Integrative modeling of inhibitor response in breast cancer cells

Thijssen, Bram

DOI
10.4233/uuid:8f52bab6-c097-4c4c-8291-3e76b0285f55

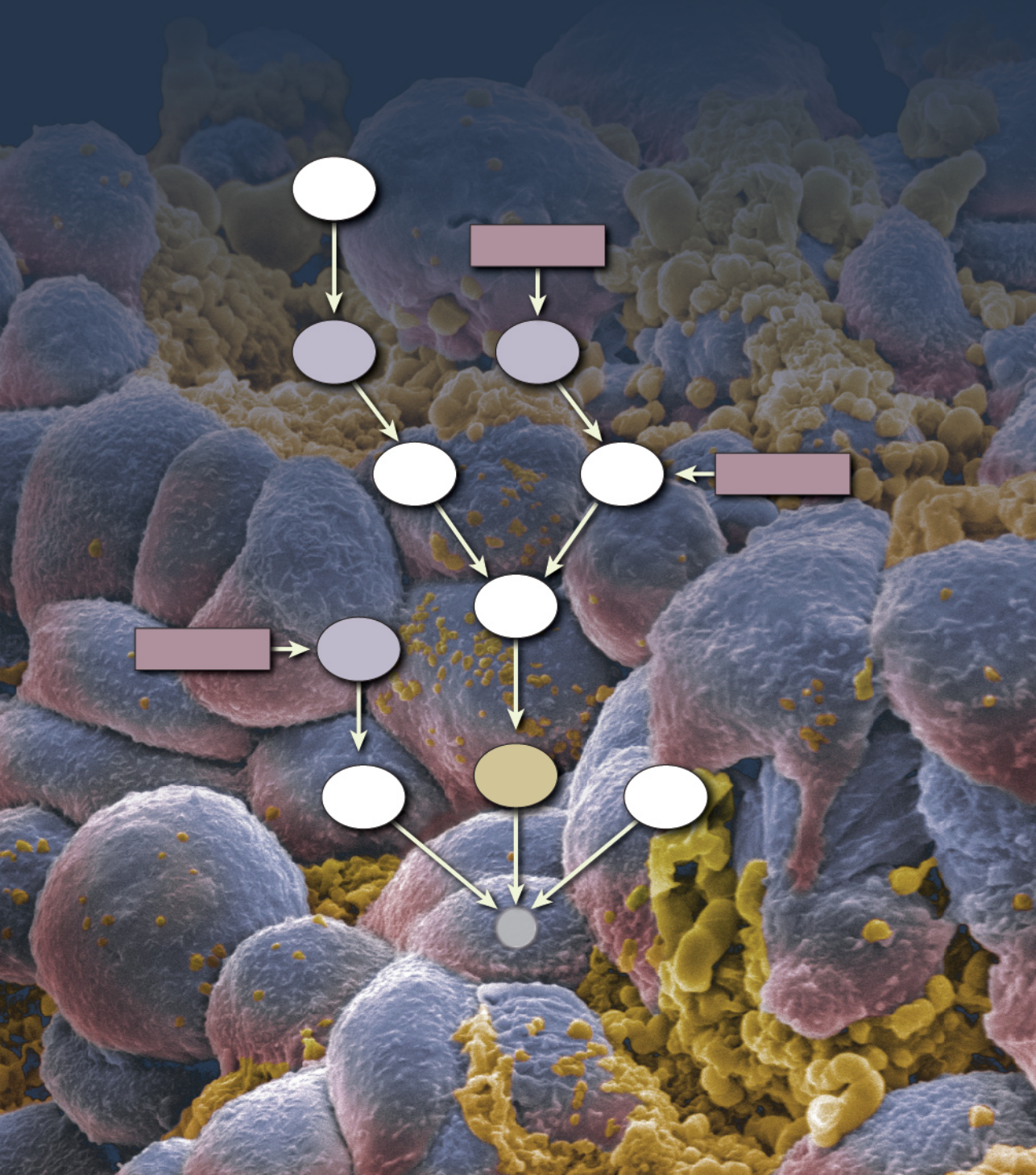
Publication date
2018

Document Version
Final published version

Citation (APA)
Thijssen, B. (2018). *Integrative modeling of inhibitor response in breast cancer cells*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:8f52bab6-c097-4c4c-8291-3e76b0285f55

Integrative modeling of inhibitor response in breast cancer cells

Bram Thijssen



INTEGRATIVE MODELING OF INHIBITOR RESPONSE IN BREAST CANCER CELLS

INTEGRATIVE MODELING OF INHIBITOR RESPONSE IN BREAST CANCER CELLS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op dinsdag 16 oktober 2018 om 15:00 uur.

door

Bram THIJSSEN

Master of Science in Computational Biology and Bioinformatics, Eidgenössische
Technische Hochschule Zürich en Universität Zürich, Zwitserland;
geboren te Gouda, Nederland.

Dit proefschrift is goedgekeurd door de promotor.

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof. dr. L.F.A. Wessels	Technische Universiteit Delft, promotor
Onafhankelijke leden:	
Prof. dr. B.M. Bakker	Rijksuniversiteit Groningen
Prof. dr. F.J. Bruggeman	Vrije Universiteit Amsterdam
Prof. dr. ir. M.J.T. Reinders	Technische Universiteit Delft
Prof. dr. M.A. van de Wiel	Vrije Universiteit Amsterdam
Dr. P. Kemmeren	Prinses Maxima Centrum
Dr. G.S. Sonke	Antoni van Leeuwenhoek - Nederlands Kanker Instituut
Prof. dr. R.C.H.J. van Ham	Technische Universiteit Delft, reservelid

Printed by: Ipskamp Printing

Front & Back: Original image of breast cancer cells undergoing programmed cell death by Annie Cavanagh. Cover design by Bram Thijssen.

