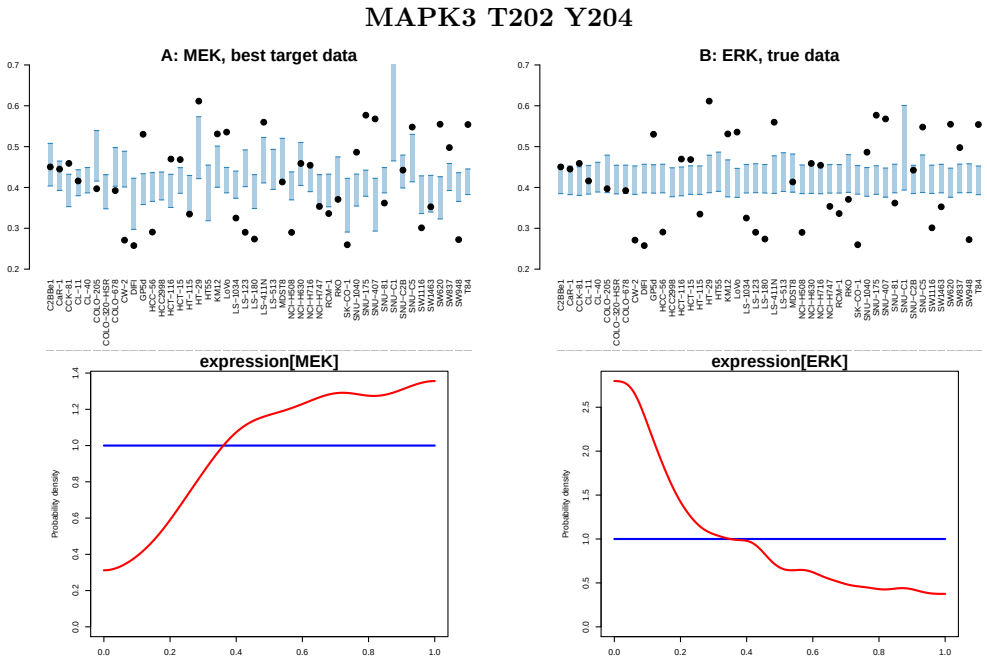


Figure 2.5: Posterior predictive and expression coefficient parameter of A: MEK and B: ERK for the MAPK/PI3K model using MAPK3 T202 Y204 as best target and true data, respectively. B and C show the same thing for the use of STAT5A S780 for ERK as best target in the MAPK/PI3K model and multi pathway model

Target sets activation data

The use of multiple targets as alternative activation data shows to be an improvement over the the true or missing data in specific cases, this can be seen in the last three column of Table 2.2. First, we did not observed a large difference between the target sets modeled as average versus the sets modeled as separate replicates in the model, however the model with separate targets took significantly longer to compute.

The target sets for AKT with the lowest number of, only the most confident, targets(n_1) can

be explained by the model, Table 2.2. But when the less confident targets(n_2 and n_3) are added to the target set the significance and correlation decrease. This difference is harder to see from the posterior and observed data graph, Figure 2.6 A, showing the modeled first and last target set. It can be observed that as the number of targets increase the overall variance of the combined targets becomes smaller, however this does not mean that the model can explain the relative variance better, as showed by the correlation.

A similar pattern but reversed is observed in ERK, Table 2.2 and Figure 2.6 B, the target set with the highest number of targets, so including targets with a low confidence score, can be explained most well.

For the nodes MTOR, S6K, and MEK none of variance of the target sets could be explained significantly. However there is a visible trend for MTOR where the correlation decreases when more targets are used, similar to AKT, and for MEK the correlation increases if the number of nodes increases, similar to ERK.

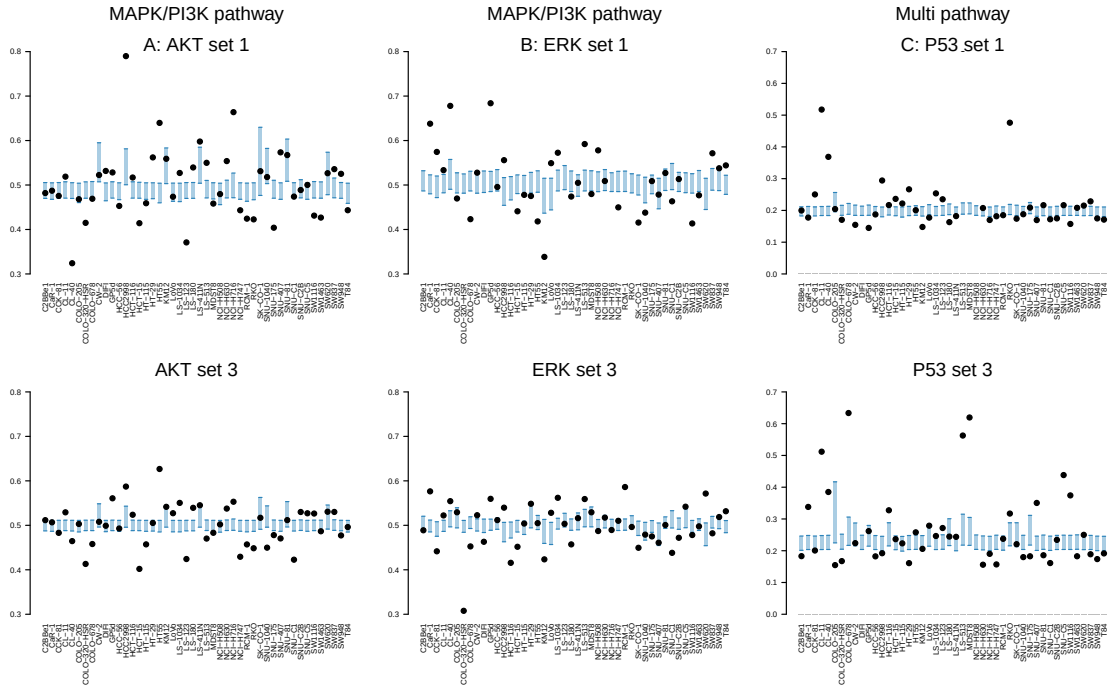


Figure 2.6: Posterior predictive of AKT and ERK in the MAPK/PI3K model and P53 in the multi pathway model, using in the first row their smallest target set, and in the second row the largest

Multi pathway model

We did the same experiment with the multiple pathway model, described at section 1.4.3. Here we focused on the five nodes, AKT, ERK, GSK3B, CTNNB, and P53, for which all except AKT we have a measured activation by PMS phosphorylation expression, for ERK(MAPK3 T202 and Y204) and GSK3B(GSK3B S9 and T7), or PMS protein expression, in case of CTNNB and P53. We run in same 8 models with true activation data, best target activation data, and six target activation data sets.

All models converged using a burn-in period of 700, sampling period of 2000, sample size of 500, and 2 time temperature adaption of 2000 samples. The correlation of the posterior predictive with the observed data points can found in Table 2.2.

True data activation data

Running the model with the true data for the nodes shows us that again the activation of ERK can not be explained, see Table 2.2, the variance of GSK3B however is well explained using its true phosphorylation data. When looking at the posterior predictive of GSK3B and marginal densities we see that this is caused by primarily the mutations on AKT, Figure 2.7, and not via the GFRA node. The decreased activation of some cell lines cannot be explained by the model. A similar

result can be seen for the CTNNB activity, Supplementary Figure 5.1 c, measured by its protein abundance. Also can be observed from its correlation that the protein expression of P53 is not sufficient to use as activation data.

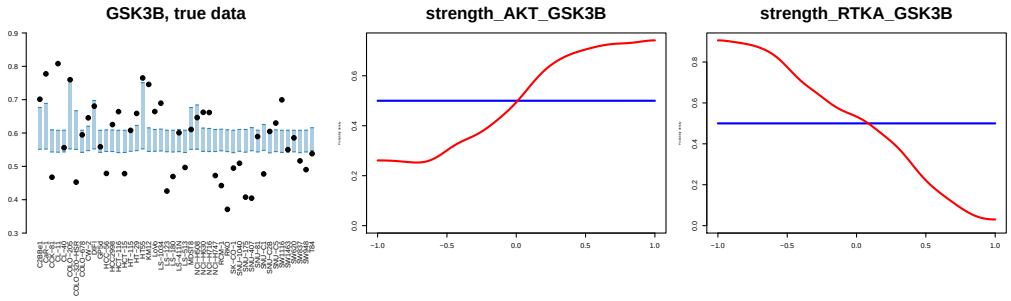


Figure 2.7: Posterior predictive of GSK3β using true data, and the marginal densities of the strength parameters between GFRA and GSK3β and AKT and GSK3β.

Best target activation data

Using the best targets for each node we see that again AKT is well explained using FOXO3. Interestingly to note is that the expression coefficient parameter is now nonnegative, see Figure 2.8 B, while in the previous pathway model it could not be determined, Figure 2.8 A.

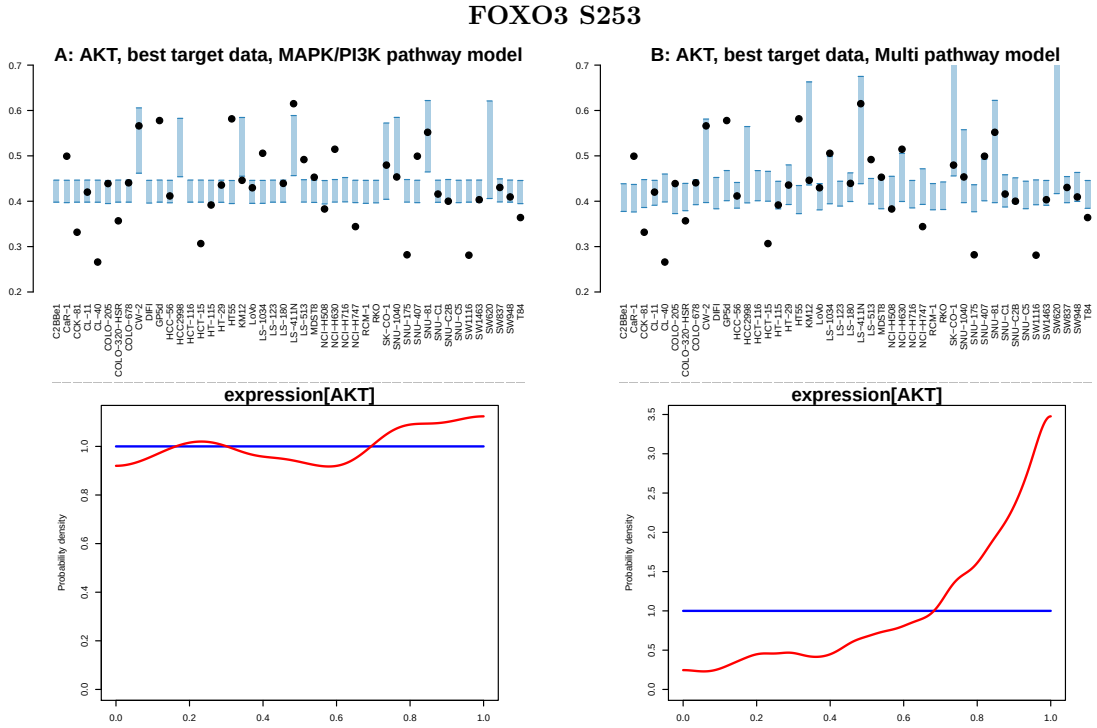


Figure 2.8: Posterior predictive and expression coefficient parameter of AKT using its best target, FOXO3 S253, as activation data for A: the MAPK/PI3K model and B: the multi pathway model

Something similar happened to ERK, while represented by the same data as in the MAPK PI3K model it has now a much better and very significant correlation, see Table 2.2 and figure 2.9 A and B, while in the previous model no variation could be explained. From the marginal density distribution of the expression coefficient in the same figure is becomes clear that now the protein expression does have a large influence.

The reason for this difference in model fit, that we now have seen in both the use of MAPK3 S202 Y204 for ERK and MEK, STAT5A S780 for ERK in both models, and FOXO3 S253 for AKT, is most likely the number and selection of mutations in the model. The MAPK PI3K model

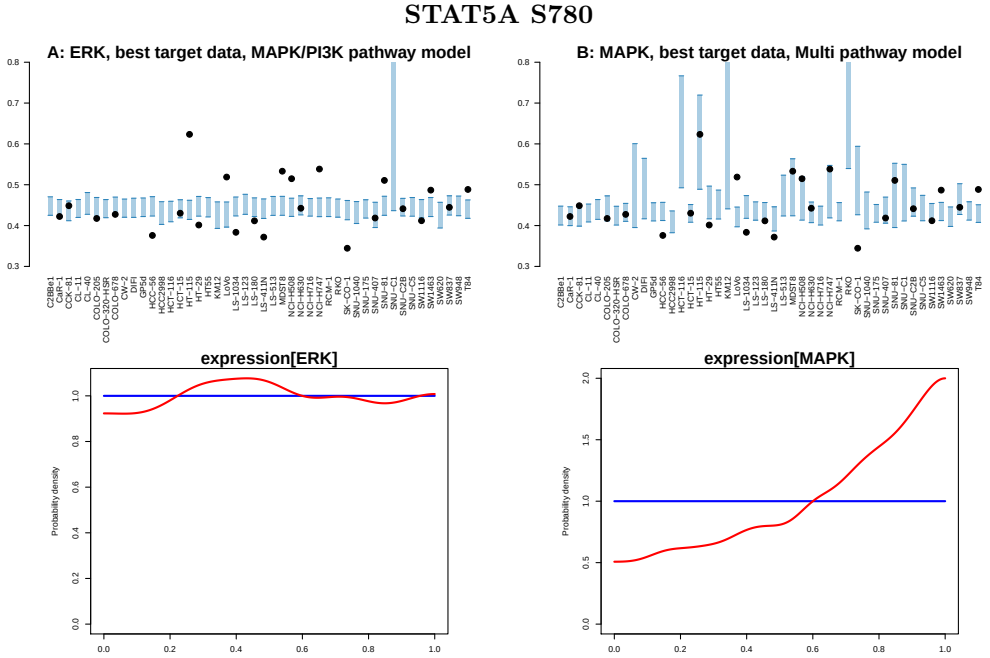


Figure 2.9: Posterior predictive and expression coefficient parameter of ERK using its best target, STAT5A S780, as data for A: the MAPK/PI3K model and B: the multi pathway model

has only 17 mutations and the multi pathway model has 23, including extra mutations that occur often in colorectal cancer.

By replacing the activation data of P53 with the gene expression of its best target RRM2B, Figure 5.1 D, the model can explain its variation and the posterior predictive and observed values have a high and very significant correlation. Doing the same for CTNNB does not work and replacing GSK3B with its best phosphorylation target yields also no success.

Target sets activation data

For the target set models, the before mentioned effect of decreasing correlation is still present for AKT, for ERK however the reversed effect has been lost. P53 is now the one that satisfies this reversed pattern, see Figure 2.6 C and Table 2.2, with increasing number of targets the variation can be better explained by the model. The other nodes, GSK3B and CTNNB, do not profit from the use of the target sets.

Model differences

Summarizing these results, we were able to replace the missing or true data for some of the nodes, including AKT, S6K, MEK and ERK in the MAPK/PI3K model and AKT, ERK, and P53 in the multi pathway model, using different approaches and selections of phosphorylation or transcription targets.

It is interesting to see that the true data of a node can be unrelated to the target replacement data, as with ERK, CTNNB and P53, and that true activation data variation can not always be explained by the model, for example ERK and MTOR in the MAPK/PI3K pathway model. This is the not case for GSK3B and CTNNB, where replacing the true data by its best target or target sets does not improve the fit.

It is not possible to say from these results what replacement approach is right, as it seems to be node specific. AKT is explained better with fewer targets, S6K and MEK perform only well with the best target as adding more targets does not improve the fit, and data replacement for P53 performs better with more targets.

Furthermore can be seen that the importance of the protein expression depends on the model and on the node it is used for, which has a significance influence on the ability of the model to explain the observed data. This is most vivid for the true activation data of ERK, where in the MAPK/PI3K pathway the expression coefficient cannot be determined, but in the multi pathway is was clearly non-negative with a improved fit. This difference can be explained by

the dissimilarities between the two pathway models, the level of abstraction and the number of mutations. By increasing the number of mutations the model has more knowledge on how to explain the activation data variance. However the true cause of why this large difference is not directly identifiable.

Incorporating the targets from the target sets via separate replicates instead of using its pre-computed averages increase the significance and the correlation slightly for almost all nodes. A reason for this is most likely the loss of information when taking the average of the set, in contrast to using all data points. However, the computation time increases with a large number of replicates, the computation of MAPK/PI3K model with target set 3 as replicates took twice as long compared to using the averages, from 35 minutes to 1 hour, Supplementary Table 5.4.

Chapter 3

Colorectal cancer drug response model

To gain insight into the signaling behavior of the colorectal cancer cell lines under treatment of different targeted drugs and to see whether with current knowledge the variance in drug response could be explained, we performed a set of computational modeling experiments, using drug response, genetic, and proteomic data. After the models are fitted to the data we can use the model to see what the mechanisms are that contribute to the drug sensitivity, their relative influence and significance, and how it relates to what is known in literature.

We also looked at the influence of using the target PMS data, replacing the missing node activation data, as described in the previous chapter, on the drug response and model fit. We chose four targeted drugs and one combination drug from our screening dataset that inhibited the MAPK and PI3K pathway. These four drugs are Trametinib (MEK inhibitor), Afatinib(EGFR,ERBB2 inhibitor) , MK2206(AKT inhibitor), and SCH772984(ERK inhibitor). The drug combination is a combination of Tramatinib and Afatinib.

3.1 Experimental setup

In total we created ten drug response models, every drug modeled separately, once using its true node activation data and once using the activation data from the previous chapter that performed most well. To be able to converge the model within reasonable time, we reduced the complexity of the MAPK/PI3K pathway model from section 1.1 by combining all the growth factor receptors to one growth factor activation node (GFRA), Figure 3.2, reducing the number of parameters. By fitting the model to the data we can see what variance of the drug response the model can explain and what it fails to explain based on the modeled pathways and data.

We compare our results, the found mechanism of resistance and sensitivity and genetic associations, with literature and features found by performing an elastic net regression on the same dataset. How the elastic net regression was performed is described below and an schematic overview of the experiments is depicted in Figure 3.1.

3.1.1 Elastic net regression

We performed elastic net regressions for every drug to select which of the genetic features were associated with drug response as measured by the half inhibitory concentrations(IC50) across the cell line panel, based on work by Garnett et al.[17][66].

We calculated the IC50 for every drug and cell line combination by fitting a four parameter sigmoid function on the dose response curve, using the R package *nplr*[67], and created a vector y of IC50s for N cell lines per drug. The feature matrix X, composed as $N \times p$, consists of the p mutations and copy number variations present in the N cell lines.

Using the vector y and matrix X we performed a 10-fold cross validated elastic net regression, with the alpha parameter set to 0.5, to optimize the lambda parameter, using R package *glmnet*[68]. With the optimized lambda the model could be estimated and the variables that are associated with drug response could be found. This procedure was repeated 100 times for each drug to find stable features and to be able to asses their weights. To obtain a measure of confidence we analyze

the frequency at which each feature is present, thus not zero, in the model in all 100 modeling iterations[17]. A frequency of 1 means that the feature was nonzero in all 100 iterations. The elastic net regressions were first performed on the full dataset of genetic features and then on the set of features also selected for the pathway model.

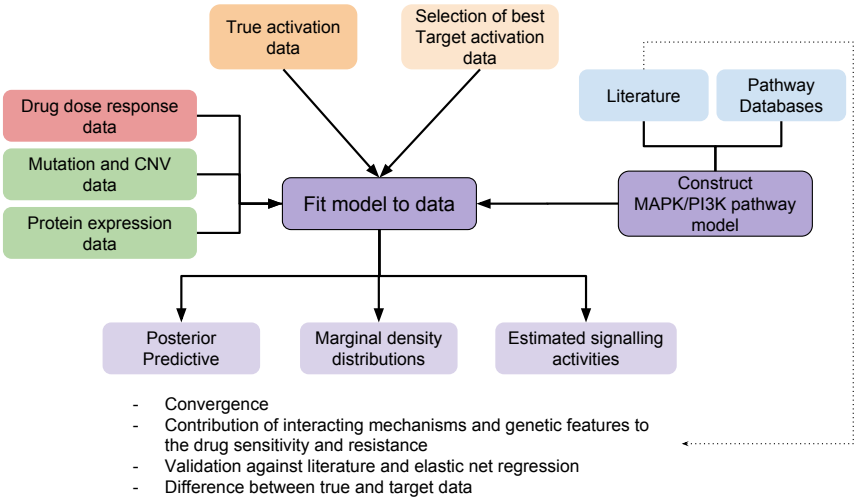


Figure 3.1: Schematic overview of the experimental setup of chapter 2.