ISBN 978-94-6186-958-6

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

CONTENTS

1

INTRODUCTION

1

1.1. VARIABILITY IN ANTICANCER DRUG RESPONSE

ONE of the central objectives of current cancer research is to design optimal, individualized treatment plans for patients. The main difficulty with this is that every patient can respond very differently to a given treatment. Some patients respond very well, while others do not respond at all, leaving the cancer to grow unimpeded. If we have a good understanding of how this variability in response arises, we will be better able to choose the optimal treatment strategy for each patient.

In some cases it is straightforward to explain why one patient responds to a particular treatment while another patient does not. An example of this is the presence of a particular mutation that is largely responsible for driving the growth of the tumor. Such a mutation can be so important, that the tumor is dependent on it for its persistence. In this situation, giving a drug that blocks the effect of this mutation should eradicate the tumor. In breast cancer, amplification of the *ERBB2* gene is such an example. And indeed, breast cancer patients who have an *ERBB2* gene amplification can respond very well to trastuzumab [1] or lapatinib [2], two drugs which target the protein encoded by this gene. Even in this case, however, there is variability in response; for some patients the tumor shrinks impressively, while other patients relapse quickly, and yet other patients do not respond at all.

This variability in response is illustrated in Figure 1.1A for lapatinib. The data shown is from a phase II clinical trial [3] in 34 patients with *ERBB2*-amplified breast cancer which has metastasized to the brain, and which is refractory to standard chemotherapy as well as to trastuzumab treatment. The response was measured by the change in lesion size on an MRI scan, after eight weeks of treatment. It is immediately clear that there is large variability in response between patients. For the patient represented by the left-most bar, the brain metastases grew by more than 150% during the eight weeks of treatment, whereas for the patient represented by the right-most bar, the brain metastases almost completely disappeared. Unfortunately, a strong response was rare in this trial, and a subsequent larger phase II trial confirmed that only few patients respond in this setting [4]. It is also unclear whether the change in lesion size is clinically relevant, as this trial did not meet the pre-specified criteria for efficacy in this setting. Nevertheless, if we knew upfront which patients would have significant decreases in lesion size upon treatment with this drug, it would likely be beneficial to give this drug to those patients.

The variability in drug response observed in patients is also seen in cancer cell lines which are cultured in vitro. Recently, several large screening efforts have profiled over 1,000 cell lines for their response to hundreds of different anticancer drugs and other molecular compounds [5–9]. These studies have all shown that cancer cell lines, including those derived from breast cancer, respond differently to many compounds. We have also profiled a smaller panel of thirty breast cancer cell lines for their response to a number of kinase inhibitors, which is discussed in detail in Chapter 4. The response to one such inhibitor, AZD8055, targeting mTORC1/2, is shown in Figure 1.1B. What is clear from this figure is that some cell lines are affected by AZD8055 treatment at very low concentrations, while other lines can continue to grow despite treatment even at high concentrations.

Can the presence of one particular mutation determine the variability in response to AZD8055? One important gene which is recurrently mutated in breast cancer is *PIK3CA*,

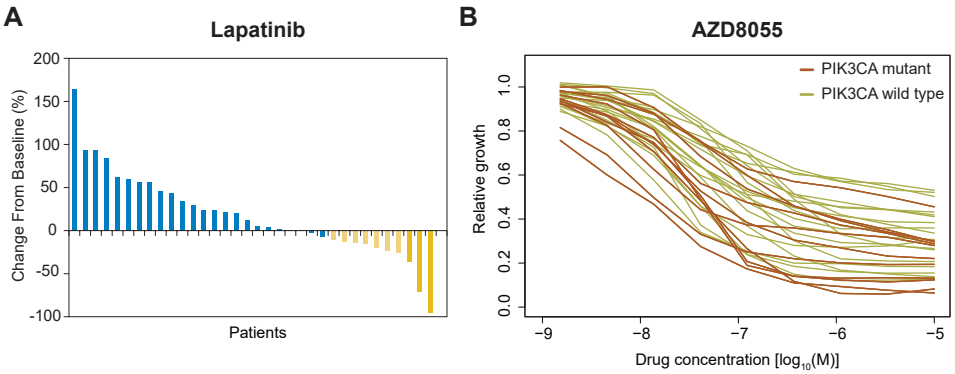


Figure 1.1: **Variability in kinase inhibitor response.** (A) Response of 34 breast cancer patients with brain metastases to the EGFR/HER2 inhibitor lapatinib; figure adapted from [3] with permission. Bars indicates the change in the sum of target lesion sizes on MRI scan after eight weeks of treatment. (B) Response of 30 breast cancer cell lines to the mTOR inhibitor AZD8055. Relative growth indicates how much each cell line can grow during three days of treatment, compared to an untreated control. Each curve indicates one cell line. This cell line data is described in more detail in Chapter 4.

the catalytic subunit of phosphoinositide 3-kinase, a kinase that signals through the mTOR pathway (discussed in more detail below). If cancer cell lines with activating PIK3CA mutations are dependent on these mutations for growth and survival, we would expect that these cell lines would be particularly sensitive to mTOR inhibitors like AZD8055. We can see in Figure 1.1B that cell lines which have such mutations (colored in brown) indeed do tend to be more sensitive than wild-type cell lines (colored in green). However, there are also *PIK3CA*-mutant cell lines that show resistance to this inhibitor, as well as wild-type cell lines which are just as sensitive to treatment as some of the mutant cell lines. It is clear that in this case, this one single factor (*PIK3CA* mutation status) cannot adequately explain the variability in response, and it is necessary to take more explanatory factors into consideration if we are to understand the response of all cell lines.

Indeed, many different factors which affect drug sensitivity and resistance have now been discovered. For example, for lapatinib, besides amplifications of *ERBB2* and mutations in genes encoding PI3K, also the loss of *PTEN* [10] or the expression of growth factors by the cancer cells themselves [11] has been found to influence drug response. For AZD8055 these factors also play a role, and yet other genetic aberrations, such as amplification of *MYC* [12], have been found to cause resistance. Given the multitude of factors that can influence drug response, it is no longer feasible to straightforwardly predict whether particular cancer cells, which may possess any combination of these factors, will be sensitive or resistant to a particular drug.

This brings us to the central topic of this thesis. At the outset I asked, given all of our knowledge, how much of the variability in drug response can we actually explain? How far can we get in explaining the variability in response across different breast cancer cell lines, if we put our available knowledge in an extensive computational model? Research into this question has two main goals. First, if we can construct models that are at least partially explanatory, they may be useful building blocks or stepping stones from which

we can create models of how patients will respond, and ultimately aid in optimizing personalized treatment plans. Second, systematically putting existing knowledge relating to drug response into a mathematical framework allows for consolidation of knowledge as well as systematic identification of gaps in that knowledge.

To address these goals, we need to construct knowledge-based models, which will be described further below. In addition, there are two important considerations that drove the choice in the modeling approach that we took. The first consideration is the inclusion of multiple, diverse data types. In other words, we wanted to generate an integrative model. There are various different measurement technologies available today with which we can profile cancer cells, and each of the resulting measurements provides us with different information. Combining these different types of data should allow us to get a more complete picture of what is happening within the cancer cells. To make this integration of data types possible, we developed novel statistical methods that can combine knowledge-based computational models with multiple data types.

The second important consideration in this thesis is the characterization of uncertainty in model parameters. Computational models typically have various parameters, and these parameters are still often set to a single value, typically the maximum likelihood value — the value that best describes all of the data. However, different parameter values may be able to explain the data just as well. If the uncertainty in parameter estimates is not taken into account, we can be lulled into a false sense of security and misinterpret which elements of the model are important. To take this into account, we characterized the full, joint uncertainty in all parameters, using Bayesian statistics.

There are many different kinds of anticancer drugs, including targeted kinase inhibitors, but also traditional chemotherapeutics such as platinum compounds or taxanes. Most recently, there has been increasing interest in the development and use of drugs that target the immune system, such as checkpoint inhibitors. In this project, we restricted ourselves to one class of drugs, the targeted kinase inhibitors. These inhibitors have a relatively well-defined mechanism of action, as they generally inhibit one particular kinase or group of kinases. If we are successful in modeling cancer cell response to kinase inhibitors, the approach could later be applied to other classes of anticancer drugs as well.

After a brief discussion of our knowledge of signaling and drug sensitivity to date, I will describe several computational modeling approaches that have been taken, and how our approach differs from them. The introduction is concluded with an overview of the scientific chapters.

1.2. ONCOGENIC SIGNALING AND DRUG SENSITIVITY

One of the defining characteristics of a tumor cell is that they are in a deregulated state of continuous growth and proliferation. Normal cells typically only divide when directed to do so by their microscopic environment, but cancer cells acquired certain mutations that cause them to grow and proliferate in the absence of an outside signal [13, 14]. Large-scale genomic studies have identified many genetic events which are recurrently found in breast cancers [15–18]. Since these mutations occur more often than would be expected by chance, it is likely that they contribute to the growth or development of the tumor.

While these genomic studies provided catalogs of important genes which are likely to be involved in cancer, more focused cell biology studies have unraveled many details of the signaling networks in which these genes are involved. Some of the major signaling events in mitogenic and survival signaling pathways have been elucidated more than twenty years ago [19, 20], and additional details of these signaling networks continue to be discovered. For example, in 2005 it was found that mTORC2 is the kinase responsible for phosphorylation of AKT on S473 [21], and more recently it was described how this phosphorylation is regulated by the mTORC2 complex member SIN1 [22]. Extensive reviews of signaling in these pathways are available [23–26].

In breast cancer, several of the most frequently occurring oncogenic mutations are centered on the MAPK and AKT pathways. I already mentioned amplifications of the *ERBB2* gene and mutations in *PIK3CA*. Several other recurrent genetic aberrations are depicted in Figure 1.2, along with a simplified overview of the signaling network. At the cell surface, receptor tyrosine kinases (RTKs) are normally activated by growth factors, but they can be aberrantly activated in cancer by gene amplification (such as for *ERBB2* and *FGFR1* in breast cancer) or through inappropriate autocrine signaling. These receptors then activate the MAPK and AKT pathways (not shown in detail). Downstream of the receptors, oncogenic driver mutations are known to occur in genes encoding PI3K and AKT, while the loss of negative regulators, such as PTEN, can also have activating effects on this pathway.