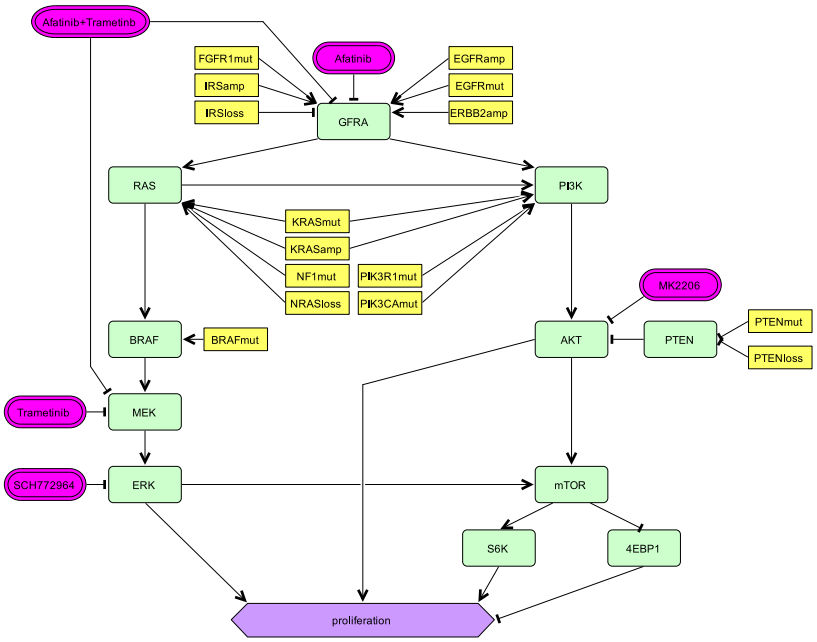


Figure 3.2: Model of signalling pathways MAPK and PI3K as used in modeling of drug response

3.2 Results

We run all 10 models using a burn-in period of 700, sampling period of 2000, sample size of 550, and 3 to 4 times temperature adaption of 2500 samples, the computation times ranged from 18 to 22 hours per model, Supplementary Table 5.4.

Due to the complexity of the signaling pathway models, reaching convergence was not an effortless job for every model, especially the models using only the true target data were hard to converge. We will go over all of the five drugs in the following sections, to report how well the model was able to explain the data and discuss the features and mechanisms that can be observed and how they are validated against literature and the elastic net model results.

Trametinib

The models involving the MEK inhibitor Trametinib were able to properly converge, Supplementary Figures 5.2 and 5.3. The traces of the parameters contain no clusters or were trapped in a single mode and thus were able to explore their solution space sufficiently. There is no autocorrelation between the samples present in most of the parameters and only a few have minimal autocorrelation. The number of round-trips of the chains, going from the lowest temperature to the highest and back again, were sufficiently high, more than 100. We can therefore conclude that the two models converged well and their parameters can be interpreted with confidence.

Both models, with true activation data and target activation data, were able to explain the response of the cell culture to Trametinib accurately for almost all cell lines, see Figure 3.3 and Figures 3.4 a and b. In the first figure is the posterior predictive depicted overlayed with the observed data point for four cell lines, we can see that the model can both explain the response of resistance cell lines, e.g. SNU-81, and sensitive cell lines, e.g. LS-513. However not all cell lines can be explained, a clear example is the sensitive cell line NCI-H716 that is predicted by the model to be resistant. Furthermore, there are also cell lines for which the model did predict the sensitivity but not the right strength, e.g. HCC-56. In the second set of figures we see an overview of the fit of the the two model for every cell line, purple cells indicate that the model predicted the cell line to be more resistant at that concentration, and in orange cells the posterior predictive is more sensitive. There is not a large difference between the two models, the true data model performs slightly better as the total sum difference of the target model is a bit larger than that of the true model, 33.47 versus 31.18.

To see what genetic features influence the drug response we analyze the marginal probability density distributions of the parameter samples, see figure 3.5. In this figures we see, besides the prior distribution (green), both the distributions of the parameters in the true data model(purple) and the target data model(orange). The distributions depict the probability of the parameter value. A distribution leaning to the left, thus 0, means the strength is low (for example figure 3.5 D purple) and to the right, 1, indicates a high influence to the proliferation (for example figure 3.5 A purple). When the distribution does not differ much from the prior then the parameter value could not be determined by the model (for example figure 3.5 A orange), note that this does not mean that the its value is zero.

So what we are looking for in these probability density distributions, when searching for important features, are the feature's strength parameters that have a nonzero probability with high confidence. The value of the strength parameter than indicates the how much of the proliferation signal arises from this genetic feature.

When we have a look at the marginal densities of these genetic features in both the true data model and target data model, we can see that there is some cases a difference in the strength parameter value. The large influence of the PIK3CA mutation (A) is captured in the true model but could not be determined in the target model. The Target data model assigns a slightly higher strength to the genetic features on the growth factor receptors, EGFR, FGFR, and IRS2 (D, E, and F, respectively). Both models roughly agree on the strength of the BRAF and PTEN mutations. The marginal densities of the other parameters can be found in Supplementary Figure 5.7.

We can compare these results of our model to the general method for finding and identifying genetic features that influence the drug response, elastic net regression. The results for the elastic net model using all genetic features can be found in Supplementary Table 5.3, and using only the features also used in our model in Table 3.1, the features with a nonzero frequency larger than 0.9 are highlighted. When the elastic net regression is performed on all genetic features, which is a common approach for finding associations[17][18][19], it fails to select the important

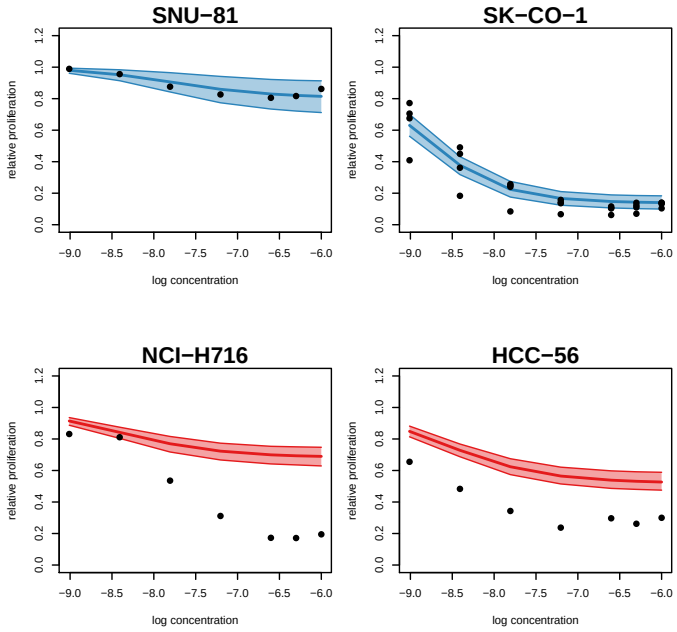
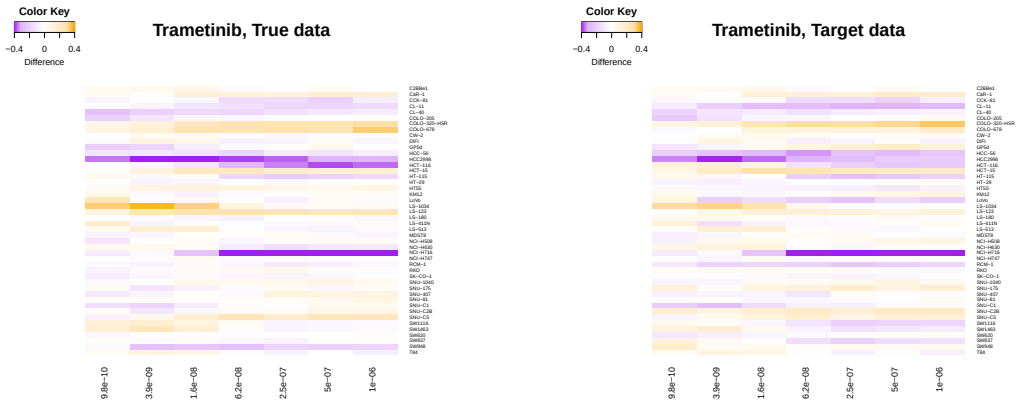


Figure 3.3: Posterior predictive (in blue) with the upper and lower boundaries a 90% confidence interval, overlayed with the observed data of the cell lines SNU-81, LS-513, NCI-H716,HCC-56 in the Trametinib target data model. The blue predictives accurately describe the measured data while the red posterior predictives are more resistant than the observed data points.



(a) (b)

Figure 3.4: Overview heat plot of absolute differences between posterior predictive and observed data points, for each cell line in the Trametinib models, over the 7 drug concentrations, from low to high

features, except for a few. When the Elastic net is performed on the set of genetic features that we selected for in our signaling pathway model, it is able to find the features with a stronger influence explaining the drug response variation.

To compare the results of the elastic net regression and the our model we cannot directly examine the regression weights in contrast to the found strengths parameter distributions of our model, as they are not the same entity. The real numbered coefficients of the genetic features in the elastic net regression describe directly the influence of the features explaining the variation in the IC50 data. The absolute value of the feature coefficient indicates the influence strength and feature coefficients with a value of 0 have no predictive power. The feature strength parameters in our model have a probability distribution between 0 to 1 and these distributions represent the strengths of the proliferation signal arising from this genetic feature. A high peak in the parameter

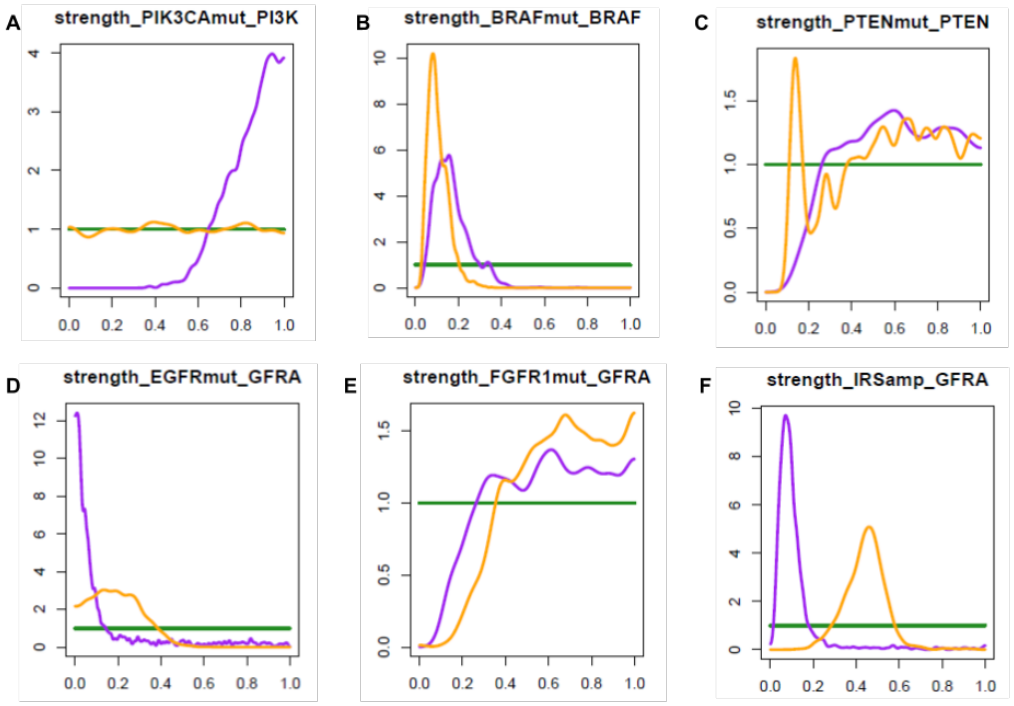


Figure 3.5: Tramatinib: Marginal density distributions of the strength parameters of A: PIK3CA, B: BRAF, C: PTEN, and D: EGFR mutations, and E: FGFR and F: IRS amplifications. The green line is the prior uniform distribution, the true data model is depicted in purple and the target data model in orange.

Table 3.1: Features weights of Elastic Net Regression using selected features. Highlighted features have a nonzero frequency higher than 0.9.

	MK2206	Trametinib	Afatinib	SCH772984	Afatinib+Trametinib
(Intercept)	**-5.568**	**-8.002**	**-5.374**	**-6.264**	**-6.548**
EGFR	0.000	**0.527**	-0.001	0.000	-0.038
KRAS	0.000	0.000	0.000	**0.033**	-0.031
BRAF	0.000	**-0.212**	0.000	**-0.183**	**-0.139**
NF1	**0.565**	**0.073**	-0.001	0.000	**0.091**
PTEN	0.000	**0.089**	0.000	**-0.087**	0.069
loss PTEN	**11.701**	0.000	0.000	0.000	0.020
PIK3CA	**-0.086**	**0.524**	**-0.136**	**0.084**	**0.093**
PIK3R1	0.000	0.000	0.000	-0.002	-0.070
loss NRAS	0.000	0.000	0.003	0.000	0.066
amp KRAS	0.008	**0.098**	0.000	0.000	0.024
loss IRS2	0.000	**0.320**	0.000	**0.150**	0.036
amp IRS2	0.000	**-0.477**	0.000	**-0.057**	**-0.206**
amp FGFR1	0.000	0.000	**-0.300**	0.000	**-0.687**
amp ERBB2	0.000	0.000	-0.003	**-0.012**	0.000
amp EGFR	0.000	0.000	**-0.329**	**-0.019**	**-0.337**

distribution implicates a high confidence in the parameter value. Although the models do not have a directly quantitative compatible results, they both give an indication on what genetic features are of influence on the growth of the cell line and an idea of the their strength. Therefore, these indications, their relative strength, and confidence are what we can compare between the two models.

If we take look at model results of the Trametinib models, we see that both the elastic net, Table 3.1, and our model, Figure 3.5, were able to identify the proliferation signal arising from the PIK3CA mutation and IRS amplification. The coefficient of the PIK3CA mutation in the

elastic net results is 0.524, which is nonzero, and the marginal density distribution of the strength of the PIK3CA mutation on PI3K indicates (for the true data model) a high influence on the proliferation, with also a high confidence. The same goes for the IRS amplification, which shows only a lower strength in the density distribution, but still a high peak of confidence. Likewise, the elastic net also indicates the predictive power of the IRS amplification.

The influence of the BRAF and EGFR mutations can be seen by both methods as well. The elastic net shows that the importance of PTEN mutations and FGFR amplifications is non existent, with a coefficient equal or close to zero. Our model gives a different result, but because it generated the distribution of the random variable parameters we can see that the strength parameters are nonzero and have a influence on the response, however their exact strength can not exactly be determined by the current model. The other marginal density distributions can be found in supplementary figure 5.7, where can be found that also a proliferation signal arises from PIK3R1 mutations and EGFR amplifications have influencing factor on the proliferation that can not be determined by the elastic net regression.

From the several mechanisms of resistance to MEK inhibition treatment described in literature, the most involved is the reactivation of proliferation by downstream MAPK signaling[69], caused by a feedback activation of EGFR in the cell lines with mutations in NRAS, MEK2, and EGFR[70][71][72][73]. Studies found that BRAF mutant cells are highly sensitive to MEK inhibition and that RAS mutant cancer cells respond by activating the PI3K pathway[72]. Where PIK3CA mutations reduce the sensitivity and PTEN mutations cause resistance to the MEK inhibitor. Our model was able to find these influences, of PIK3CA, BRAF, and PTEN(weak), without having prior knowledge on the importance of these mutations. From the marginal densities we see that the PIK3CA mutations and PIK3RI amplifications influence the proliferation signal, indicating the presence of the resistance mechanism. Our model could not completely determine the mechanism of PTEN mutations causing resistance, using the current data, only that its influence is not absent.

To see the difference in signaling and proliferation caused by genetic features between the cell lines we use the signaling estimates of the nodes, which are the estimated values of the signaling activities of the nodes, $A_{i,j}$, and r_j in case of proliferation. In Figure 3.6 we show a scatter plot of the signaling estimates of the proliferation under treatment with Trametinib against the untreated condition. These signaling estimates of the treated condition are not based on measured activation of the molecule under treatment, but inferred by the model using the measured activation data from the untreated condition and the drug response data of the treated condition. From the figure can be observed that the BRAF mutated cell lines (in green) are sensitive to the targeted drug Trametinib and the PI3KCA and PTEN mutated cell lines (in purple and orange, respectively) are resistant, validating the literature.

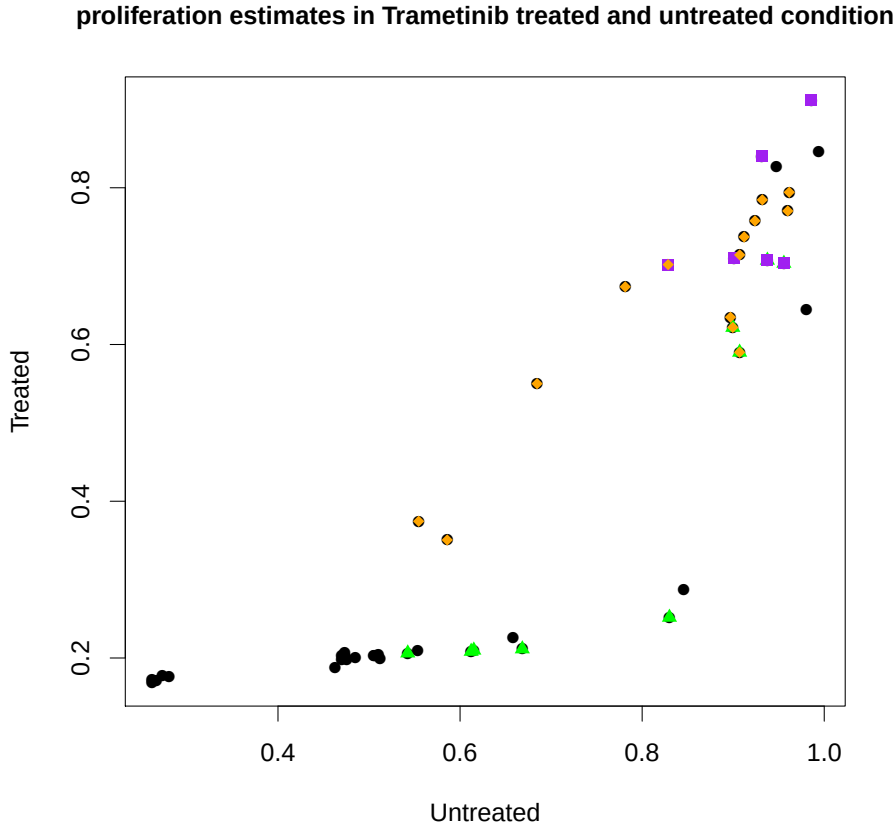


Figure 3.6: Signalling estimate for the proliferation node in the trametinib true data model, points in purple, orange, and green are respectively cell lines with mutations in PIK3CA, PTEN, and BRAF.

MK2206

The target activation data model with AKT inhibitor MK2206 was able to properly converge, although in the true activation data model some parameters kept being stuck in modes and was therefore not able to completely sample the solution space. Nonetheless the autocorrelation was still small, but the number of round trips was low for the true data model. The parameter density distributions of the true data model must consequently be interpreted with a lower confidence.

Both models still do a good job predicting the drug response, Figure 5.5 A. The total absolute difference between the posterior predictive and observed data was slightly higher in the target model than the true model, 25.43 and 23.36 respectively.

From the marginal densities, Figure 3.7 and Supplementary Figure 5.8, we can learn that the mutations and amplifications in the growth factor receptors, including EGFR and FGFR have an influence on the proliferation and so do the NF1, PIK3CA, and BRAF mutations. The two models do not always agree on the parameter densities, but the target model had a better convergence and is therefore better trusted.