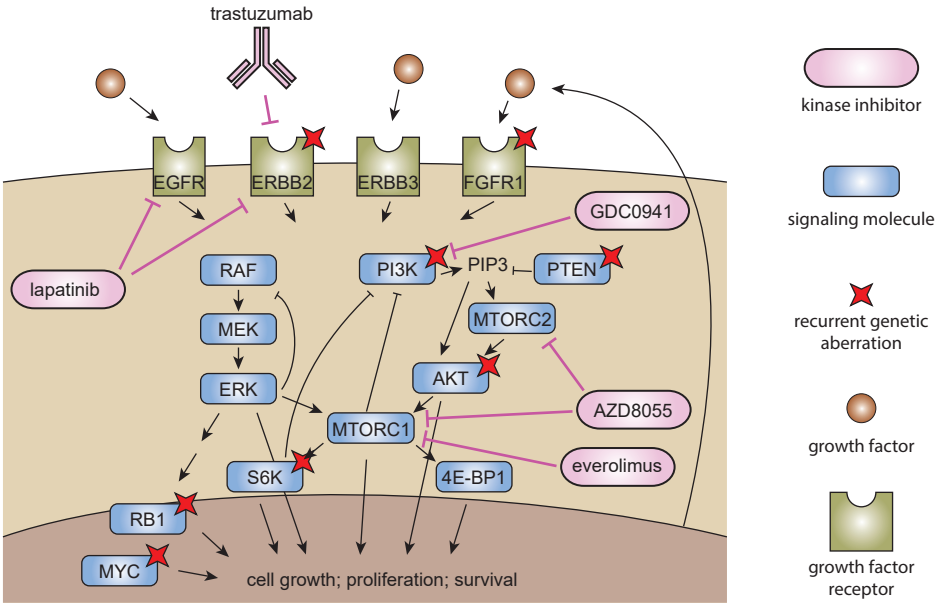


Figure 1.2: **Simplified overview of oncogenic signaling in the MAPK and AKT pathways in breast cancer.** Blue boxes indicate signaling proteins. Red stars incidate recurrent genetic aberrations in breast cancer, including mutations and gene amplifications and losses. Pink items are anticancer agents either used in clinical practice or which are in (pre-)clinical development, including several kinase inhibitors and one antibody drug.

Given the importance of these signaling pathways, many candidate drugs have been developed to target them at different points. Three drugs which are used in clinical practice are lapatinib (EGFR/HER2 inhibitor), trastuzumab (anti-HER2 antibody) and everolimus (an allosteric mTORC1 inhibitor), while other compounds, including second-generation mTOR inhibitors, such as AZD8055, and PI3K inhibitors, like GDC0941, are in pre-clinical or clinical development. Everolimus has been found to prolong progression-free survival in hormone receptor positive breast cancer [27], although this benefit did not extend to an improved overall survival [28]. In *ERBB2*-amplified breast cancer, there is only modest benefit of everolimus [29, 30]. One downside of everolimus is that feedback pathways can result in re-activation of AKT and mTOR, which may be responsible for the limited therapeutic benefit of the inhibitor [31]. To prevent this re-activation, second-generation mTOR inhibitors have been developed which inhibit mTORC2 as well as mTORC1 [32]. For example, the mTORC1/2 inhibitor AZD2014, an analog of AZD8055 with optimized pharmacokinetic properties, has passed phase I trials [33] and is currently in multiple phase II trials, although it has to be noted that such potent mTOR inhibitors have significant side effects [34]. PI3K inhibitors have also been suggested to be a promising treatment for PI3K-mutant tumors, although the PI3K-inhibitor GDC0941 did not show a benefit in two separate phase II clinical trials [35, 36].

These clinical results indicate that, despite the likely importance of genetic aberrations in these pathways, a blockade of one signaling molecule using kinase inhibitors has variable or limited effect. A better understanding of which mutations and signaling pathways are most important in driving the growth and survival of the tumor may allow us to better intervene with kinase inhibitors.

1.3. COMPUTATIONAL MODELING OF KINASE INHIBITOR RESPONSE

Various approaches have been taken to construct computational models of anticancer drug response. These can be broadly classified in two categories: 'black box', machine learning models and knowledge-based, mechanistic models. The black box models, such as elastic net regression and random forests, have been used to uncover new factors which are associated with drug sensitivity or resistance [5, 6, 8, 37]. These approaches do not employ our existing knowledge of cell biology however, and are therefore not a good tool for testing whether our knowledge can explain the variability in response. Among the knowledge-based, mechanistic models, four approaches are most directly related to the topic of this thesis.

Saez-Rodriguez et al. [38] used a logic modeling framework to model signaling in four hepatocellular carcinoma cell lines and in primary hepatocytes. They profiled the phosphorylation levels of sixteen proteins in these cells, in various stimulated conditions before and after treatment with three kinase inhibitors. The inferred model highlighted several signaling events which are different between primary and transformed cells. With this approach, they discovered that the IKK-inhibitor TPCA-1 also inhibits the phosphorylation of STAT3 by JAK2 [38]. A later report indicated that this effect may be due to the inhibitor binding STAT3 rather than inhibiting JAK2 [39].

Klinger et al. [40] used an extended version of modular response analysis (MRA, [41]) to quantify signaling in six colon cancer cell lines. This framework is based on a linearization around steady state, and uses intervention experiments to quantify the local

interactions between nodes, based on global measurements. The authors extended MRA by using a maximum likelihood approach to allow estimation of the response coefficient in a setting where not all nodes had been perturbed [42], and by searching for an optimal, sparse network structure [40]. This network structure optimization is done by iteratively removing edges from a starting network. They measured phosphorylation of eight proteins after stimulation with growth factors and inhibition with four kinase inhibitors. With this approach, they discovered that MEK inhibition leads to EGFR-mediated activation of AKT, an important mechanism which had concurrently been discovered through functional genetic screening [43].