The elastic net regression find only the NF1 mutation and a small influence of the PIK3CA mutation, Table 3.1, besides the large influencing loss of PTEN. The importance of loss of PTEN is not recapitulated by our model, as it is not able to determine its value, most likely due to the fact that only one cell line has this feature, what probably also causes the high coefficient value in the elastic net result. Because the PI3K pathway is inhibited by the drug, activating mutation in the MAPK pathway lead to higher proliferation and thus resistance to the treatment, Figure 3.8.

Our findings do not completely agree with the literature, which states that cell lines with loss of PTEN or PIK3CA mutations are significantly more sensitive to MK-2206[74][75]. This mechanism can only roughly be verified in our models for the PIK3CA mutations, see Figure 3.8, where the estimates of the proliferation are depicted for the cell lines with and without the mutation or

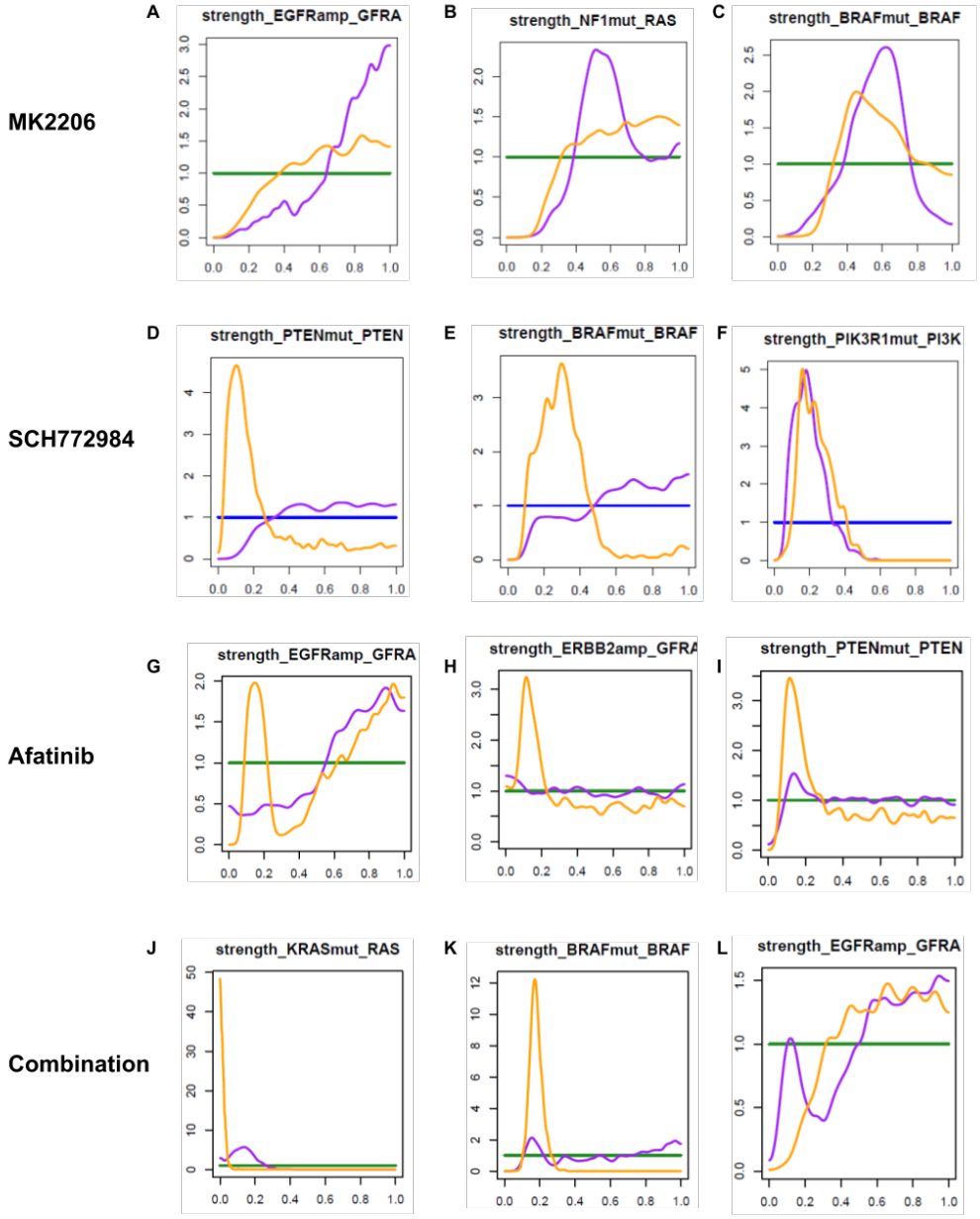


Figure 3.7: Sets of 3 marginal density distributions, for the models MK2206, SCH772984, and the drug combination, of the strength parameters of A: Loss of EGFR, B: FGFR amplification, C: BRAF mutation, D: PTEN , E: BRAF, and F: PIK3R1 mutations, G: EGFR and H: ERBB2 amplificatin, I: PTEN mutatin, J: KRAS mutation, K: BRAF mutation, L: EGFR amplification. The green line is the prior uniform distribution, the true data model is depicted in purple and the target data model in orange.

amplification. The loss of PTEN and PTEN mutations however are more slightly more resistant, according to our model. Furthermore it is demonstrated by other studies that the PI3K pathway is a effective therapeutic target for NF1-mutant cancer cells[76], however this effect is not recapitulated by our model as there is not a difference between the NF-1 mutant and non mutant cell lines, while the elastic net does find the NF1 mutation to be of influence on the drug response. Other studies found that KRAS mutant cancer cell Lines are unresponsive to MEK Inhibitors[77], which our model validates as the KRAS mutated cell lines are estimated more resistant. From the density distribution we can also see that the ERK, 4EBP1, and AKT nodes have the highest direct strength on the proliferation, while the strength of S6K is relatively low, Supplementary Figure 5.8.

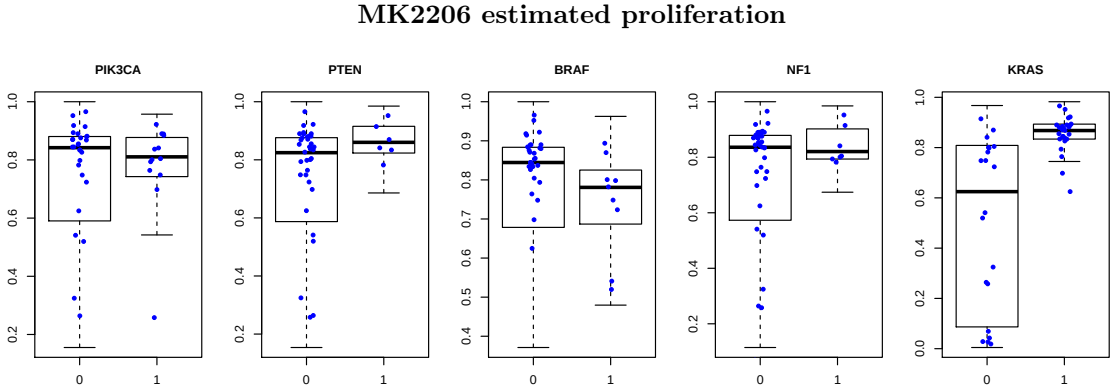


Figure 3.8: Signalling estimate boxplots for the proliferation node in the MK2206 target data model under treated conditions, for 5 different features, PIK3CA, PTEN, BRAF, NF1, and KRAS mutations. The right graphs are the cell lines with the feature and the left graphs are without, the vertical axis is the estimated proliferation.

SCH772984

The ERK inhibitor (SCH772984) models converged well, no modes are visible in the traces and only a few parameters have little autocorrelation. The difference heat maps of the SCH772984 models can be found in supplementary Figure 5.5 B, the total difference for the true model is again is bit lower 24.74, than the target data model, 26.18. The models were both able to predict a large number of the cell line responses, except for NCI−H716 and HCT−116.

Based on the marginal densities we can derive that the proliferation signal arises from features including, BRAF, EGFR, PTEN, and PIK3R1 mutations, loss of NRAS, and EGFR amplifications, Figure 3.7 and Supplementary Figure 5.9. The elastic net regression model found only the BRAF mutation and IRS2 amplification to be the important factors, and with a weaker strength, the PTEN and PIK3CA mutations. Our model similarly found the influence of the BRAF and PTEN mutations, but was not able to determine the importance of IRS2 amplification and found PIK3CA mutations to be more important than PIK3R1.

Resistance to ERK inhibition is caused by the absence of ERK-dependent negative feedback activating RAS and PI3K signaling[78]. Therefore enhanced PI3K pathway signaling, by PIK3CA, PI3KR1 and PTEN mutations, are associated with resistance to the ERK inhibitor[79], so are RAS signaling enhancers, such as KRAS and NRAS aberrations. We can find these mechanisms back in the signaling estimates of the proliferation, Figure 3.9, shown for the features in the PI3K pathway, PIK3CA, PTEN, and KRAS mutations, and NRAS loss. It can be observed that there are KRAS mutants that are sensitive to the drug, but also a large cluster that is resistant. Furthermore is it interesting that the nodes that have a high direct influence on proliferation are ERK and S6K, and the strength of 4EBP1 on proliferation is very low, Supplementary Figure 5.9, which is the opposite of the MK2206 models where 4EBP1 was most determining.

Afatinib

Both Afatinib models were hard to converge, especially the true data model, which exhibits a large amount of autocorrelation between the parameter samples and in the traces of modes are clearly visible, Supplementary Figure 5.4. The target data model did much better, modes where only visible in a single parameter and the autocorrelation was low.

Afatinib is a very effective drug, as only two cell lines are truly resistant, making it hard for the model to the explain the small variance between the cell lines. The models also fail to predict the resistance for these two cell lines, C2BBe1 and COLO−678. Furthermore the posterior predictive fails for a set of cell lines to follow the precise dose response, supplementary Figure 5.6 A.

If we look at the marginal density distribution it is surprising that there is not a big difference between the two models, knowing that the true model did not converge well. As expected for the EGFR/ERBB3 inhibitor, the EGFR amplification has a influence on the proliferation signal, figure 3.7 G, causing slight resistance, 3.10. Besides small influences of PTEN and PIK3CA mutations and ERBB2 amplification, all genetic features have zero influence on the drug response. The elastic net regression found the importance of EGFR amplifications, PIK3CA mutations,

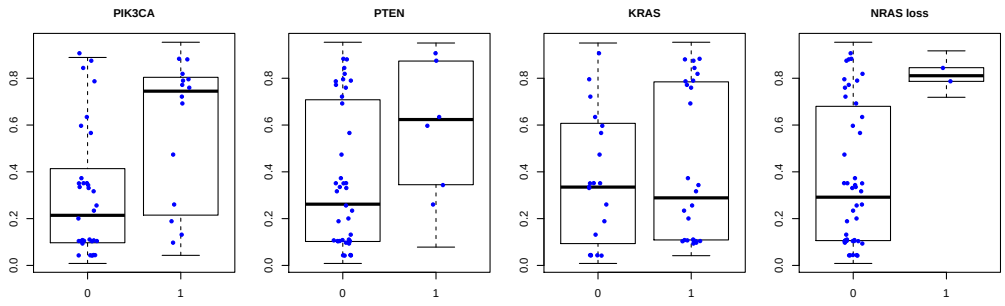


Figure 3.9: Signalling estimate boxplots for the proliferation node in the SCH772984 true data model under treated conditions, for 4 different features, PIK3CA, PTEN, and KRAS mutations, and loss of NRAS. The right graphs are the cell lines with the feature and the left graphs are without, the vertical axis is the estimated proliferation.

and FGFR amplifications. Only the latter was not captured by our model to be of importance. The effect of ERBB2(HER2) overexpression on the Afatinib treatment was also found by other studies[80][81][82]. EGFR mutations did demonstrate promising clinical efficacy in colorectal and lung cancer patients[83][84][85], this effect is also captured by our model. Reported resistance to EGFR inhibitors caused by PIK3CA[86] mutations could no be verified, the PIK3CA mutated cell lines where even estimated to be more sensitive according by our model.

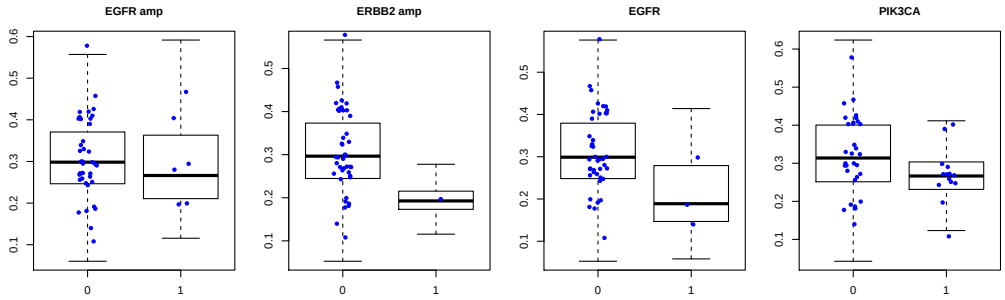


Figure 3.10: Signalling estimate boxplots for the proliferation node in the Afatinib target data model under treated conditions, for 4 different features, EGFR and ERBB2 amplification, EGFR, and PIK3CA mutations. The right graphs are the cell lines with the feature and the left graphs are without, the vertical axis is the estimated proliferation.

Combination of Trametinib and Afatinib

The drug combination model was difficult to converge, for both models some parameters still got stuck in certain modes, which is visible in the traces and the autocorrelation. A reason for this could be that there is not enough variation in the drug response data, because the number of cell lines that was resistant to this combination of Trametinib and Afatinib was limited, only COLO−678 and NCI−H630, all other cell lines had more than 50% proliferation reduction. Both models were able to predict the drug response well, supplementary Figure 5.6 B. But failed to predict the resistance of the only resistant cell lines.

Almost all genetic features had a low but nonzero strength, with relative high values for EGFR amplification, PIK3CA and PTEN mutations, and IRS2 amplification, Figure 3.7 and Supplementary Figure 5.11. A similar result is been recovered by the elastic net regression, where also a lot of features have low coefficients and higher values for BRAF mutations and EGFR and FGFR1 amplifications.

The combination of the two drugs is successful because the feedback reactivation to EGFR when treated with only Trametinib is blocked by the addition of the EGFR and ERBB3 inhibitor Afatinib, which also suppresses AKT signaling[77]. Sun et al.[77] describe that MEK inhibitors,

Trametinib and Selumetinib, show strong synergy with Afatinib in different KRAS mutated colon and lung cancer cell lines. However, the influence of AKT mutations on the proliferation was very small in our models, see Figure 3.7 J. They found also that, besides KRAS mutations, the expression of ERBB3 was predictive for the response. We could not verify this with our model, because we combined all growth factor receptor nodes. The features that had a small influence on the proliferation, PIK3CA, EGFR, PTEN, and NF1 activating mutations, all affect the PI3K pathway and are therefore biomarkers for the increased resistance to the EGFR and MEK inhibition, Figure 3.11.

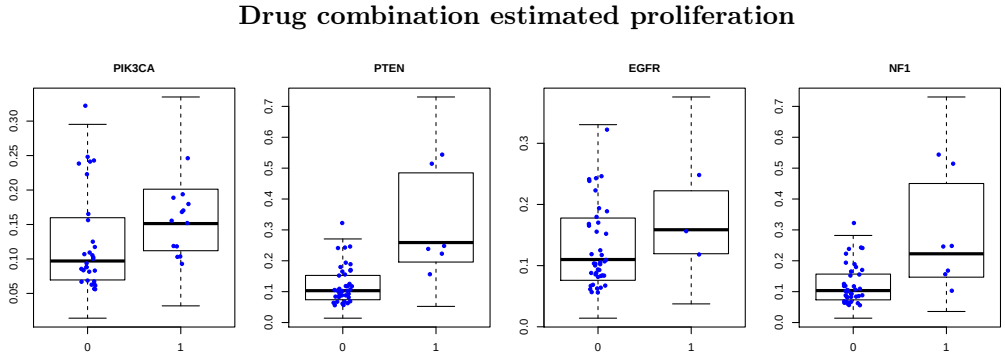


Figure 3.11: Signalling estimate boxplots for the proliferation node in the combination target data model under treated conditions, for 4 different features, PIK3CA, EGFR, PTEN, and NF1 activating mutations. The right graphs are the cell lines with the feature and the left graphs are without, the vertical axis is the estimated proliferation.

Findings and drug differences

Almost all models were able to converge, the best converged and therefore most credible models are the trametinib and SCH772984 models. The MK2206, Afatinib, and drug combination models had some trouble but their results were still interpretable.

The mechanisms of resistance to the different drugs were visible and could be verified against the literature and compared to a elastic net regression model. Most resistance is caused by activating mutations and amplifications in the pathway that is not inhibited by the drug. These features create an high proliferation signal that neutralizes the therapy. If the cancer cell relies on over expression of a gene, such as EGFR or ERBB2 amplification, than inhibiting these nodes will reduce the proliferation of the cells, as seen in Afatinib.

The differences between the true data and the replacement target data set are reflected most in the marginal density distributions. Due to the use of different target data that tries to reconstruct the true activation of the node the genetic features sometimes have a different influence on the proliferation than under the use of the true activation data. However for most parameters in the models the target and true data generate similar parameter distributions.

It is not easy to say whether our model outperforms the elastic net regression in finding the genetic features that are associated with resistance or sensitivity of the cell lines to a certain targeted drug, as our model is much more complex and takes days to compute where the elastic net is done within seconds. However, instead of one coefficient, our model provides the probability density distribution of the parameter value from which the strength and the confidence of the strength can be determined. Furthermore, the pathway model also indicates if a parameter cannot be estimated by the model. The elastic net is more sensitive to outliers and assigns high coefficients to features that are only present in a single cell line, while the pathway model takes this into account.

An overview of the sum squared error of all models on every cell line is visualized in Figure 3.12. Trametinib has trouble fitting more cell lines than SCH772984, while they both inhibit the same pathway. Afatinib and the drug combination of Afatinib and Trametinib cannot properly fit the same three cell lines, but the model was better able to explain the variation in the drug combination than in Afatinib alone.

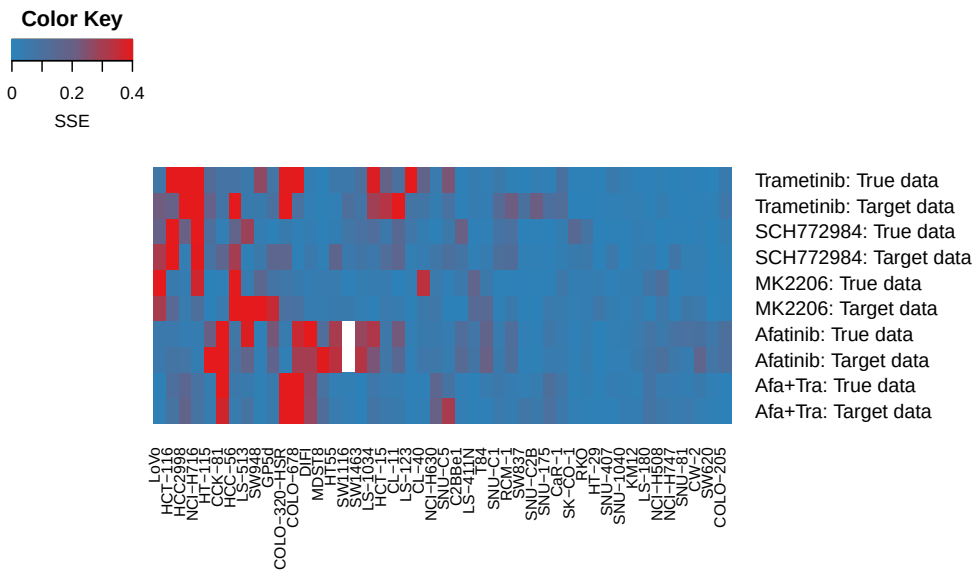


Figure 3.12: Heatmap of the sum squared error of all cell lines in each drug model, clustered on the cell lines.

Chapter 4

Colorectal cancer subtype analysis

In this last chapter we delve into the differences between colorectal cancer subtypes. These subtypes are used to characterize the cancer commonly based on the tumors molecular characteristics and are clusters of CRC cells that are genetically and sometimes functionally similar. We will look at three ways of clustering the CRC cell lines, classification in two groups by micro satellite stability, in four groups by the consensus molecular subtype clustering, or three groups using the colorectal cancer subtypes of De Sousa et. al[87].