While the previous two modeling frameworks describe cell lines at steady state, Kirouac et al. [44] constructed a model describing dynamics over time. They constructed a simplified representation of signaling by ErbB family receptors, AKT and ERK, and estimated signaling strengths from phosphorylation measurements of five proteins, after stimulation with growth factors and inhibition with two drugs in 25 different combinations of concentrations, in a single cell line. This analysis confirmed the importance of transcriptional feedback from AKT to ERBB3. The model further indicated that a triple combination of trastuzumab, lapatinib and the bi-specific antibody MM-111 should be more effective than a combination of a MEK and an AKT inhibitor. In an in vivo model with BT-474 cells injected into mice, the triple combination was indeed more effective at controlling tumor growth, with less toxicity as judged by animal weight.

Very recently, Eduati et al. [45] combined a discrete logical framework with dynamic modeling. In this case, fourteen colorectal cancer cell lines were used, and fourteen epitopes were measured in 43 conditions. These results will be discussed more extensively in the discussion of this thesis.

Together, these four studies show that computational modeling of signaling in cancer cells can be useful to better understand response to anticancer drugs and that they can allow for discovery of factors associated with response. However, all four of these approaches also had several limitations, as outlined below.

The logic modeling approach of Saez-Rodriguez et al. [38], although it scales relatively well to larger models, is limited by using a binary description. From Figure 1.1 it is clear that the variability in response is continuous, with a wide spectrum of response, rather than a clear dichotomy. To adequately describe the variability, therefore, a continuous framework seems necessary. The model of Kirouac et al. [44] did use continuous variables (in addition to modeling dynamics over time), but in this case the model was fairly small, encompassing only four intracellular signaling nodes. Furthermore, all four approaches mentioned here used only a single data type, protein phosphorylation, to infer the signaling strengths. This limits the insight that can be obtained, as, for example, protein abundance or the presence of mutations and gene amplifications is not taken into account. Finally, the number of cell lines used was limited (5, 6, 1 or 14 cell lines, respectively). It is therefore unclear whether these models would be able to describe variability across a larger panel of cell lines.

In this thesis, by combining a relatively detailed model, along with continuous variables, multiple types of measurement data and a larger cell line panel (30 cell lines), we aimed to arrive at a more detailed understanding of the variability in response to kinase inhibitors.

1

1.4. ANALYZING UNCERTAINTY

As mentioned earlier, an important consideration in this thesis is the analysis of uncertainty. Many computational models rely on parameters to describe the system of interest. In some situations, values for the parameters can be derived from biochemical reasoning or dedicated experiments. Usually however, parameter values are obtained by fitting the model to the available data. It is now increasingly appreciated that the uncertainty in the parameter values needs to be considered [46–48]. Although a single parameter value often provides the unique best fit for the data, other parameter values may describe the data almost equally well. Characterizing the range or distribution of parameter values that adequately describe the data allows for more robust model predictions.

In this thesis, we have chosen to use Bayesian statistics for the characterization of uncertainty. Bayesian inference can be computationally demanding, but it has several benefits. First, it allows us to include prior information about parameter values when this is available. For example, the concentration at which drugs inhibit their target has often been measured using in vitro kinase assays, and this information can be used in a semi-informative prior. At the same time, the inference still allows for deviations from this semi-informative prior if other data support alternative values. A second benefit, compared to profile likelihoods for example, is that the Bayesian inference allows the characterization of the joint uncertainty in all parameters together. Particularly in chapters 5 and 6, we will see several situations where the joint uncertainty provided additional information. Third, the inference naturally allows for the inclusion of multiple different types of data. Each dataset can be used to further update the probability distribution of the parameters.

Figure 1.3 shows the general approach we used for inferring the parameters. A prior probability was specified, which was then updated based on several datasets obtained using different types of measurements (three datasets shown here). This gave us a posterior distribution describing which parameter values are consistent with all data together. In the example case shown in the figure, viability data of cells treated with the kinase in-

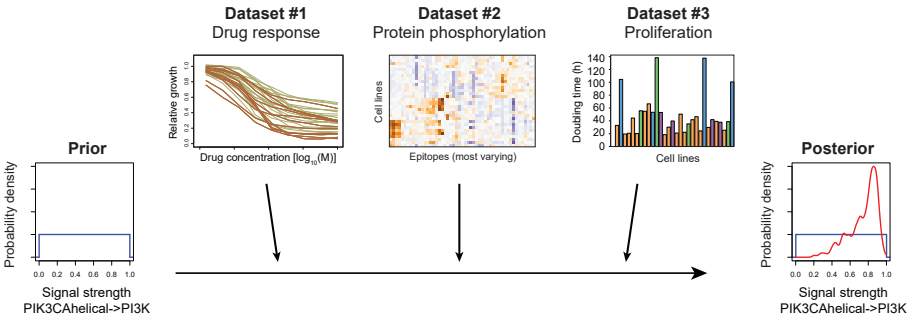


Figure 1.3: **Outline of the integrative Bayesian analysis used in this thesis.** The prior probability distribution for all parameters (shown here for only 1 parameter, specifically the proliferative signal strength arising from mutations in the helical domain of *PIK3CA*) is updated based on multiple types of measurements, to give a posterior distribution describing the most likely values for the parameters.

hibitor AZD8055, along with protein phosphorylation data and the growth rates of all the cell lines in untreated conditions, was used to infer the signaling strengths. Although mutation and copy number data is not shown here, these were included as well, but as constraints rather than as inference data (see Chapter 4 for details). In this example, the parameter describing how strongly mutations in the helical domain of the *PIK3CA* gene activate the PI3K protein is shown. Together, the data indicate that this mutation indeed has a strong effect, in line with the association of *PIK3CA* mutations with mTOR inhibitor sensitivity [49].