All these three sub typing methods of CRC distinguish the cell lines based on there genetic and molecular characteristics. Therefore it is interesting to see whether the pathway mechanisms and responses to targeted drugs can also be differentiated between these subtypes.

To test this we will use the modeled estimates of the signaling node activities in the different pathways of our model. By comparing the activation of the nodes between the cell lines, classified with a certain subtype, we can see whether one or more subtypes have a higher or lower activated signaling pathway compared to the other subtypes.

We will investigate how large these differences are and try to see if a subtype of a cell line could be recognized based on its pathway activations and whether the specific subtypes are more sensitive or resistant to a chosen targeted drug.

4.1 Experimental setup

For these experiments we use the multi pathway model, see section 1.1, using the best activation data for the model components found in chapter 2.

The classification of the micro satellite stability of the cell lines was obtained along with the data from the GDSC dataset. The Consensus Molecular Subtype classification of the cell lines was generated using the CMS prediction algorithm of Guinney et al.[48] on the gene expression data, consisting of four subtypes, CMS1 till CMS4. The algorithm was both run on the normalized and not normalized gene expression. The third cell line clustering method is the Colon Cancer Subtypes (CCS) by de Sousa et. al.[87], which clusters the cell lines in three different subtypes, based on the expression of 146 genes.

To see whether the subtypes differ we look at the differences between the distributions of the estimated signaling activation data points from the cell lines per subtype. These data points are the means of the 2000 samples that where generated for each activation estimate. We asses the differences between the distributions of the subtypes using anova and Tukey HSD to find means that are significantly different from each other.

It must be noted that with 46 cell lines, which is a large number for signaling pathway modeling but not for statistical analysis between the cell lines, the difference between the the points is not easily significant. Especially when the number of cell lines per subtype set is even lower and decreases with the number of subtypes used per clustering method.

Four models were estimated, one with no drugs, similar as in chapter 2 as steady state, and three with the targeted inhibitors MK2206, Trametinib, and SCH772984. By adding the targeted drugs to this model we are also able to observe whether the response data could be explained better or worse with the larger model consisting of more pathways and whether the specific subtypes are more sensitive or resistant to the treatment.

We will focus on the activation estimates of the following nodes, MAPK, GFRA, AKT, S6K, 4EBP1, CTNNB, GSK3B, SMAD, P53, and the proliferation.

4.2 Results

4.2.1 Untreated model

The untreated multi pathway model using the best target data converged well, with no traces nor autocorrelation. We will look separately at the three sub typing approaches, starting with the micro satellite instability.

Micro satellite stability

The difference between the micro satellite stable and instable cell lines for the ten model nodes is illustrated in Figure 4.1. It can be observed that the largest differences between the subtypes are in the PI3K and WNT pathway. However none the differences between the means of subtypes are significant. We can therefore only report on the visible trends.

It is interesting to see that the estimated proliferation, without treatment, is lower for the micro satellite instable cell lines, which are characterized by a high mutation load. The signaling trough AKT seems to be higher for the MSI cell lines compared to the MSS cell lines, with the effect that the S6K is activated and 4EBP1 inhibited. The activity of β-catenin and P53 seems to be higher for the MSI cell lines, which corresponds to the fact that in MSI cell lines β-catenin and APC, are one of most frequently altered genes[88]. There is no visible difference between the subtypes in the MAPK pathway, a reason for this could be that MSI cell lines have a high number of BRAF mutations and MSS cell lines have often mutations of the KRAS oncogene[88][89].

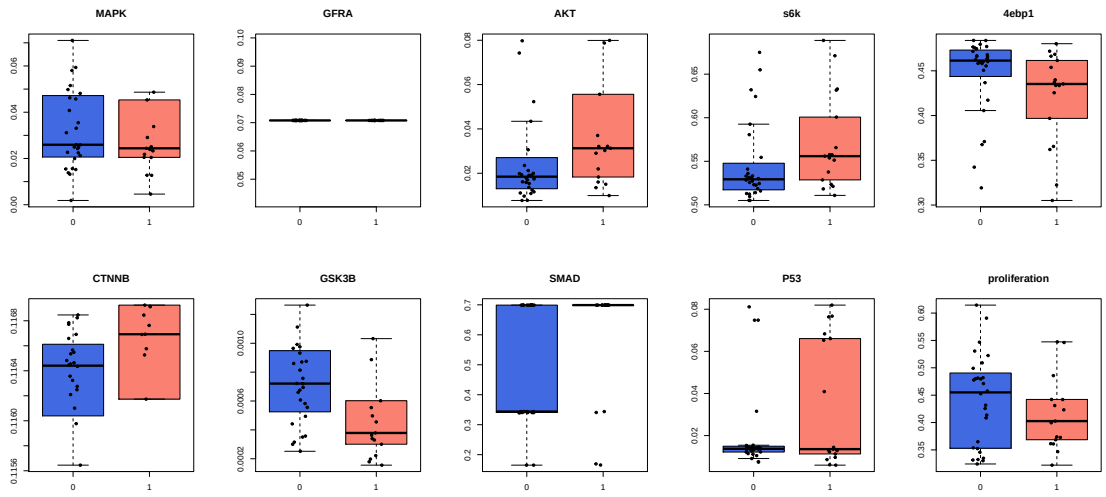


Figure 4.1: Signalling estimates of seven pathway nodes and the proliferation, classified on the cell line's micro satellite instability status. The blue and red box plot represent respectively the MSS and MSI subtype.

Consensus molecular subtypes

To difference between the four concensus molecular subtypes was not very large, Figure 4.2, but still it is possible to capture the mechanisms that are at play. From Dienstmann et. al.[47] we know that the CMS2 tumors have a higher activation of the WNT pathway, which seems to be recapitulated by the model, see CTNNB and GSK3B.

The CMS1 subtype is characterized by its micro satellite instability and enrichment for BRAF mutations, causing a high activation in the MAPK pathway[47]. Similarly the CMS3 tumors are enriched for KRAS-activating mutations and mutation in PI3K[90], resulting in a higher activation of the MAPK and AKT pathway.

The CMS4 cell lines seem to have a low overall signaling, the variance of the points is often small except for the proliferation node. CMS4, together with CMS2, consist of a majority of microsatellite stable cell lines[47], which have similar to the MSI/MSS subtype above a lower AKT activation and a larger variance in the proliferation.

Also the CMS4 tumors are characterized by the activation of pathways related to the epithelial to mesenchymal transition[47], such as the TGFβ pathway, however this mechanism could not be recapitulated by the model.

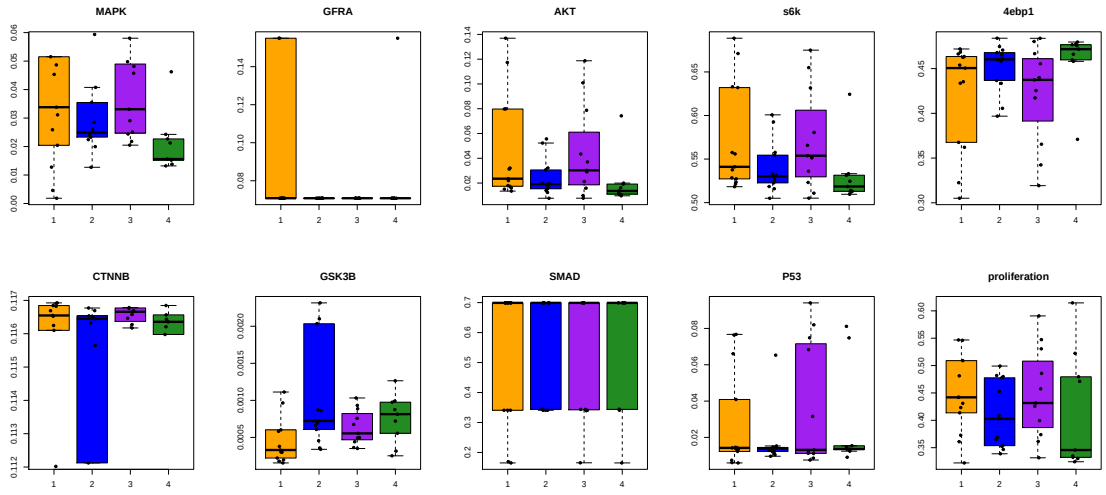


Figure 4.2: Signalling estimates of seven pathway nodes and the proliferation, classified on the Concensus Molecular Clustering subtypes. The box plots are from right to left, CMS1 to 4.

Colorectal Cancer Subtypes

The classification by de Sousa et al.[87] has three different subtypes, CCS1, characterized by KRAS and TP53 mutations in chromosomal instable tumors, CCS2, represented by MSI and CpG island methylator phenotype (CIMP) cell lines, and CCS3, which contained a large proportion of patients with BRAF and KRAS mutations and is heterogeneous with respect to MSI and CIMP. The differences between the subtypes is not large, Figure 4.3. Only at GSK3B has CCS3 a bit lower activity and its proliferation is slightly higher.

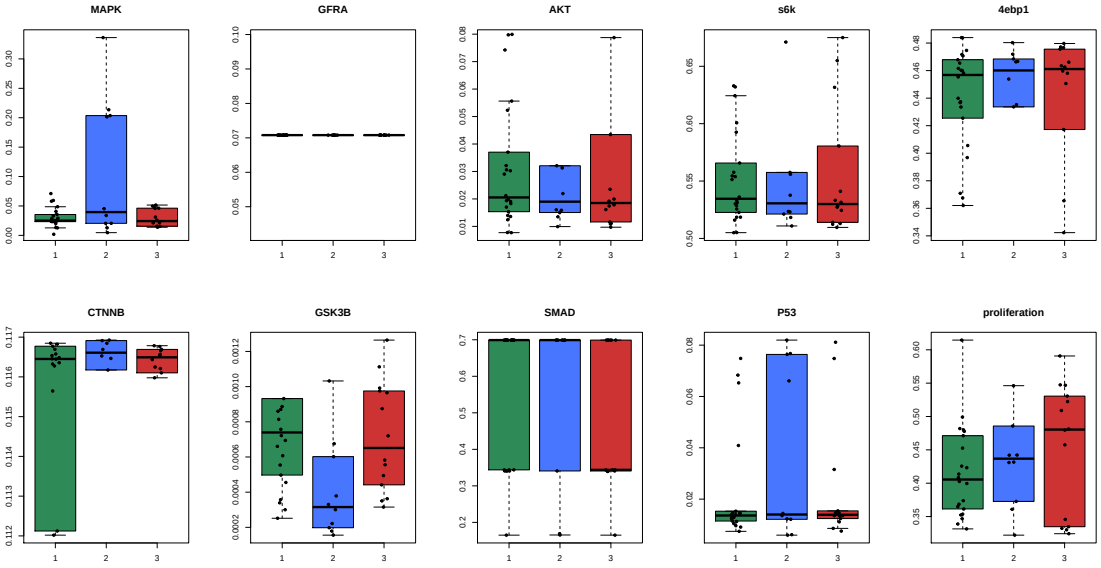


Figure 4.3: Signalling estimates of seven pathway nodes and the proliferation, classified on the Colorectal Cancer Subtypes[87]. The box plots are from right to left, CCS1 to 3.

We can conclude that the untreated steady state model is able to distinct some differences between the subtypes of the three classification methods and the mechanisms can be explained by the literature. However, the trends are still small and mostly insignificant.

4.2.2 Treated model

Continuing with the treated models, the Trametinib model was able to converge properly. But not the MK2206 and SCH772984 models, from which several parameters were trapped in a single or more modes. We will therefore not extensively discuss their results as the confidence in the parameter values is low.

The first thing that can be observed is that the multi pathway model of Trametinib did a better job explaining the drug response data than the MAPK/PI3K models did. Every cell line was predicted accurately, Figure 4.4, both resistant and sensitive cell lines, with an exception for two cell lines with a bit lower response than predicted.

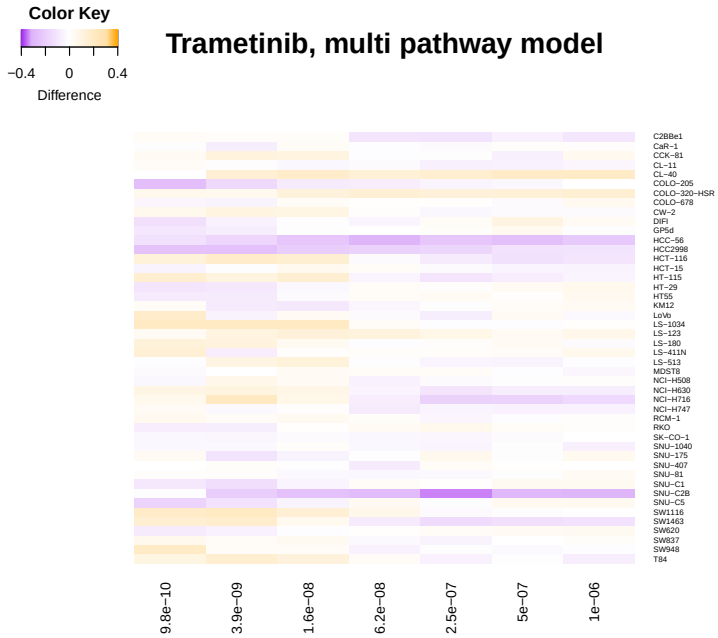


Figure 4.4: Overview heat plot of absolute differences between posterior predictive and observed data points, for each cell line in the Trametinib model, over the 7 drug concentrations.

As we can see from the marginal density distribution of the Trametinib model, see Supplementary Figure 5.12, the mutations in the added pathways have an large influence on the proliferation, which is the explanation for the better model fit. Most notable are the SMAD2 mutations in the TGFB pathway, and the RNF43 mutation in the WNT pathway, and the loss of IRS2. Furthermore, the features that were found in the MAPK/PI3K pathway in chapter 3 for the Trametinib model do now also influence the proliferation signal but with a weaker strength, such as the PTEN mutations and losses, KRAS, EGFR mutations and FGFR1 amplifications.

These associations can be found back in literature, as RNF43 mutations occur in roughly 18% of CRCS, are mutually exclusive with APC mutations[91], and leads to the activation of the WNT pathway, a thus create an increase in proliferation. Mutations in the SMAD2 tumor suppressor also cause a gain in proliferation[92].

Subtype sensitivity

To gain insight into the sensitivity of the separate subtypes to the targeted drugs we compared the estimated activations of the proliferation node between the cell line subtype sets. In Figure 4.5 we depicted these comparisons for the Trametinib model between the subtypes of the three classification methods, the micro satellite instability, the consensus molecular subtypes, and the Colorectal Cancer Subtypes. The differences for the MK2206 and SCH772984 model are depicted in Figures 4.6 and 4.7, however on must keep in mind that these models where not well converged and thus confidence is low.

For all three drugs the MSI subtype has a higher proliferation, thus is more resistant, than the MSS cell line, but it has also a larger variation. As expected is the proliferation for both the MSI and MSS lower for all treatments compared to the untreated model, where the MSI had a slightly lower proliferation than MSS.

In the consensus clustering, the CMS1 cell lines are the most resistant to the three targeted drugs, specially in the Trametinib and SCH772984 model. Which is caused by the enrichment of MSI in the CMS1 cell lines[47]. The other three subtypes behave roughly the same as the variation between the subtypes is small with a few outliers.

The Colorectal Cancer Subtypes are significantly($p < 0.01$) different in proliferation under Trametinib and SCH772984, where the CCS2 subtype is more resistant compared to CCS1 and CCS3. There is no visible difference between CCS1 and CCS3. The CCS2 subtype consists of almost only MSI cell lines and is able to capture the effect of the MSI better that MSI classification

itself. By creating the CCS3 subtype, that is distinct from the two major entities CIN and MSI, de Sousa et. al.[87] has successfully defined a subtype of MSI cell lines that we show here are evidently resistant to these targeted inhibitors.

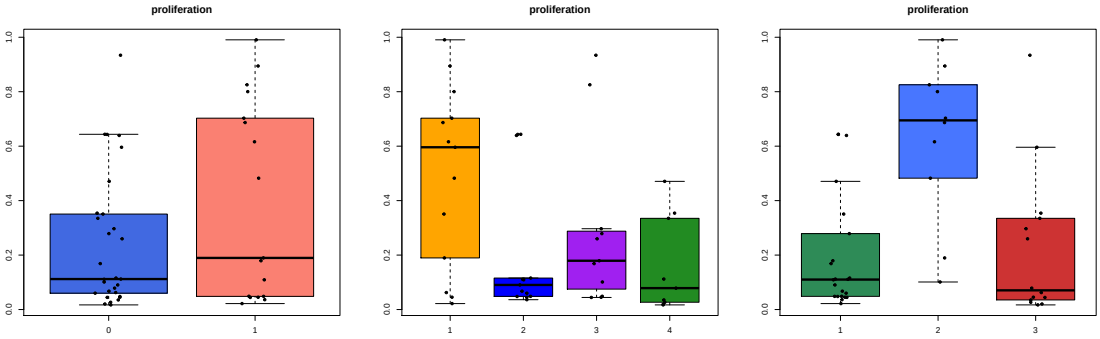


Figure 4.5: Model estimates of the proliferation in the Trametib model, clustered on the three CRC classification approaches, from left to right, MSI, CMS, and CCS.

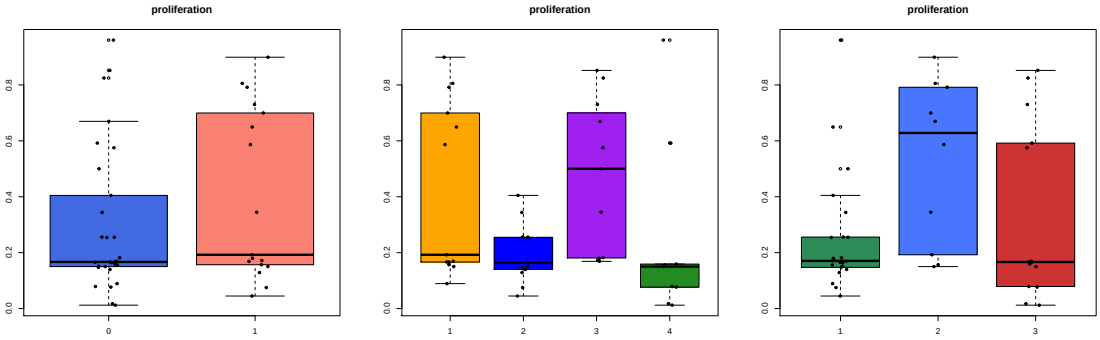


Figure 4.6: Model estimates of the proliferation in the MK2206 model, clustered on the three CRC classification approaches, from left to right, MSI, CMS, and CCS.

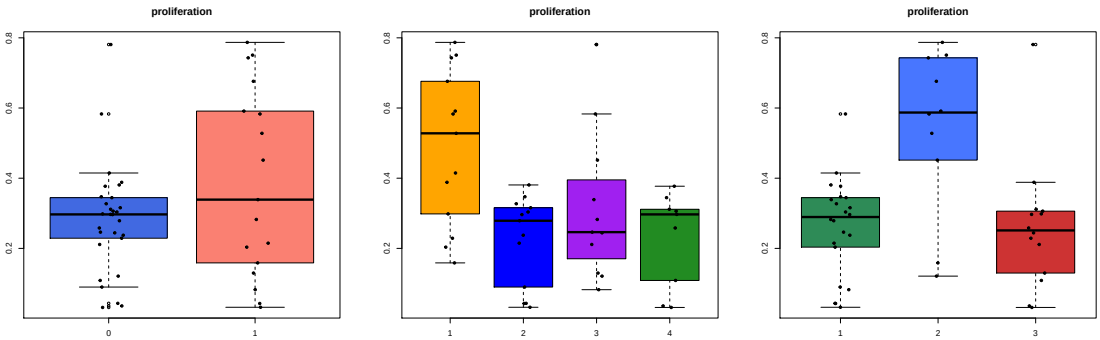


Figure 4.7: Model estimates of the proliferation in the SCH772984 model, clustered on the three CRC classification approaches, from left to right, MSI, CMS, and CCS.

Chapter 5

Discussion

In this thesis we have described our work on the computational modeling of drug response in colorectal cancer cell lines using a signaling pathway model approach and Bayesian inference parameter sampling, which shows that by combining multi type data and computational modeling, a detailed understanding of variability in targeted drug sensitivity can be obtained.