A challenge with Bayesian inference is obtaining an accurate representation of the posterior distribution. To estimate the posterior distribution, we used several variants of Monte Carlo sampling. Since the conception of the Metropolis-Hastings algorithm [50, 51], a specific type of Monte Carlo sampling, many variants and improvements have been developed, such as adaptive proposal distributions [52], parallel tempering [53] and temperature optimization [54]. An alternative approach to these Markov chain-based methods are the sequential Monte Carlo samplers [55], which also continue to be improved upon [56]. In cosmology and physics, nested sampling using the MultiNest algorithm is a popular method for Bayesian inference [57]. Before starting an inference, it is not always clear which of these methods will perform best, and we therefore considered each of these approaches.

1.5. OUTLINE OF THIS DISSERTATION

At the start of this project, existing software packages for Bayesian inference were either too inefficient, or not sufficiently flexible, to accommodate the type of models we aimed to use to model kinase inhibitor response. We therefore developed a novel, high performance software package, BCM, which provides multi-threaded and high performance implementations of several sampling algorithms, and allows inference with arbitrary models. **Chapter 2** introduces BCM, where we show that, in the test cases we considered, it is more efficient than existing software packages. In this way, BCM made it feasible to perform the inferences presented in the remaining chapters.

Before studying kinase inhibitor response, we made a detour to cell cycle regulation in yeast. This is a well-studied area of biology, where various computational models have already been built, and multiple datasets are available. We used this area to develop the Bayesian framework with which we can integrate multiple datasets to constrain the unknown parameters in a model. In **Chapter 3**, we first illustrate how we iterated over several model versions to find a model that can adequately describe the datasets separately. We then show that by combining multiple datasets, we can reject a specific hypothesis, which could not be done by using any of the datasets separately. This showed that combining datasets can be useful for understanding a biological system, and may therefore be a fruitful approach to studying kinase inhibitor response as well.

With the high performance software package and capability to integrate datasets in hand, we turned to kinase inhibitor response modeling in **Chapter 4**. In this chapter, we describe the extensive characterization of a panel of thirty breast cancer cell lines, and developed a novel integrative analysis method which we call Inference of Signaling Activity (ISA). With this approach, we constructed a model that can describe a large part of the variability in drug response, as well as highlight cases where our knowledge is

insufficient to do so. The model then guided us to identify a novel mechanism involved in mTOR inhibitor sensitivity, namely that overexpression of 4E-BP1 leads to enhanced sensitivity to these inhibitors.

One of the main assumptions in ISA was the absence of feedback signaling events. Although models without explicit feedback were indeed capable of describing a large part of the variability in response, it is known that feedback signaling can play an important role in cellular signaling networks. In **Chapter 5**, we therefore extended ISA to be able to include feedback signaling. We also added an integrated capability for batch correction, which allowed the inclusion of additional datasets from different sources. Using this extended framework, we explored which data is most useful to infer the activity of feedback events, and delineated the most likely feedback activities in the MAPK and AKT pathways, given four different datasets.

The computational cost of Bayesian inference for models with many parameters is typically large, due to the challenges in characterizing a high-dimensional parameter space. At the same time, we would like to include as many details of intracellular signaling as possible, which increases the number of parameters. In **Chapter 6**, we wondered whether it is possible to improve the efficiency of inference with multiple datasets by breaking the inference into separate steps, using one dataset at a time. After comparing several different approximation methods to describe the intermediate posterior distributions, we show that sequential inference can indeed decrease the number of model evaluations that are needed to do the inference, albeit at a cost in precision.

The thesis is concluded with a discussion of the implications of these results and how this research could be extended in order to enable precision medicine in the future.

REFERENCES

[1] E. H. Romond, E. A. Perez, J. Bryant, V. J. Suman, C. E. Geyer, *et al., Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer.* The New England journal of medicine **353**, 1673 (2005).

[2] C. E. Geyer, J. Forster, D. Lindquist, S. Chan, C. G. Romieu, *et al., Lapatinib plus capecitabine for HER2-positive advanced breast cancer.* The New England journal of medicine **355**, 2733 (2006).

[3] N. U. Lin, L. A. Carey, M. C. Liu, J. Younger, S. E. Come, *et al., Phase II trial of lapatinib for brain metastases in patients with human epidermal growth factor receptor 2-positive breast cancer,* Journal of Clinical Oncology **26**, 1993 (2008).

[4] N. U. Lin, V. Diéras, D. Paul, D. Lossignol, C. Christodoulou, *et al., Multicenter phase II study of lapatinib in patients with brain metastases from HER2-positive breast cancer,* Clinical Cancer Research **15**, 1452 (2009).

[5] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, *et al., Systematic identification of genomic markers of drug sensitivity in cancer cells.* Nature **483**, 570 (2012).

[6] F. Iorio, T. A. Knijnenburg, D. J. Vis, J. Saez-Rodriguez, U. Mcdermott, *et al., A Landscape of Pharmacogenomic Interactions in Cancer,* Cell **166**, 740 (2016).

[7] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. a. Margolin, *et al., The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature **483**, 603 (2012).

[8] B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, *et al.*, *Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset,* Cancer Discovery (2015).

[9] L. M. Heiser, A. Sadanandam, W.-L. Kuo, S. C. Benz, T. C. Goldstein, *et al.*, *Subtype and pathway specific responses to anticancer compounds in breast cancer.* Proceedings of the National Academy of Sciences of the United States of America **109**, 2724 (2012).