Furthermore we developed a method for reconstructing and replacing the missing activation data, measured by protein mass spectrometry of the signaling pathway nodes and we assessed the differences in estimated pathway activation and proliferation between the commonly used colorectal cancer classification subtypes.

In chapter 2 we analyzed the performance of our activation data replacement method on the MAPK and PI3K pathway and Multi pathway model, by comparing the different combinations of phosphorylation and transcription targets to reconstruct the activation signal of the parent node.

We were able to reconstruct this signal for a selection of signaling nodes. In the MAPK/PI3K pathway we found that for the nodes AKT, S6K, MEK and ERK the activation data from the selected best target or combined selection of targets could be explained by the model. We observed that for some nodes only the first best target or a selection of the first best targets was the most predictive data(S6k, MEK, ERK, and AKT). However, for other nodes the model could better explain the activation signal when the number of selected targets increased, which is the case for P53. Furthermore the effect of abundance expression is important to be able to predict the data well. For example, the MAPK3 T202 and Y204 phosphorylation expression cannot be explained when used as activation data for ERK itself, with the protein abundance of ERK as expression variable, but it can be explained by the model when it is used as data for MEK, which has MEK protein abundance as expression data. Whether the effect of the expression is an influencing factor also depends on the model itself, as the multi pathway model can explain the best target data of AKT better than the MAPK/PI3K model. The effect of the protein abundance has been known to be an important factor for tumor proliferation[93].

The method knows limitations, most important is that the node must have a set of targets, phosphorylated proteins or transcription genes, to select from. To create a set of targets one must rely on the current knowledge and prediction algorithms, and manual curation is often needed, making it a laborious job. Moreover, the number of targets that is needed is hard to predict, as some nodes perform better with more and others with less, therefore it is required to test multiple combinations to find the optimum target set for a specific node.

In chapter 3 we used the MAPK/PI3K pathway model to explain the drug response data of five different targeted inhibitors, using both the true activation protein mass spectrometry data of the nodes and the best replacement activation data that was found by experiments in chapter 2. The model was able to converge properly for the drugs Trametinib, MK2206, and SCH772984, the other two, Afatinib and the combination of Afatinib and Trametinib, were not able to completely converge in reasonable time with the given sample size and sampler settings. The models could explain the drug dose response for a large number of cell lines and also showed which cell line behavior could not be explained and to what extent. From the marginal density distributions of the parameters the influence of all genetic features on the proliferation signal can be analyzed. We compared the found influence of these genetic aberrations to the features found by a elastic net

regression on the same set of features and the drug IC50s. The elastic net finds similar results but less than our model and it does not give precise information on the confidence of the coefficients.

We were able to verify the found resistance and sensitivity mechanisms of the model against the literature. But it is hard to say whether the model was able to find novel associations and mechanisms that were not, to our knowledge, described in literature. The found associations with the largest strengths are already characterized in the literature, and weaker associations must be handled with caution before drawing large conclusions, as the number of cell lines is still limited.

There are some limitation to this modeling approach. First it must be noted that these models are a simplified representation of the real biological structures and mechanisms in the cancer cells. The most important simplifications are the absence of time and not including feedback signaling. Whether the model based on data from in vitro experiments on cancer cell lines with only short term drug response is reliable enough to say something about the effect of the drug in cancer patients can be debated, but it does give functional insight into the signaling mechanisms inside the cell lines. The effect of a later relapse is not taken into account, adding a dimension of time and feedback mechanisms should give more insight into this, however the computational costs would become significantly higher. As the number of parameters and the model complexity increases sampling the high dimensional solution space becomes the biggest challenge.

Sampling the solution space is already a demanding job using the current model. It was found by Thijssen et. al.[38] that the PTMCMC sampling is a efficient way of searching the parameter space in these pathway drug response models, especially using the adaptive temperature, adaptive proposals, and parameter blocking. However, the models still did not always converge, and parameters were traps in local modes, that hindered efficient sampling to gain enough confidence on the true parameter value. Furthermore, because estimating one model took already longer than 20 hours, cross validation of the model would take weeks and was therefore not attempted. Therefore, Bayesian inference using parameter sampling on these models still remains a large problem that has to be solved with smarter sampling algorithms and more computational power, as reaching convergence for some of the larger models would require 10 or 20 times more samples. Another solution would be using more informative priors for the parameter distributions, however one of the bottlenecks in the development of predictive mathematical models for signaling networks is the lack of experimental data and literature that describes these interaction strengths[29].

Another important limitation for inferring treatment sensitivity using measurements from cultured cell lines and and in vitro experiments, is that these data lack predictive power with respect to clinical trials[94]. The underlying causes are not precisely known, but are most likely due to alterations in biologic properties, such as changes in genetic information, growth rate, invasion potential, and loss of specific cell populations[95][94]. However, the study of cell line cultures is still essential for the anti-cancer drug development[95].

Using the estimated node activation proliferation we could distinct in chapter 4 the different subtypes of colorectal cancer in the Multi pathway model. We looked at a steady state model and three drug models. By using a larger multi pathway model with more mutations and copy number variations than the MAPK/PI3k pathway model the drug dose response data could be better explained, almost all cell lines were well predicted for the Trametinib model. We found that the micro satellite instable cell lines had on average a lower steady state proliferation, activated AKT and P53 pathway and an inactivated WNT pathway. The difference between the four consensus molecular subtypes was smaller, but mechanisms described in literature could be verified by the model results. Likewise, the Colorectal Cancer Subtypes by de Sousa et al.[87] showed results similar to the MSI/MSS subtypes, which is logical given that the subtypes are also based on the MSI status. Due to the low number of cell lines the differences are not always significant, the conclusions that can be made from these experiments are therefore not far reaching and must be considered as trends. Incorporating more cell lines would make it possible to better distinguish different subtypes or develop a new subtype clustering system based on the estimated pathway activities. These molecular classifications could help the patient care by aiding the selection of the most appropriate treatment[90].

Regarding other future work, the main problem of the signaling pathway modeling approach is the estimation of the model parameters. We used PTMCM sampling to sample from the posterior predictive, which does a good job for small models but when the model become more complex it gets very difficult and takes a very long computation time to reach sufficient convergence. Therefore

research into different parameter sampling techniques that can sample this complex high dimensional spaces efficiently is much needed. Another sampling method that could be looked into is for example sequential Monte-Carlo sampling, which also has been improved recently[96].

We found that the abundance expression of the signaling node has a large influence on how well the model can explain the variance of the activation data. It is to us not completely clear why in some models the influence of the abundance expression is larger than others. It would therefore be interesting to study the source of these differences and the effect of protein abundance of the model fit.

Furthermore, nowadays targeted treatments are increasingly often given in combinations of multiple drugs to inhibit multiple pathway at once[97]. Being able to model these different combination of targeted drugs is thus of much interest. We were able to model the combination of Trametinib and Afatinib using our current model, showing that is this modeling technique is capable of explaining multiple drugs at once. Future research should therefore also focus on uncovering the mechanisms underlying different combination treatments.

Finally, we did not have access to growth data of the untreated cell lines, forcing us to estimate the untreated proliferation of the cell lines without experimental measurements. Adding this data in the future to the model would make the estimates more accurate and meaningful, making it easier to compare the untreated cell lines and cell lines under treatment of the target inhibitors.

We can conclude that using multiple measured data types of a large colorectal cancer cell line panel integrated in a mathematical signaling pathway model, based on current knowledge, can give insight into the biological mechanisms underlying targeted drug sensitivity, that can verify and possibly supplement findings in literature. Furthermore, it is possible to reconstruct missing node activation data based on its phosphorylation or transcription targets and use the model estimated pathway activation to roughly distinct different colorectal cancer subtypes and their response to targeted therapy. We believe that understanding the behavior of signaling pathways in cancer cells under targeted treatment will be a valuable step in the way to precision and personalized medicine.

5.1 Acknowledgements

I would first like to thank Bram Thijssen for introducing me to this project and for the useful comments, remarks and his engagement in my master thesis. I thank the Netherlands Cancer Institute, for offering me a place to work and for letting me use their computational servers and other facilities, and the Welcome Trust Sanger Institute, for providing the protein mass spectrometry and drug dose response data. Furthermore would like to I thank Erwin Bakker for his supervision at Universiteit Leiden. Also I thank Daniel Vis for his help with the Sanger datasets and Gergana Bounova for providing the colorectal cancer subtype data.

Bibliography

[1] P. Favoriti, G. Carbone, M. Greco, F. Pirozzi, R. E. M. Pirozzi, and F. Corcione, "Worldwide burden of colorectal cancer: a review," *Updates in surgery*, vol. 68, no. 1, pp. 7–11, 2016.

[2] R. L. Siegel, K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. Meester, A. Barzi, and A. Jemal, "Colorectal cancer statistics, 2017," *CA: a cancer journal for clinicians*, vol. 67, no. 3, pp. 177–193, 2017.

[3] K. D. Miller, R. L. Siegel, C. C. Lin, A. B. Mariotto, J. L. Kramer, J. H. Rowland, K. D. Stein, R. Alteri, and A. Jemal, "Cancer treatment and survivorship statistics, 2016," *CA: a cancer journal for clinicians*, vol. 66, no. 4, pp. 271–289, 2016.

[4] Q. Mo, S. Wang, V. E. Seshan, A. B. Olshen, N. Schultz, C. Sander, R. S. Powers, M. Ladanyi, and R. Shen, "Pattern discovery and cancer gene identification in integrated cancer genomic data," *Proceedings of the National Academy of Sciences*, vol. 110, no. 11, pp. 4245–4250, 2013.

[5] M. Niepel, M. Hafner, E. A. Pace, M. Chung, D. H. Chai, L. Zhou, B. Schoeberl, and P. K. Sorger, "Profiles of basal and stimulated receptor signaling networks predict drug response in breast cancer lines," *Science signaling*, vol. 6, no. 294, 2013.

[6] M. Huang, A. Shen, J. Ding, and M. Geng, "Molecularly targeted cancer therapy: some lessons from the past decade," *Trends in pharmacological sciences*, vol. 35, no. 1, pp. 41–50, 2014.

[7] F. Eduati, V. Doldàn-Martelli, B. Klinger, T. Cokelaer, A. Sieber, F. Kogera, M. Dorel, M. J. Garnett, N. Blüthgen, and J. Saez-Rodriguez, "Drug resistance mechanisms in colorectal cancer dissected with cell type–specific dynamic logic models," *Cancer research*, 2017.

[8] A. S. Iyer, H. U. Osmanbeyoglu, and C. S. Leslie, "Computational methods to dissect gene regulatory networks in cancer," *Current Opinion in Systems Biology*, 2017.

[9] R. J. Sullivan and K. T. Flaherty, "Resistance to braf-targeted therapy in melanoma," *European journal of cancer*, vol. 49, no. 6, pp. 1297–1304, 2013.

[10] A. Marusyk and K. Polyak, "Tumor heterogeneity: causes and consequences," *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, vol. 1805, no. 1, pp. 105–117, 2010.

[11] A. Sartore-Bianchi, M. Martini, F. Molinari, S. Veronese, M. Nichelatti, S. Artale, F. Di Nicolantonio, P. Saletti, S. De Dosso, L. Mazzucchelli, *et al.*, "Pik3ca mutations in colorectal cancer are associated with clinical resistance to egfr-targeted monoclonal antibodies," *Cancer research*, vol. 69, no. 5, pp. 1851–1857, 2009.

[12] N. McGranahan and C. Swanton, "Clonal heterogeneity and tumor evolution: past, present, and the future," *Cell*, vol. 168, no. 4, pp. 613–628, 2017.

[13] D. Fey, M. Halasz, D. Dreidax, S. P. Kennedy, N. Rauch, A. Garcia Munoz, R. Pilkington, M. Fischer, W. Kolch, B. N. Kholodenko, *et al.*, "Signaling pathway models as biomarkers: Patient-specific simulations of jnk activity predict the survival of neuroblastoma patients," 2015.

[14] D. Silverbush, S. Grosskurth, D. Wang, F. Powell, B. Gottgens, J. Dry, and J. Fisher, "Cell-specific computational modeling of the pim pathway in acute myeloid leukemia," *Cancer research*, 2016.

[15] R. Chuang, B. A. Hall, D. Benque, B. Cook, S. Ishtiaq, N. Piterman, A. Taylor, M. Vardi, S. Koschmieder, B. Gottgens, *et al.*, "Drug target optimization in chronic myeloid leukemia using innovative computational platform," *Scientific reports*, vol. 5, 2015.

[16] G. Riddick, H. Song, S. Ahn, J. Walling, D. Borges-Rivera, W. Zhang, and H. A. Fine, "Predicting in vitro drug sensitivity using random forests," *Bioinformatics*, vol. 27, no. 2, pp. 220–224, 2010.

[17] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares, *et al.*, "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, no. 7391, p. 570, 2012.

[18] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, *et al.*, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.

[19] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, *et al.*, "A landscape of pharmacogenomic interactions in cancer," *Cell*, vol. 166, no. 3, pp. 740–754, 2016.

[20] A. Daemen, O. L. Griffith, L. M. Heiser, N. J. Wang, O. M. Enache, Z. Sanborn, F. Pepin, S. Durinck, J. E. Korkola, M. Griffith, *et al.*, "Modeling precision treatment of breast cancer," *Genome biology*, vol. 14, no. 10, p. R110, 2013.

[21] Z. Dong, N. Zhang, C. Li, H. Wang, Y. Fang, J. Wang, and X. Zheng, "Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection," *BMC cancer*, vol. 15, no. 1, p. 489, 2015.

[22] J. K. Lee, D. M. Havaleshko, H. Cho, J. N. Weinstein, E. P. Kaldjian, J. Karpovich, A. Grimshaw, and D. Theodorescu, "A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery," *Proceedings of the National Academy of Sciences*, vol. 104, no. 32, pp. 13086–13091, 2007.

[23] J. E. Staunton, D. K. Slonim, H. A. Coller, P. Tamayo, M. J. Angelo, J. Park, U. Scherf, J. K. Lee, W. O. Reinhold, J. N. Weinstein, *et al.*, "Chemosensitivity prediction by transcriptional profiling," *Proceedings of the National Academy of Sciences*, vol. 98, no. 19, pp. 10787–10792, 2001.

[24] B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, N. E. Bodycombe, C. K. Soule, J. Gould, *et al.*, "Harnessing connectivity in a large-scale small-molecule sensitivity dataset," *Cancer discovery*, vol. 5, no. 11, pp. 1210–1223, 2015.

[25] K. Vougas, M. Krochmal, T. Jackson, A. Polyzos, A. Aggelopoulos, I. S. Pateras, M. Liontos, A. Varvarigou, E. O. Johnson, V. Georgoulias, *et al.*, "Deep learning and association rule mining for predicting drug response in cancer. a personalised medicine approach.," *bioRxiv*, p. 070490, 2017.

[26] T. S. Christensen, A. P. Oliveira, and J. Nielsen, "Reconstruction and logical modeling of glucose repression signaling pathways in saccharomyces cerevisiae," *BMC systems biology*, vol. 3, no. 1, p. 7, 2009.

[27] A. Saadatpour and R. Albert, "A comparative study of qualitative and quantitative dynamic models of biological regulatory networks," *EPJ Nonlinear Biomedical Physics*, vol. 4, no. 1, pp. 1–13, 2016.

[28] G. Bidkhori, A. Moeini, and A. Masoudi-Nejad, "Modeling of tumor progression in nsclc and intrinsic resistance to tki in loss of pten expression," *PloS one*, vol. 7, no. 10, p. e48004, 2012.

[29] N. Sulaimanov, M. Klose, H. Busch, and M. Boerries, "Understanding the mtor signaling pathway via mathematical modeling," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2017.

[30] B. Klinger, A. Sieber, R. Fritsche-Guenther, F. Witzel, L. Berry, D. Schumacher, Y. Yan, P. Durek, M. Merchant, R. Schäfer, *et al.*, "Network quantification of egfr signaling unveils potential for targeted combination therapy," *Molecular systems biology*, vol. 9, no. 1, p. 673, 2013.

[31] L. Calzone, L. Tournier, S. Fourquet, D. Thieffry, B. Zhivotovsky, E. Barillot, and A. Zinovyev, "Mathematical modelling of cell-fate decision in response to death receptor engagement," *PLoS computational biology*, vol. 6, no. 3, p. e1000702, 2010.

[32] P. Zhu, H. M. Aliabadi, H. Uludağ, and J. Han, "Identification of potential drug targets in cancer signaling pathways using stochastic logical models," *Scientific reports*, vol. 6, p. 23078, 2016.

[33] N. Bonzanni, A. Garg, K. A. Feenstra, J. Schütte, S. Kinston, D. Miranda-Saavedra, J. Heringa, I. Xenarios, and B. Göttgens, "Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model," *Bioinformatics*, vol. 29, no. 13, pp. i80–i88, 2013.

[34] N. Le Novere, "Quantitative and logic modelling of gene and molecular networks," *Nature Reviews. Genetics*, vol. 16, no. 3, p. 146, 2015.

[35] A. Ryll, J. Bucher, A. Bonin, S. Bongard, E. Gonçalves, J. Saez-Rodriguez, J. Niklas, and S. Klamt, "A model integration approach linking signalling and gene-regulatory logic with kinetic metabolic models," *Biosystems*, vol. 124, pp. 26–38, 2014.

[36] D. C. Kirouac, J. Y. Du, J. Lahdenranta, R. Overland, D. Yarar, V. Paragas, E. Pace, C. F. McDonagh, U. B. Nielsen, and M. D. Onsum, "Computational modeling of erbb2-amplified breast cancer identifies combined erbb2/3 blockade as superior to the combination of mek and akt inhibitors," *Sci Signal*, vol. 6, no. 288, p. ra68, 2013.

[37] N. Lartillot, "Parameter estimation: optimizing versus conditioning," aug 2014.

[38] B. Thijssen, T. M. Dijkstra, T. Heskes, and L. F. Wessels, "Bcm: toolkit for bayesian analysis of computational models using samplers," *BMC systems biology*, vol. 10, no. 1, p. 100, 2016.

[39] S. Boellner and K.-F. Becker, "Reverse phase protein arrays—quantitative assessment of multiple biomarkers in biopsies for clinical use," *Microarrays*, vol. 4, no. 2, pp. 98–114, 2015.

[40] C. J. Creighton and S. Huang, "Reverse phase protein arrays in signaling pathways: a data integration perspective," *Drug design, development and therapy*, vol. 9, p. 3519, 2015.

[41] F. G. Strathmann and A. N. Hoofnagle, "Current and future applications of mass spectrometry to the clinical laboratory," *American journal of clinical pathology*, vol. 136, no. 4, pp. 609–616, 2011.

[42] M. A. Gillette and S. A. Carr, "Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry," *Nature methods*, vol. 10, no. 1, pp. 28–34, 2013.

[43] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 4036–4048, 2006.

[44] S. F. Gull and G. J. Daniell, "Image reconstruction from incomplete and noisy data," *Nature*, vol. 272, no. 5655, pp. 686–690, 1978.

[45] H. Xie, K. T. McDonnell, and H. Qin, "Surface reconstruction of noisy and defective data sets," in *Proceedings of the conference on Visualization'04*, pp. 259–266, IEEE Computer Society, 2004.

[46] T. K. Dey and S. Goswami, "Provable surface reconstruction from noisy samples," in *Proceedings of the twentieth annual symposium on Computational geometry*, pp. 330–339, ACM, 2004.

[47] R. Dienstmann, L. Vermeulen, J. Guinney, S. Kopetz, S. Tejpar, and J. Tabernero, "Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer," *Nature Reviews Cancer*, vol. 17, no. 2, pp. 79–92, 2017.

[48] J. Guinney, R. Dienstmann, X. Wang, A. De Reyniès, A. Schlicker, C. Soneson, L. Marisa, P. Roepman, G. Nyamundanda, P. Angelino, *et al.*, "The consensus molecular subtypes of colorectal cancer," *Nature medicine*, vol. 21, no. 11, p. 1350, 2015.

[49] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, *et al.*, "Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells," *Nucleic acids research*, vol. 41, no. D1, pp. D955–D961, 2012.

[50] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to mcmc for machine learning," *Machine learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[51] H. G. Katzgraber, S. Trebst, D. A. Huse, and M. Troyer, "Feedback-optimized parallel tempering monte carlo," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2006, no. 03, p. P03018, 2006.