[10] P. J. A. Eichhorn, M. Gili, M. Scaltriti, V. Serra, M. Guzman, *et al.*, *Phosphatidylinositol 3-kinase hyperactivation results in lapatinib resistance that is reversed by the mTOR/phosphatidylinositol 3-kinase inhibitor NVP-BEZ235.* Cancer research **68**, 9221 (2008).

[11] T. R. Wilson, J. Fridlyand, Y. Yan, E. Penuel, L. Burton, *et al.*, *Widespread potential for growth-factor-driven resistance to anticancer kinase inhibitors.* Nature **487**, 505 (2012).

[12] N. Ilic, T. Utermark, H. R. Widlund, and T. M. Roberts, *PI3K-targeted therapy can be evaded by gene amplification along the MYC-eukaryotic translation initiation factor 4E (eIF4E) axis,* Proceedings of the National Academy of Sciences **108**, E699 (2011).

[13] D. Hanahan and R. Weinberg, *The hallmarks of cancer,* Cell **100**, 57 (2000).

[14] P. Blume-Jensen and T. Hunter, *Oncogenic kinase signalling,* Nature **411**, 355 (2001).

[15] The Cancer Genome Atlas Network, *Comprehensive molecular portraits of human breast tumours.* Nature **490**, 61 (2012).

[16] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, *et al.*, *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.* Nature **486**, 346 (2012).

[17] S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, *et al.*, *Landscape of somatic mutations in 560 breast cancer whole-genome sequences,* Nature **534**, 1 (2016).

[18] E. Rheinbay, P. Parasuraman, J. Grimsby, G. Tiao, J. M. Engreitz, *et al.*, *Recurrent and functional regulatory mutations in breast cancer,* Nature **547**, 55 (2017).

[19] E. Nishida and G. Yukiko, *The MAP kinase cascade is essential for diverse signal transduction pathways,* Trends in Biochemical Science **18**, 128 (1993).

[20] S. R. Datta, A. Brunet, and M. E. Greenberg, *Cellular survival: A play in three akts,* Genes and Development **13**, 2905 (1999).

[21] D. D. Sarbassov, D. A. Guertin, S. M. Ali, and D. M. Sabatini, *Phosphorylation and regulation of Akt/PKB by the rictor-mTOR complex,* Science **307**, 1098 (2005).

[22] P. Liu, W. Gan, Y. R. Chin, K. Ogura, J. Guo, *et al.*, *PtdIns(3,4,5)P3-Dependent Activation of the mTORC2 Kinase Complex,* Cancer Discovery **5**, 1194 (2015).

[23] M. A. Lemmon and J. Schlessinger, *Cell signaling by receptor tyrosine kinases.* Cell **141**, 1117 (2010).

[24] Pearson G, Robinson F, Gibson TB, Xu B-E, Karandikar M, Berman K, and Cobb MH, *Mitogen-activated protein(MAP) Kinase pathways: Regulation and Physiological Functions,* Endocrine Reviews **22**, 153 (2001).

[25] M. Shimobayashi and M. N. Hall, *Making new contacts: The mTOR network in metabolism and signalling crosstalk,* Nature Reviews Molecular Cell Biology **15**, 155 (2014).

[26] R. Zoncu, A. Efeyan, and D. M. Sabatini, *mTOR: from growth signal integration to cancer, diabetes and ageing.* Nature reviews. Molecular cell biology **12**, 21 (2011).

[27] D. A. Yardley, S. Noguchi, K. I. Pritchard, H. A. Burris, J. Baselga, *et al.*, *Everolimus plus exemes-*

tane in postmenopausal patients with HR+ breast cancer: BOLERO-2 final progression-free survival analysis, Advances in Therapy **30**, 870 (2013).

[28] M. Piccart, G. Hortobagyi, M. Campone, K. Pritchard, S. Noguchi5, *et al., Everolimus plus exemestane for hormone receptor-positive (HR+), human epidermal growth factor receptor-2-negative (HER2-) advanced breast cancer (BC): overall survival results from BOLERO-2,* Annals of Oncology **25**, 2357 (2014).

[29] F. André, R. O'Regan, M. Ozguroglu, M. Toi, B. Xu, *et al., Everolimus for women with trastuzumab-resistant, HER2-positive, advanced breast cancer (BOLERO-3): A randomised, double-blind, placebo-controlled phase 3 trial,* The Lancet Oncology **15**, 580 (2014).

[30] S. A. Hurvitz, F. Andre, Z. Jiang, Z. Shao, M. S. Mano, *et al., Combination of everolimus with trastuzumab plus paclitaxel as first-line treatment for patients with HER2-positive advanced breast cancer (BOLERO-1): A phase 3, randomised, double-blind, multicentre trial,* The Lancet Oncology **16**, 816 (2015).

[31] K. E. O'Reilly, F. Rojo, Q. B. She, D. Solit, G. B. Mills, *et al., mTOR inhibition induces upstream receptor tyrosine kinase signaling and activates Akt,* Cancer Research **66**, 1500 (2006).

[32] D. Benjamin, M. Colombi, C. Moroni, and M. N. Hall, *Rapamycin passes the torch: a new generation of mTOR inhibitors.* Nature reviews. Drug discovery **10**, 868 (2011).

[33] B. Basu, E. Dean, M. Puglisi, A. Greystoke, M. Ong, *et al., First-in-human pharmacokinetic and pharmacodynamic study of the dual m-TORC 1/2 inhibitor AZD2014,* Clinical Cancer Research **21**, 3412 (2015).