[52] D. Turek, P. de Valpine, C. J. Paciorek, C. Anderson-Bergman, *et al.*, "Automated parameter blocking for efficient markov chain monte carlo sampling," *Bayesian Analysis*, vol. 12, no. 2, pp. 465–490, 2017.

[53] H. Haario, E. Saksman, J. Tamminen, *et al.*, "An adaptive metropolis algorithm," *Bernoulli*, vol. 7, no. 2, pp. 223–242, 2001.

[54] S. D. Markowitz and M. M. Bertagnolli, "Molecular basis of colorectal cancer," *New England Journal of Medicine*, vol. 361, no. 25, pp. 2449–2460, 2009.

[55] A. S. Sameer, "Colorectal cancer: molecular mutations and polymorphisms," *Frontiers in oncology*, vol. 3, 2013.

[56] E. Castellano and J. Downward, "Ras interaction with pi3k: more than just another effector pathway," *Genes & cancer*, vol. 2, no. 3, pp. 261–274, 2011.

[57] Y. Shi, "Structural insights on smad function in tgfβ signaling," *Bioessays*, vol. 23, no. 3, pp. 223–232, 2001.

[58] M. Giannakis, E. Hodis, X. J. Mu, M. Yamauchi, J. Rosenbluh, K. Cibulskis, G. Saksena, M. S. Lawrence, Z. R. Qian, R. Nishihara, *et al.*, "Rnf43 is frequently mutated in colorectal and endometrial cancers," *Nature genetics*, vol. 46, no. 12, pp. 1264–1266, 2014.

[59] J. A. McCubrey, L. S. Steelman, F. E. Bertrand, N. M. Davis, M. Sokolosky, S. L. Abrams, G. Montalto, A. B. D'Assoro, M. Libra, F. Nicoletti, *et al.*, "Gsk-3 as potential target for therapeutic intervention in cancer," *Oncotarget*, vol. 5, no. 10, p. 2881, 2014.

[60] P. Liu, M. Begley, W. Michowski, H. Inuzuka, M. Ginzberg, D. Gao, P. Tsou, W. Gan, A. Papa, B. M. Kim, *et al.*, "Cell-cycle-regulated activation of akt kinase by phosphorylation at its carboxyl terminus," *Nature*, vol. 508, no. 7497, p. 541, 2014.

[61] G. Hatzivassiliou, J. R. Haling, H. Chen, K. Song, S. Price, R. Heald, J. F. Hewitt, M. Zak, A. Peck, C. Orr, *et al.*, "Mechanism of mek inhibition determines efficacy in mutant kras-versus braf-driven cancers," *Nature*, vol. 501, no. 7466, p. 232, 2013.

[62] D. J. Templeton, "Protein kinases: getting neked for s6k activation," *Current Biology*, vol. 11, no. 15, pp. R596–R599, 2001.

[63] R. Linding, L. J. Jensen, G. J. Ostheimer, M. A. van Vugt, C. Jørgensen, I. M. Miron, F. Diella, K. Colwill, L. Taylor, K. Elder, *et al.*, "Systematic discovery of in vivo phosphorylation networks," *Cell*, vol. 129, no. 7, pp. 1415–1426, 2007.

[64] H. Horn, E. M. Schoof, J. Kim, X. Robin, M. L. Miller, F. Diella, A. Palma, G. Cesareni, L. J. Jensen, and R. Linding, "Kinomexplorer: an integrated platform for kinome biology studies," *Nature methods*, vol. 11, no. 6, pp. 603–604, 2014.

[65] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, *et al.*, "Transfac® and its module transcompel®: transcriptional gene regulation in eukaryotes," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D108–D110, 2006.

[66] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[67] F. Commo and B. M. Bot, "R package nplr n-parameter logistic regressions," 2016.

[68] T. Hastie and J. Qian, "Glmnet vignette," 2014.

[69] M. Bernhardt, E. Orouji, L. Larribere, C. Gebhardt, and J. Utikal, "Efficacy of vemurafenib in a trametinib-resistant stage iv melanoma patient," *Clinical Cancer Research*, vol. 20, no. 9, pp. 2498–2499, 2014.

[70] A. Prahallad, C. Sun, S. Huang, F. Di Nicolantonio, R. Salazar, D. Zecchin, R. L. Beijersbergen, A. Bardelli, and R. Bernards, "Unresponsiveness of colon cancer to braf (v600e) inhibition through feedback activation of egfr," *Nature*, vol. 483, no. 7387, p. 100, 2012.

[71] G. V. Long, C. Fung, A. M. Menzies, G. M. Pupo, M. S. Carlino, J. Hyman, H. Shahheydari, V. Tembe, J. F. Thompson, R. P. Saw, *et al.*, "Increased mapk reactivation in early resistance to dabrafenib/trametinib combination therapy of braf-mutant metastatic melanoma," *Nature communications*, vol. 5, p. 5694, 2014.

[72] S. Wee, Z. Jagani, K. X. Xiang, A. Loo, M. Dorsch, Y.-M. Yao, W. R. Sellers, C. Lengauer, and F. Stegmeier, "Pi3k pathway activation mediates resistance to mek inhibitors in kras mutant cancers," *Cancer research*, vol. 69, no. 10, pp. 4286–4293, 2009.

[73] J. A. McCubrey, L. S. Steelman, W. H. Chappell, S. L. Abrams, R. A. Franklin, G. Montalto, M. Cervello, M. Libra, S. Candido, G. Malaponte, *et al.*, "Ras/raf/mek/erk and pi3k/pten/akt/mtor cascade inhibitors: how mutations can result in therapy resistance and how to overcome resistance," *Oncotarget*, vol. 3, no. 10, p. 1068, 2012.

[74] T. Sangai, A. Akcakanat, H. Chen, E. Tarco, Y. Wu, K.-A. Do, T. W. Miller, C. L. Arteaga, G. B. Mills, A. M. Gonzalez-Angulo, *et al.*, "Biomarkers of response to akt inhibitor mk-2206 in breast cancer," *Clinical Cancer Research*, vol. 18, no. 20, pp. 5816–5828, 2012.

[75] H. Hirai, H. Sootome, Y. Nakatsuru, K. Miyama, S. Taguchi, K. Tsujioka, Y. Ueno, H. Hatch, P. K. Majumder, B.-S. Pan, *et al.*, "Mk-2206, an allosteric akt inhibitor, enhances antitumor efficacy by standard chemotherapeutic agents or molecular targeted drugs in vitro and in vivo," *Molecular cancer therapeutics*, vol. 9, no. 7, pp. 1956–1967, 2010.

[76] C. F. Malone, J. A. Fromm, O. Maertens, T. DeRaedt, R. Ingraham, and K. Cichowski, "Defining key signaling nodes and therapeutic biomarkers in nf1-mutant cancers," *Cancer discovery*, vol. 4, no. 9, pp. 1062–1073, 2014.

[77] C. Sun, S. Hobor, A. Bertotti, D. Zecchin, S. Huang, F. Galimi, F. Cottino, A. Prahallad, W. Grernrum, A. Tzani, *et al.*, "Intrinsic resistance to mek inhibition in kras mutant lung and colon cancer through transcriptional induction of erbb3," *Cell reports*, vol. 7, no. 1, pp. 86–93, 2014.

[78] M. S. Carlino, J. R. Todd, K. Gowrishankar, B. Mijatov, G. M. Pupo, C. Fung, S. Snoyman, P. Hersey, G. V. Long, R. F. Kefford, *et al.*, "Differential activity of mek and erk inhibitors in braf inhibitor resistant melanoma," *Molecular oncology*, vol. 8, no. 3, pp. 544–554, 2014.

[79] T. K. Hayes, N. F. Neel, C. Hu, P. Gautam, M. Chenard, B. Long, M. Aziz, M. Kassner, K. L. Bryant, M. Pierobon, *et al.*, "Long-term erk inhibition in kras-mutant pancreatic cancer is associated with myc degradation and senescence-like growth suppression," *Cancer cell*, vol. 29, no. 1, pp. 75–89, 2016.

[80] I. De Pauw, A. Wouters, J. Van den Bossche, M. Peeters, P. Pauwels, V. Deschoolmeester, J. Vermorken, and F. Lardon, "Preclinical and clinical studies on afatinib in monotherapy and in combination regimens: Potential impact in colorectal cancer," *Pharmacology & therapeutics*, vol. 166, pp. 71–83, 2016.

[81] S. M. Leto, F. Sassi, I. Catalano, V. Torri, G. Migliardi, E. R. Zanella, M. Throsby, A. Bertotti, and L. Trusolino, "Sustained inhibition of her3 and egfr is necessary to induce regression of her2-amplified gastrointestinal carcinomas," *Clinical Cancer Research*, vol. 21, no. 24, pp. 5519–5531, 2015.

[82] S.-S. Guan, J. Chang, C.-C. Cheng, T.-Y. Luo, A.-S. Ho, C.-C. Wang, C.-T. Wu, and S.-H. Liu, "Afatinib and its encapsulated polymeric micelles inhibits her2-overexpressed colorectal tumor cell growth in vitro and in vivo," *Oncotarget*, vol. 5, no. 13, p. 4868, 2014.

[83] V. A. Miller, V. Hirsh, J. Cadranel, Y.-M. Chen, K. Park, S.-W. Kim, C. Zhou, W.-C. Su, M. Wang, Y. Sun, *et al.*, "Afatinib versus placebo for patients with advanced, metastatic non-small-cell lung cancer after failure of erlotinib, gefitinib, or both, and one or two lines of chemotherapy (lux-lung 1): a phase 2b/3 randomised trial," *The lancet oncology*, vol. 13, no. 5, pp. 528–538, 2012.

[84] L. V. Sequist, J. C.-H. Yang, N. Yamamoto, K. O'Byrne, V. Hirsh, T. Mok, S. L. Geater, S. Orlov, C.-M. Tsai, M. Boyer, *et al.*, "Phase iii study of afatinib or cisplatin plus pemetrexed in patients with metastatic lung adenocarcinoma with egfr mutations," *Journal of clinical oncology*, vol. 31, no. 27, pp. 3327–3334, 2013.

[85] J.-P. H. Machiels, R. I. Haddad, J. Fayette, L. F. Licitra, M. Tahara, J. B. Vermorken, P. M. Clement, T. Gauler, D. Cupissol, J. J. Grau, *et al.*, "Afatinib versus methotrexate as second-line treatment in patients with recurrent or metastatic squamous-cell carcinoma of the head and neck progressing on or after platinum-based therapy (lux-head & neck 1): an open-label, randomised phase 3 trial," *The Lancet Oncology*, vol. 16, no. 5, pp. 583–594, 2015.

[86] W. De Roock, B. Claes, D. Bernasconi, J. De Schutter, B. Biesmans, G. Fountzilas, K. T. Kalogeras, V. Kotoula, D. Papamichael, P. Laurent-Puig, *et al.*, "Effects of kras, braf, nras, and pik3ca mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis," *The lancet oncology*, vol. 11, no. 8, pp. 753–762, 2010.

[87] E. M. Felipe De Sousa, X. Wang, M. Jansen, E. Fessler, A. Trinh, L. P. De Rooij, J. H. De Jong, O. J. De Boer, R. Van Leersum, M. F. Bijlsma, *et al.*, "Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions," *Nature medicine*, vol. 19, no. 5, pp. 614–618, 2013.

[88] D. L. Worthley and B. A. Leggett, "Colorectal cancer: molecular features and clinical opportunities," *The Clinical Biochemist Reviews*, vol. 31, no. 2, p. 31, 2010.

[89] D. J. Weisenberger, K. D. Siegmund, M. Campan, J. Young, T. I. Long, M. A. Faasse, G. H. Kang, M. Widschwendter, D. Weener, D. Buchanan, *et al.*, "Cpg island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with braf mutation in colorectal cancer," *Nature genetics*, vol. 38, no. 7, p. 787, 2006.

[90] N. Rodriguez-Salas, G. Dominguez, R. Barderas, M. Mendiola, X. García-Albéniz, J. Maurel, and J. F. Batlle, "Clinical relevance of colorectal cancer molecular subtypes," *Critical reviews in oncology/hematology*, vol. 109, pp. 9–19, 2017.

[91] R. B. Corcoran, C. E. Atreya, G. S. Falchook, E. L. Kwak, D. P. Ryan, J. C. Bendell, O. Hamid, W. A. Messersmith, A. Daud, R. Kurzrock, *et al.*, "Combined braf and mek inhibition with dabrafenib and trametinib in braf v600–mutant colorectal cancer," *Journal of clinical oncology*, vol. 33, no. 34, pp. 4023–4031, 2015.

[92] J. Xu and L. Attisano, "Mutations in the tumor suppressors smad2 and smad4 inactivate transforming growth factor β signaling by targeting smads to the ubiquitin–proteasome pathway," *Proceedings of the National Academy of Sciences*, vol. 97, no. 9, pp. 4820–4825, 2000.

[93] D. Silvera, S. C. Formenti, and R. J. Schneider, "Translational control in cancer," *Nature reviews. Cancer*, vol. 10, no. 4, p. 254, 2010.

[94] M. Hidalgo, F. Amant, A. V. Biankin, E. Budinská, A. T. Byrne, C. Caldas, R. B. Clarke, S. de Jong, J. Jonkers, G. M. Mælandsmo, *et al.*, "Patient-derived xenograft models: an emerging platform for translational cancer research," *Cancer discovery*, vol. 4, no. 9, pp. 998–1013, 2014.

[95] J.-P. Gillet, A. M. Calcagno, S. Varma, M. Marino, L. J. Green, M. I. Vora, C. Patel, J. N. Orina, T. A. Eliseeva, V. Singal, *et al.*, "Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance," *Proceedings of the National Academy of Sciences*, vol. 108, no. 46, pp. 18708–18713, 2011.

[96] T. L. T. Nguyen, F. S. Septier, G. W. Peters, and Y. Delignon, "Efficient sequential monte-carlo samplers for bayesian inference," 2017.

[97] D. Decaudin and C. Le Tourneau, "Combinations of targeted therapies in human cancers," *Aging (Albany NY)*, vol. 8, no. 10, p. 2258, 2016.

[98] N. Hay and N. Sonenberg, "Upstream and downstream of mtor," *Genes & development*, vol. 18, no. 16, pp. 1926–1945, 2004.

Supplementary Figures and Tables

Chapter 1

Table 5.1: Overview of the parameter used in the model and their prior distributions

Parameters	Description Prior	Prior
Node Activation		
b_i	Base signal	Uniform($a = 0$,$b = 1$)
$s_{k,i}$	Signal strength	Uniform($a = 0$,$b = 5$)
$s_{mut,m,i}$	Mutation signal strength	Uniform($a = 0$,$b = 1$)
p_i	Protein expression coefficient	Uniform($a = 0$,$b = 1$)
Proliferation		
k_k	Signal to proliferation strength	Exponential($\lambda = 25$)
$C_{IC50,n}$	Drug affinity for target (x = in vitro IC50 + 1)	Normal($\mu = x$, $\sigma = 1$) – log10 scale
h_n	Logsteepness of dose-response effects	Uniform($a = x$,$b = 1$) – log10 scale
K_n	Max inhibition by drug	Uniform($a = 0$,$b = 1$)
Data		
σ_i	Variance for PMS data	Exponential($\lambda = 5$)
g_i	Background or base signal for PMS data	Uniform($a = 0$, $b = 1$)
σ_i	Variance for drug response data (cell titer)	Exponential($\lambda = 10$)

Chapter 2

Table 5.2: True data and target data of all pathway nodes used in the missing data reconstruction experiments

Node	MEK	ERK/MAPK	AKT	MTOR	S6K	GSK3B	CTNNB	P53
n1	2	5	6	3	3	4	4	4
n2	2	9	10	5	7	9	9	6
n3	4	13	15	7	13	14	12	15
	True activation data:							
	-	MAPK3_T202_Y204	-	MTOR_S2448[98]	-	GSK3B_S9_T7	Protein Expr	Protein Expr
	Ranked Targets data with decreasing confidence:							
	MAPK3_T202_Y204	MYC_S62_T58_S64	FOXO3_S253	EIF4EBP1_S65_T68_T70	MTOR_S2448	MYC_T58_S64	CCND1	RRM2B
	MAPK1_T185_Y187	ELK1_S324	FOXO1_S319	EIF4EBP1_T70	RPS6_S236_S240	MYC_S62_T58_S64	CCND2	MDM2
	HSF1_S320_S326_S314	RAF1_S301	FOXO1_S256	EIF4EBP1_S35_S44_T36...	RPS6_S242_S244	CTNNB1_S45	CCND3	GDF15
	KRT8_S71_S74_T67	STAT5A_S780	BAD_S99	RPTOR_S859_S863	EIF4B_S422	DPYSL2_S518_T514	BIRC5	BTG2
		BRAF_T401	MTOR_S2448	AMOTL2_S759	IRS1_S1101	PPP1R2_S77_S87_T89_T92	JUN	DDB2
		RPTOR_S859_S863	AKT1S1_T246	DEPTOR_S280(293,299)	IRS1_S270	EIF4EBP2_T37_T41_T45_T46	MYC	GADD45A
		ATF2_T71	GSK3B_S9_T7	MAF1_S60_S65_S68_S70	IRS1_S636_S629	JUN_S243	SOX17	PLK3
		KRT8_S425_S432_S436...	RAF1_S257_S259_T260		CAD_S1859	CTPS1_S571_S573_S574...	PPARD	RPS27L
		SP1_S59_S42	GSK3A_S21		NCBP1_S22_T21	PTK2_S722	CLDN1	TNFRSF10B
		RPS6KA1_S380	MAP2K4_S80		NCBP1_S7	EZH2_S363_S366_T367	EP300	TRIAP1
		MYB_S532	IRS1_S636_S629		NDRG2_S332_S338_T334	FBXO4_S12	PLAU	ZMAT3
		RRAS2_S186	PDCD4_S457		NDRG2_S50	NFKB2_S222	NOS2	BAX
		SPHK2_S387	BRAF_S364_S365		URI1_S372	KAT5_S90_S86		PGF
			PIKFYVE_S307			UNG_T60		POLH
			CHEK1_S280					PPM1D

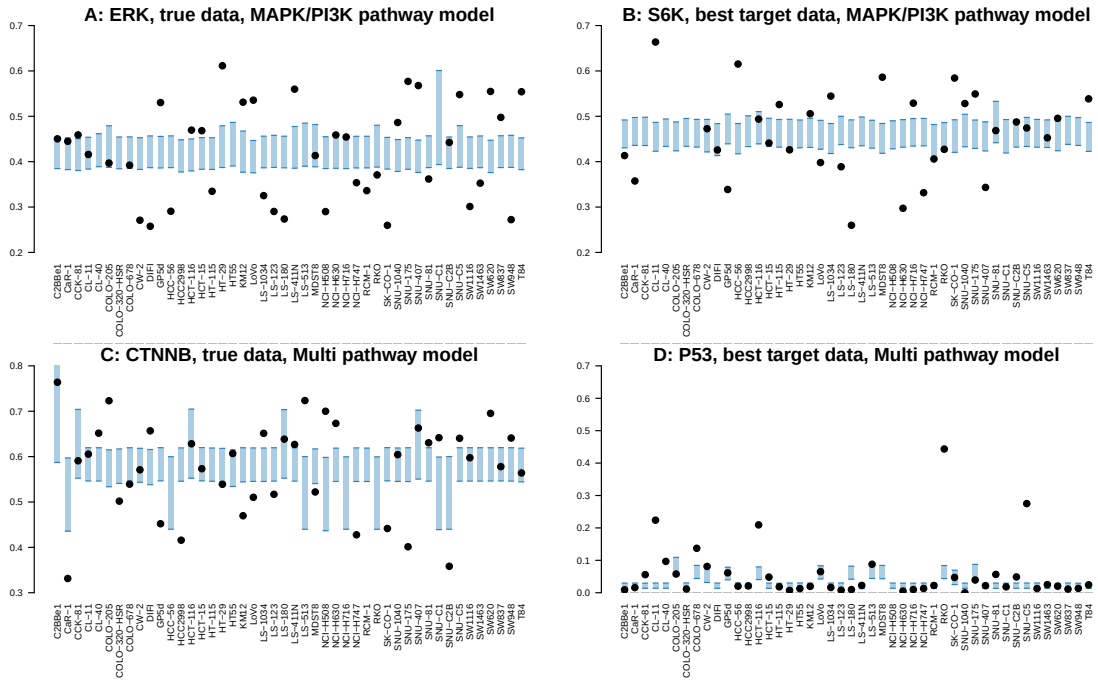


Figure 5.1: Posterior predictives with overlay observed data for the MAPK/PIK3CA pathway nodes A: ERK and B: S6K, and multi pathway nodes C: CTNNB, and D: P53.

Chapter 3

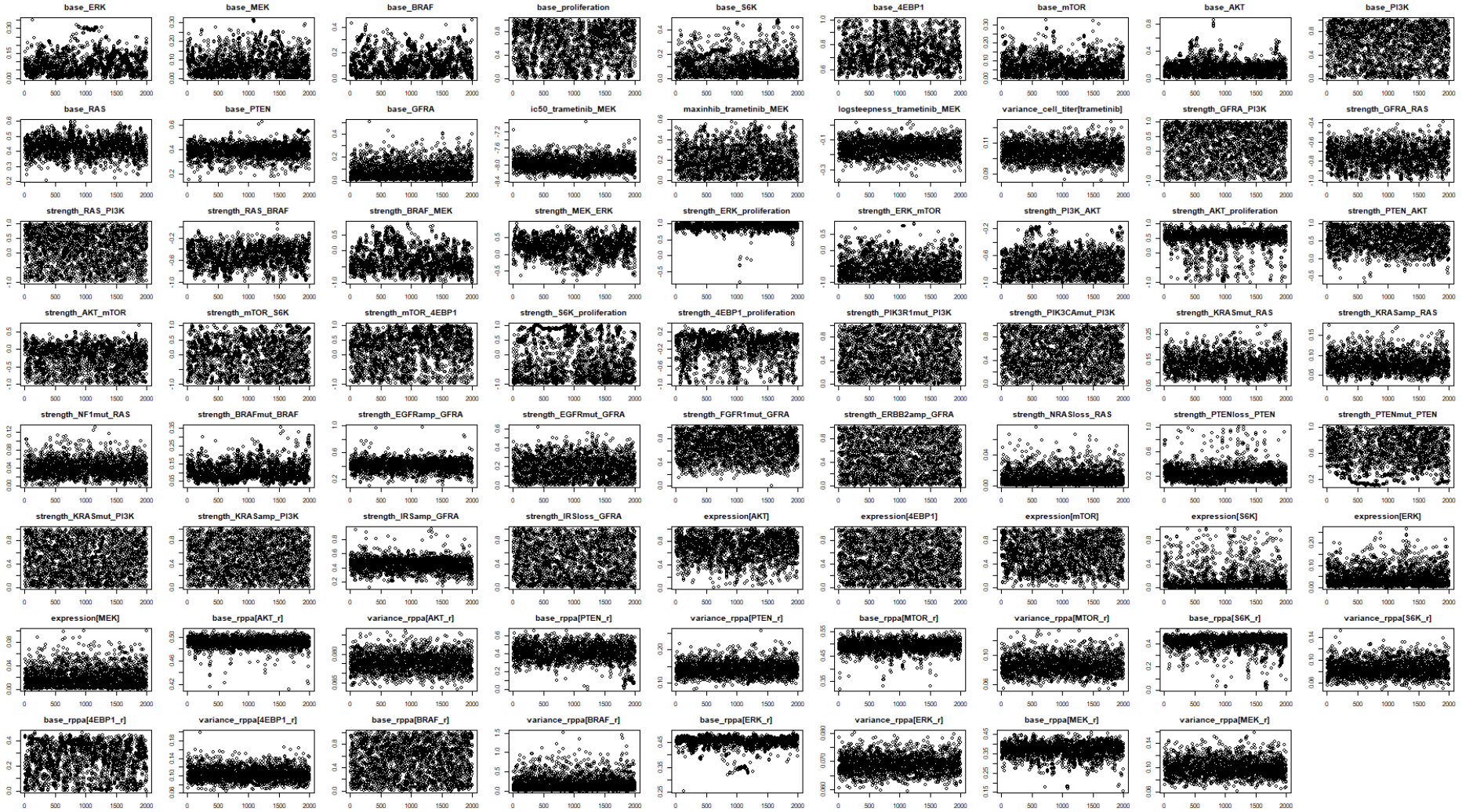


Figure 5.2: Sampling traces of all parameters in the MAPK/PI3K model with Trametinib and target node activation data. The parameters did not got stuck into one or two value or mode, indicating that the model converged well.

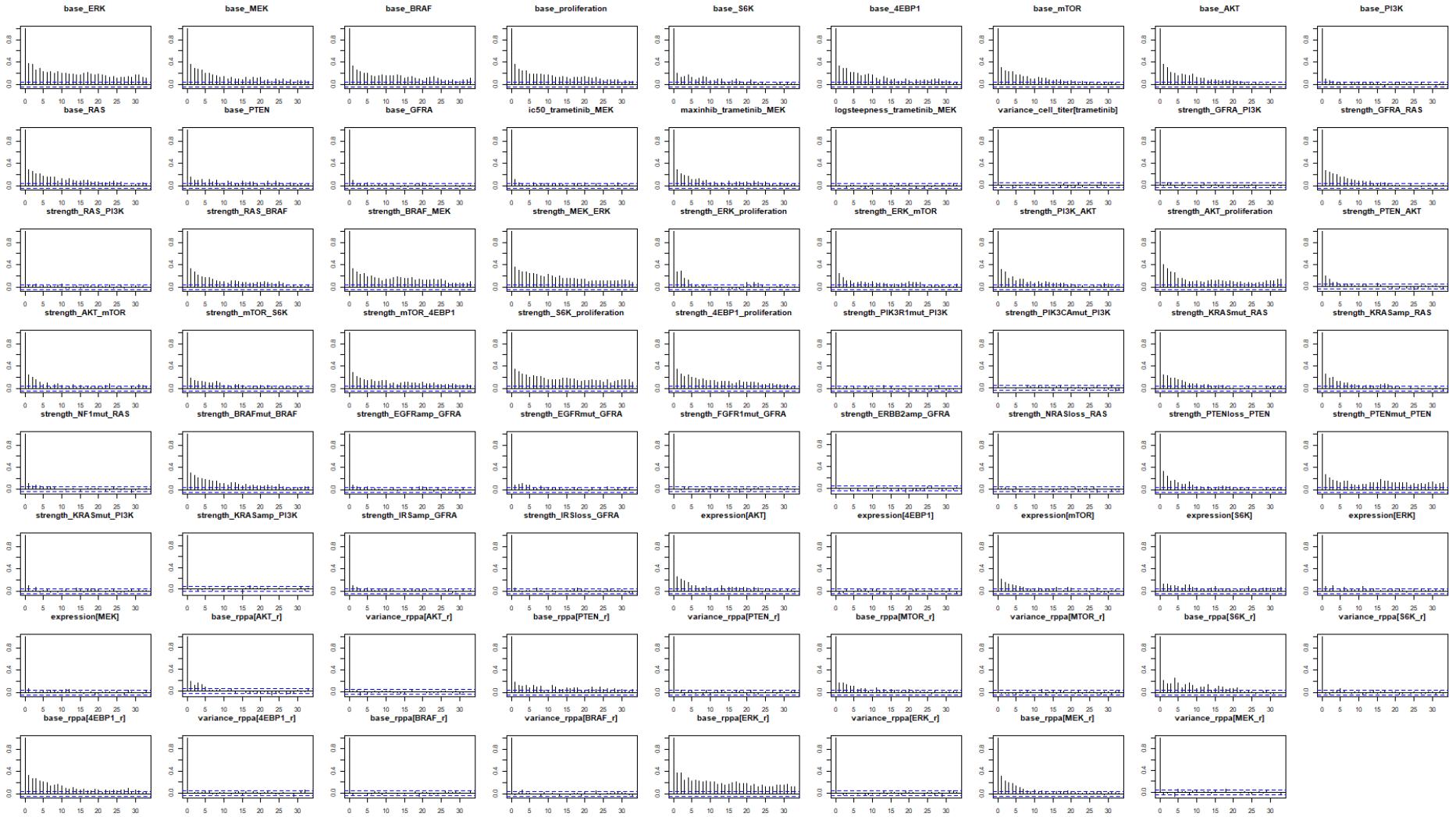


Figure 5.3: Autocorrelation of all parameters in the MAPK/PI3K model with Trametinib and target node activation data. The parameters did end up with large autocorrelation, meaning that the model converged well.

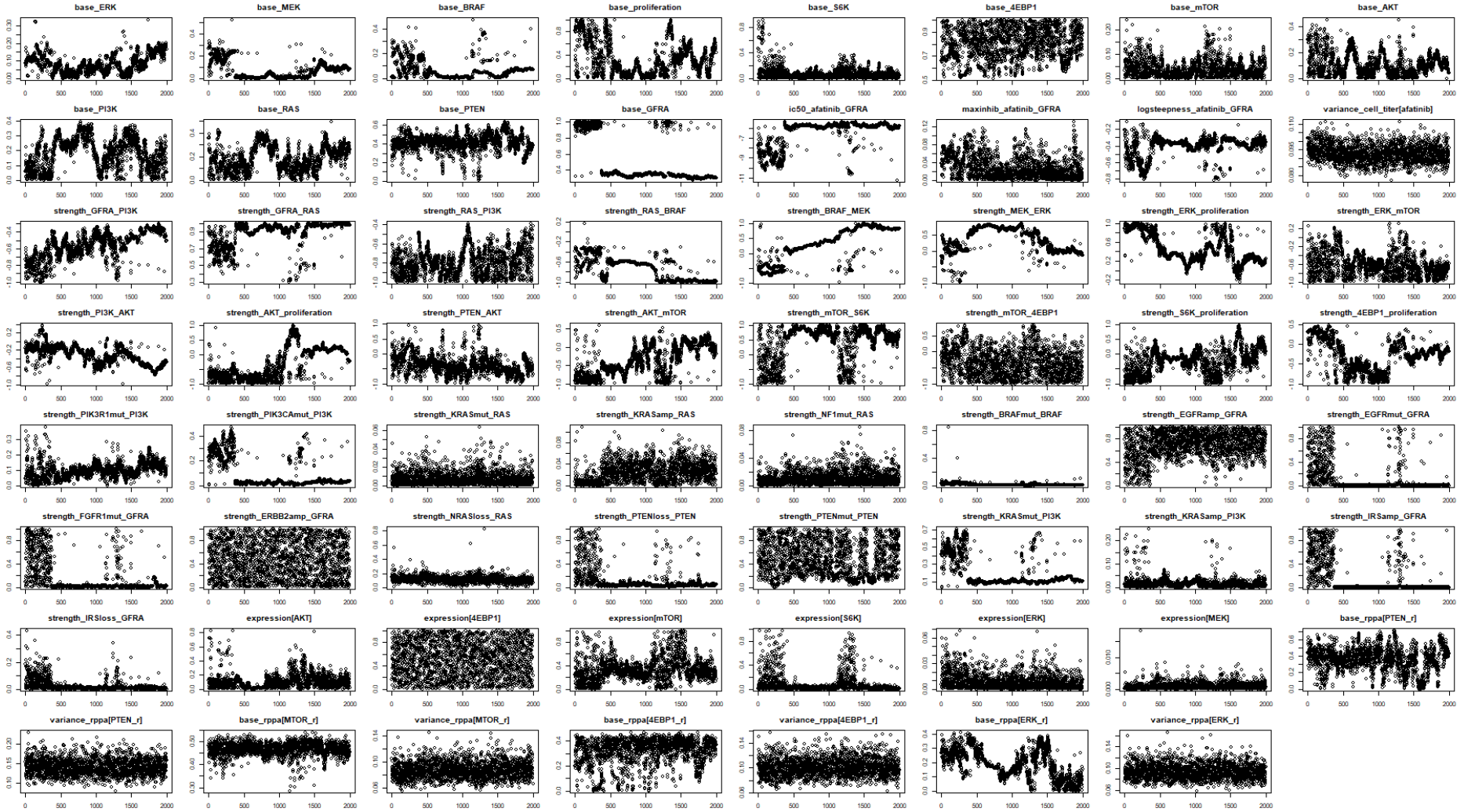
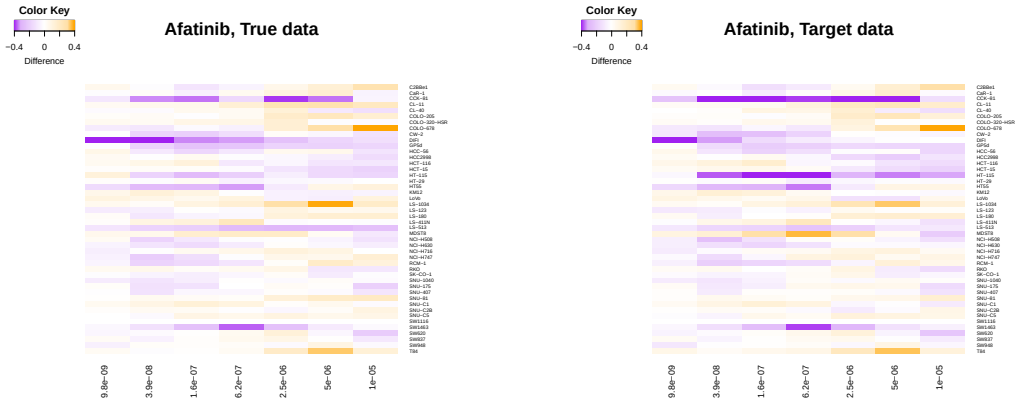


Figure 5.4: Sampling traces of all parameters in the MAPK/PI3K model with Afatinib and true node activation data. This is an example of bad convergence, where the parameters get stuck in modes, and fail to sample from effectively. The autocorrelation is very high for these parameters.

Table 5.3: Features weights of Elastic Net Regression using all features

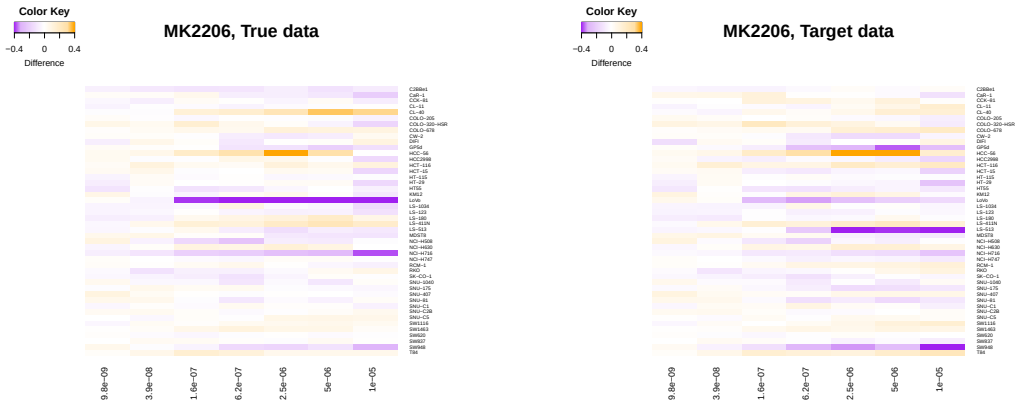
	MK2206	Trametinib	Afatinib	SCH772984	Afatinib+Trametinib
(Intercept)	-5.324	-8.096	-5.483	-6.304	-6.718
EGFR	0.000	0.000	0.000	0.000	0.000
KRAS	0.000	0.000	0.000	0.000	0.000
BRAF	0.000	0.000	0.000	-0.081	-0.003
NF1	0.000	0.004	0.000	-0.001	0.000
PTEN	0.000	0.000	0.000	-0.010	0.000
loss PTEN	**1.666**	0.000	0.000	0.000	0.002
PIK3CA	0.000	**0.250**	0.000	0.001	0.012
PIK3R1	0.000	0.000	0.000	0.000	0.000
loss NRAS	0.000	0.000	0.000	0.000	0.000
amp KRAS	0.000	0.000	0.000	0.000	0.000
loss IRS2	0.000	**0.102**	0.000	0.002	0.000
amp IRS2	0.000	-0.017	0.000	0.000	-0.055
amp FGFR1	0.000	0.000	0.000	0.000	**-0.176**
amp ERBB2	0.000	0.000	-0.005	0.000	0.000
amp EGFR	0.000	0.000	0.000	0.000	**-0.112**
...					
354 more					

(a)



(b)

Figure 5.5: Overview heat plot of differences between posterior predictive and observed data points, for each cell line over the 7 drug concentrations for the drugs a: MK2206 and b: SCH772984



(a)

(b)

Figure 5.6: Overview heat plot of differences between posterior predictive and observed data points, for each cell line over the 7 drug concentrations for the drugs a: Afatinib, and b: Drug combination

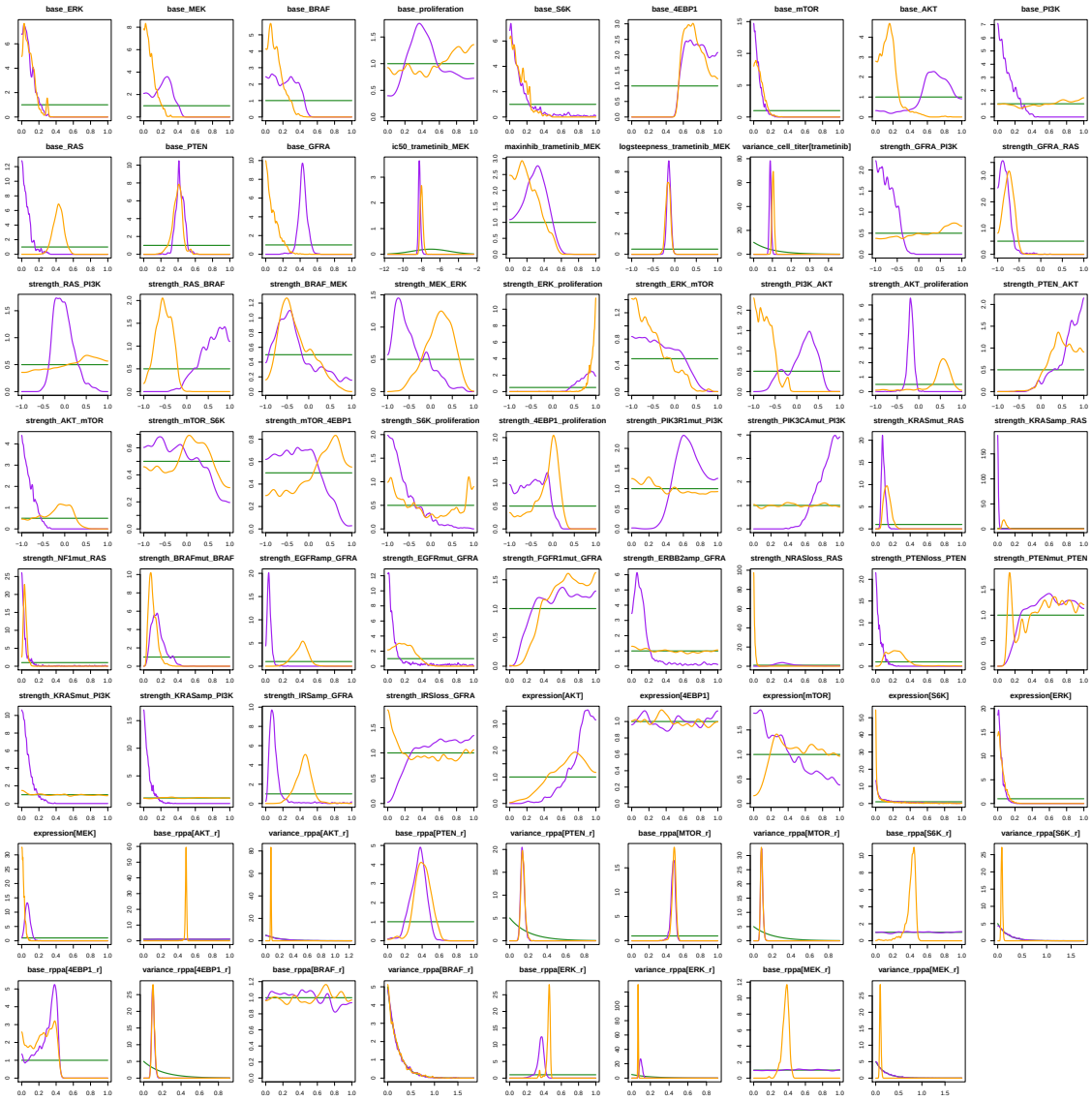


Figure 5.7: Marginal density distributions of the parameters in the Trametinib models. The target data model distributions are orange and the true data model's are purple. The prior distribution is depicted in green.

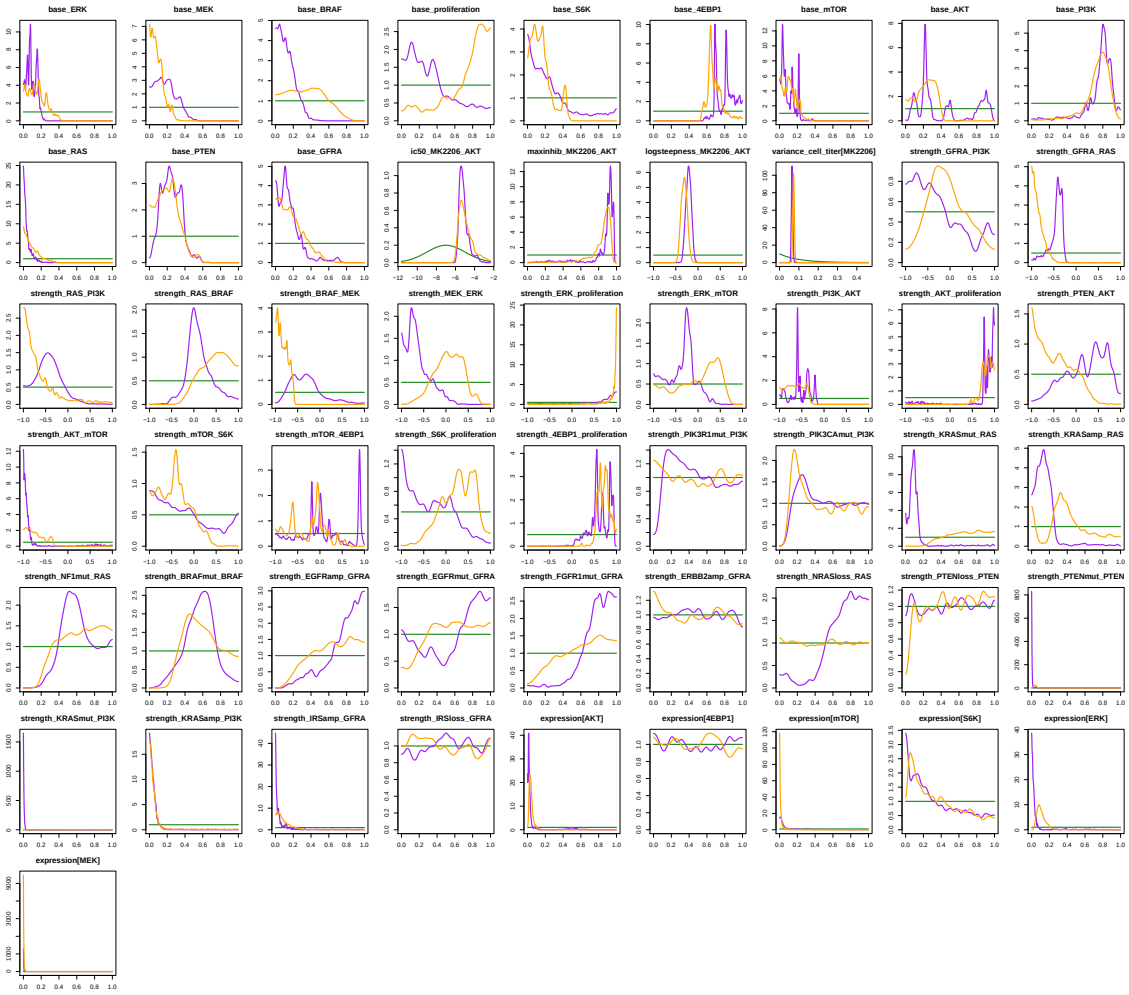


Figure 5.8: Marginal density distributions of the parameters in the MK2206 models. The target data model distributions are orange and the true data model's are purple. The prior distribution is depicted in green.

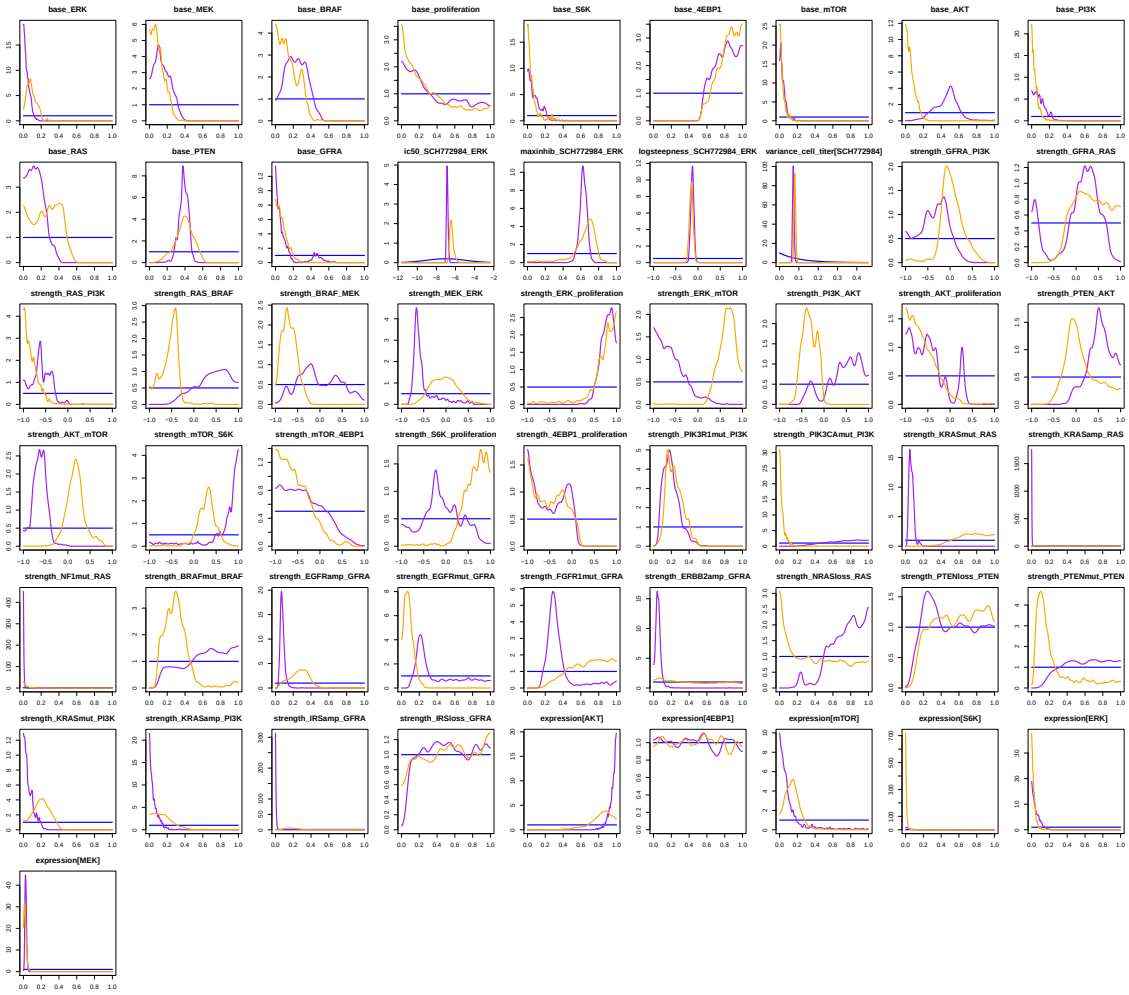


Figure 5.9: Marginal density distributions of the parameters in the SCH772984 models. The target data model distributions are orange and the true data model's are purple. The prior distribution is depicted in green.

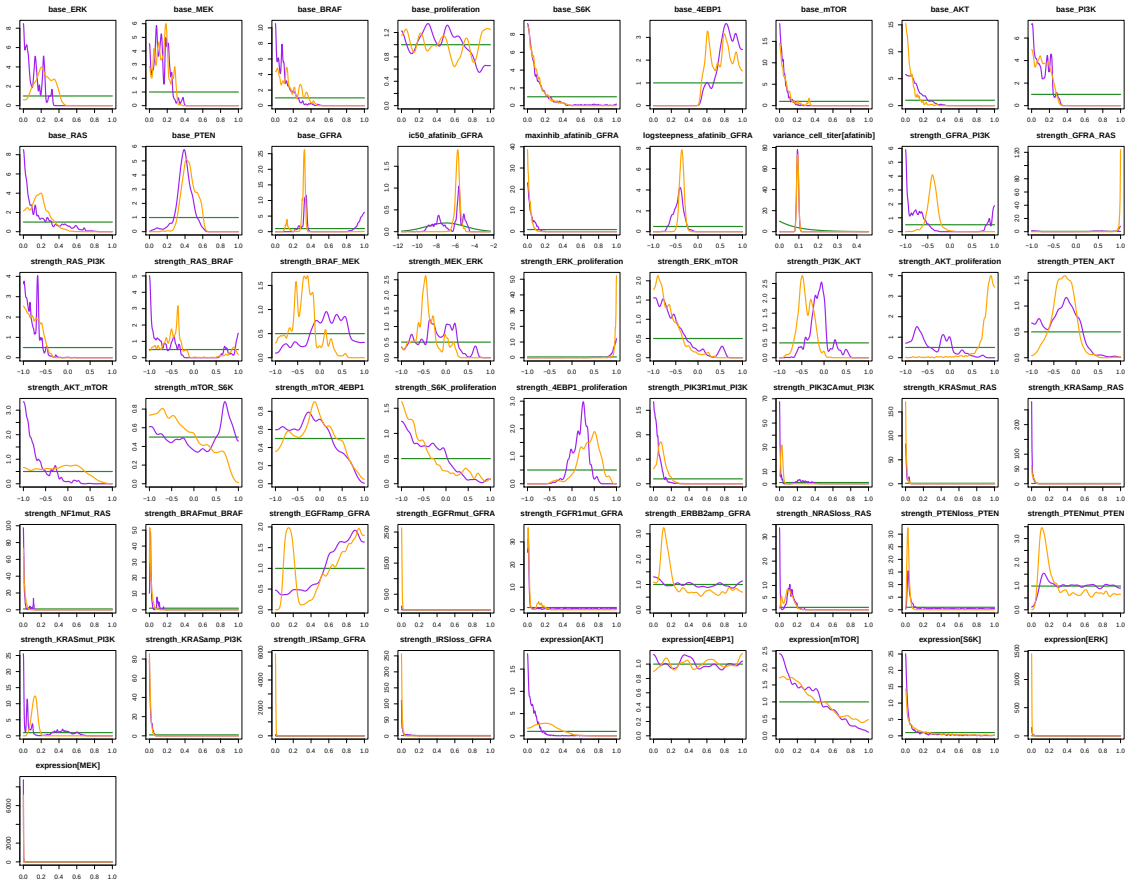


Figure 5.10: Marginal density distributions of the parameters in the Afatinib models. The target data model distributions are orange and the true data model's are purple. The prior distribution is depicted in green.

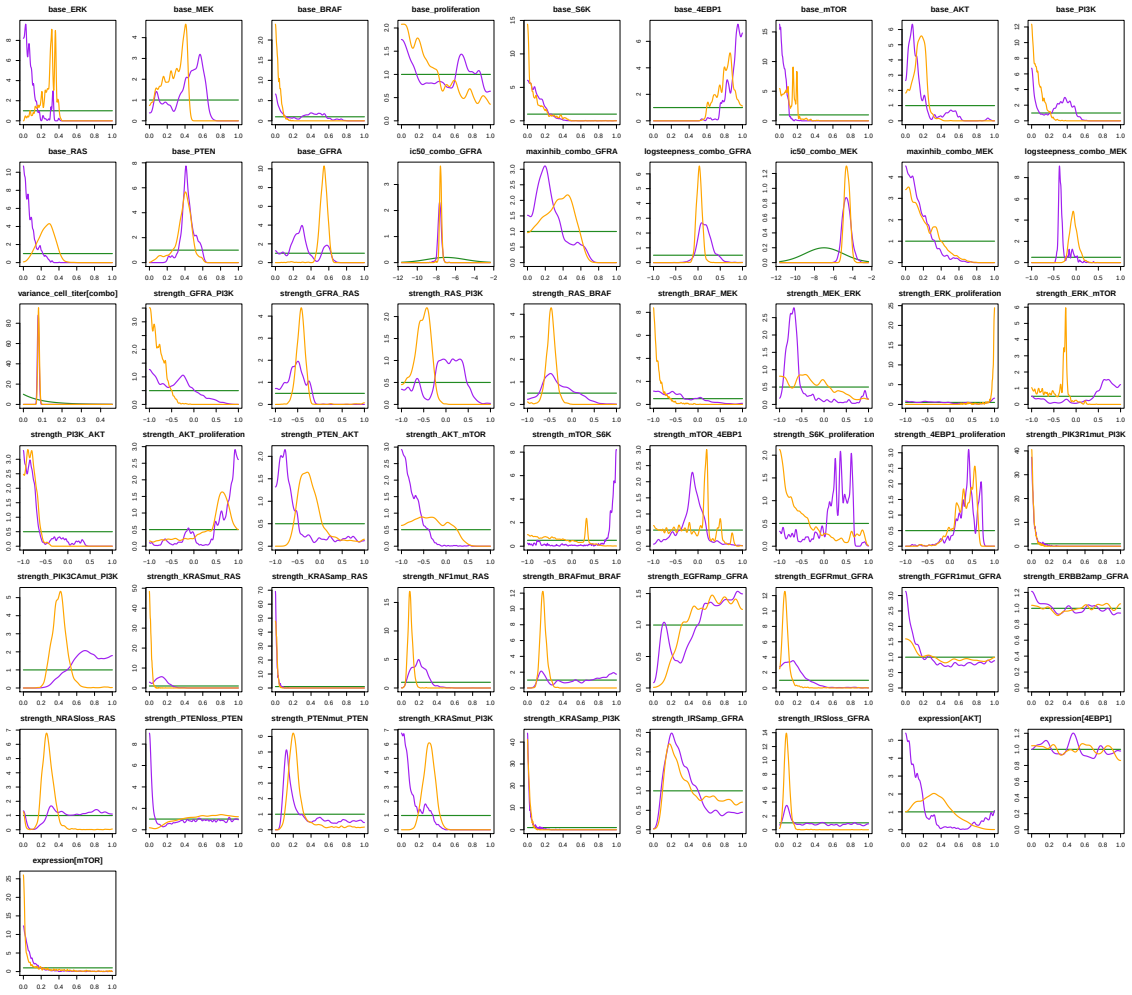


Figure 5.11: Marginal density distributions of the parameters in the drug combination models. The target data model distributions are orange and the true data model's are purple. The prior distribution is depicted in green.

Chapter 4

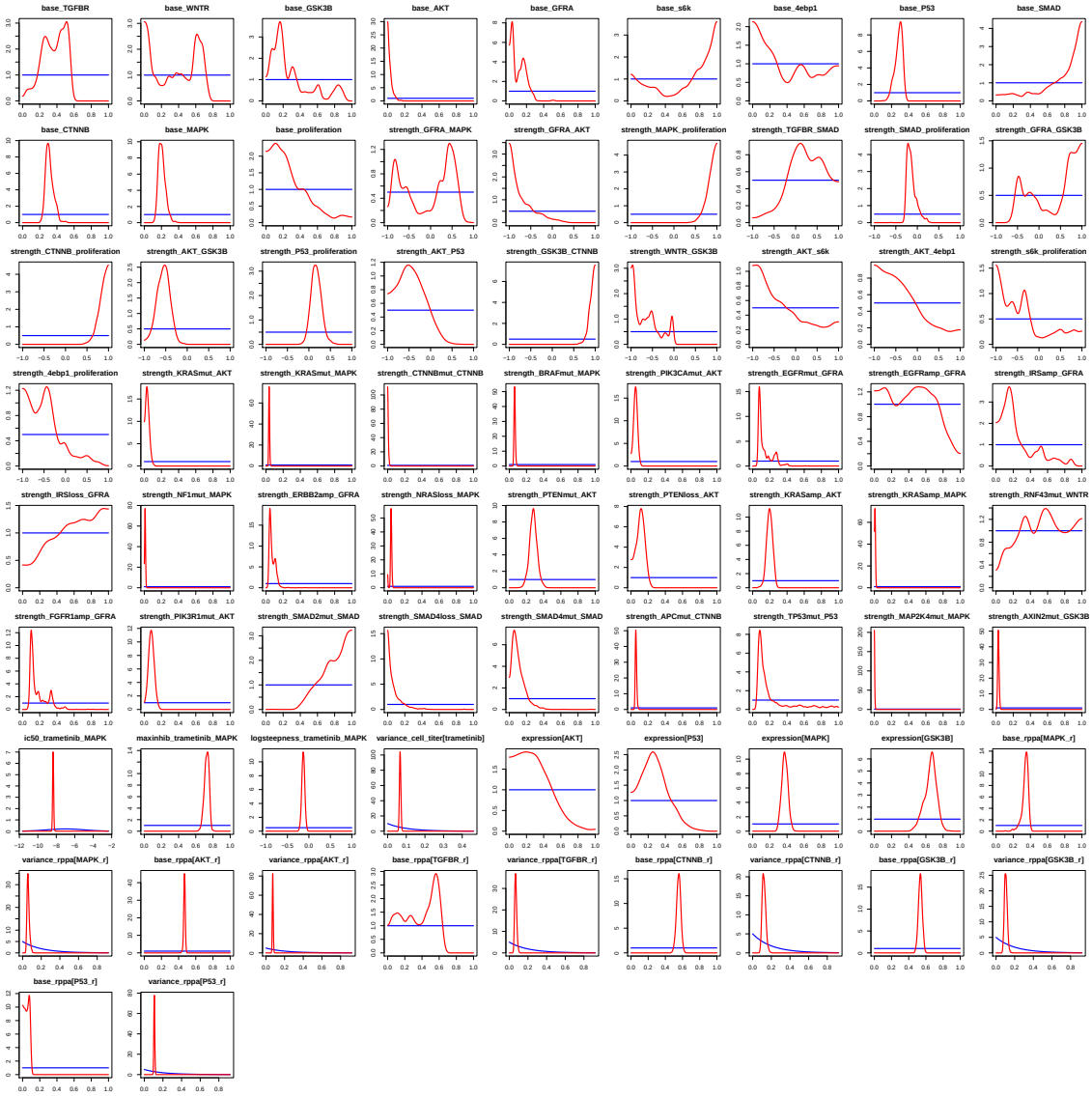


Figure 5.12: Marginal density distributions of the parameters in the Trametinib multi pathway model.

Table 5.4: Overview of all run models described in this thesis with computational times and convergence.

CH2		Comp time	Convergence
MAPK/PI3K pathway	True data	00:34:07	Very good
	Best target data	00:37:21	Very good
	n1 target sets, combined	00:36:40	Very good
	n2 target sets, combined	00:36:36	Very good
	n3 target sets, combined	00:36:51	Very good
	n1 target sets, separate	00:48:12	Very good
	n2 target sets, separate	00:54:20	Very good
	n3 target sets, separate	01:00:53	Very good
Multi Pathway	True data	01:50:59	Very good
	Best target data	02:04:18	Very good
	n1 target sets, combined	02:09:28	Very good
	n2 target sets, combined	02:09:31	Very good
	n3 target sets, combined	02:08:47	Very good
	n1 target sets, separate	02:44:24	Very good
	n2 target sets, separate	03:17:22	Very good
	n3 target sets, separate	03:53:02	Very good
CH3			
MAPK/PI3K pathway	Trametinib, true data	21:26:23	Good
	Trametinib, target data	21:35:37	Good
	MK2206, true data	17:51:56	Medium
	MK2206, target data	17:38:37	Good
	SCH772984, true data	18:23:21	Medium
	SCH772984, target data	21:43:11	Good
	Afatinib, true data	16:58:43	Bad
	Afatinib, target data	20:22:06	Medium
	Combination, true data	21:33:08	Bad
	Combination, target data	20:54:54	Medium
CH4			
Multi pathway	untreated, target data	02:38:42	Very good
	Trametinib, target data	17:11:36	Good
	MK2206, target data	16:31:02	Bad
	SCH772984, target data	17:49:53	Bad