[34] J. Xie, X. Wang, and C. G. Proud, *mTOR inhibitors in cancer therapy,* F1000Research **5**, 2078 (2016).

[35] P. Vuylsteke, M. Huizing, K. Petrakova, R. Roylance, R. Laing, *et al., Pictilisib PI3Kinase inhibitor (a phosphatidylinositol 3-kinase [PI3K] inhibitor) plus paclitaxel for the treatment of hormone receptor-positive, HER2-negative, locally recurrent, or metastatic breast cancer: interim analysis of the multicentre, placebo-controlled, phase II randomised PEGGY study,* Annals of Oncology **27**, 2059 (2016).

[36] I. E. Krop, I. A. Mayer, V. Ganju, M. Dickler, S. Johnston, *et al., Pictilisib for oestrogen receptor-positive, aromatase inhibitor-resistant, advanced or metastatic breast cancer (FERGI): a randomised, double-blind, placebo-controlled, phase 2 trial,* The Lancet Oncology **17**, 811 (2016).

[37] A. Daemen, O. L. Griffith, L. M. Heiser, N. J. Wang, O. M. Enache, *et al., Modeling precision treatment of breast cancer,* Genome Biology **14** (2013).

[38] J. Saez-Rodriguez, L. Alexopoulos, M. Zhang, M. K. Morris, D. A. Lauffenburger, and P. K. Sorger, *Comparing signaling networks between normal and transformed hepatocytes using discrete logical models,* Cancer research **71**, 5400 (2011).

[39] J. Nan, Y. Du, X. Chen, Q. Bai, Y. Wang, *et al., TPCA-1 Is a Direct Dual Inhibitor of STAT3 and NF- B and Regresses Mutant EGFR-Associated Human Non-Small Cell Lung Cancers,* Molecular Cancer Therapeutics **13**, 617 (2014).

[40] B. Klinger, A. Sieber, R. Fritsche-Guenther, F. Witzel, L. Berry, *et al., Network quantification of EGFR signaling unveils potential for targeted combination therapy,* Molecular Systems Biology **9** (2013).

[41] B. N. Kholodenko, A. Kiyatkin, F. J. Bruggeman, E. Sontag, H. V. Westerhoff, and J. B. Hoek, *Untangling the wires: A strategy to trace functional interactions in signaling and gene networks,*

Proceedings of the National Academy of Sciences **99**, 12841 (2002).

[42] I. Stelniec-Klotz, S. Legewie, O. Tchernitsa, F. Witzel, B. Klinger, *et al.*, *Reverse engineering a hierarchical regulatory network downstream of oncogenic KRAS.* Molecular systems biology **8**, 601 (2012).

[43] A. Prahallad, C. Sun, S. Huang, F. Di Nicolantonio, R. Salazar, *et al.*, *Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR.* Nature **483**, 100 (2012).

[44] D. C. Kirouac, J. Y. Du, J. Lahdenranta, R. Overland, D. Yarar, *et al.*, *Computational modeling of ERBB2-amplified breast cancer identifies combined ErbB2/3 blockade as superior to the combination of MEK and AKT inhibitors.* Science signaling **6**, ra68 (2013).

[45] F. Eduati, V. Doldàn-Martelli, B. Klinger, T. Cokelaer, A. Sieber, *et al.*, *Drug resistance mechanisms in colorectal cancer dissected with cell type-specific dynamic logic models,* Cancer Research **77**, 3364 (2017).

[46] S. Marino, I. B. Hogue, C. J. Ray, and D. E. Kirschner, *A methodology for performing global uncertainty and sensitivity analysis in systems biology,* Journal of Theoretical Biology **254**, 178 (2008).

[47] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, *Universally sloppy parameter sensitivities in systems biology models,* PLoS Computational Biology **3**, 1871 (2007), 0701039 .

[48] J. Vanlier, C. Tiemann, P. Hilbers, and N. van Riel, *Parameter uncertainty in biochemical models described by ordinary differential equations.* Mathematical biosciences **246**, 305 (2013).

[49] V. Serra, B. Markman, M. Scaltriti, P. J. a. Eichhorn, V. Valero, *et al.*, *NVP-BEZ235, a dual PI3K/mTOR inhibitor, prevents PI3K signaling and inhibits the growth of cancer cells with activating PI3K mutations.* Cancer research **68**, 8022 (2008).

[50] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller, *Equation of state calculations by fast computing machines,* The journal of chemical physics **21**, 1087 (1953).

[51] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications,* Biometrika **57**, 97 (1970).

[52] H. Haario, E. Saksman, and J. Tamminen, *An Adaptive Metropolis Algorithm,* Bernoulli **7**, 223 (2001).

[53] C. J. Geyer, *Markov Chain Monte Carlo Maximum Likelihood,* in *Proceedings of the 23rd Symposium Interface*, 1 (1991) pp. 156–163.

[54] H. G. Katzgraber, S. Trebst, D. A. Huse, and M. Troyer, *Feedback-optimized parallel tempering Monte Carlo,* Journal of Statistical Mechanics: Theory and Experiment **2006**, P03018 (2006), 0602085v3 .

[55] P. Del Moral, A. Doucet, and A. Jasra, *Sequential Monte Carlo samplers,* Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**, 411 (2006).

[56] P. Del Moral, A. Doucet, and A. Jasra, *An adaptive sequential Monte Carlo method for approximate Bayesian computation,* Statistics and Computing **22**, 1009 (2011).

[57] F. Feroz, M. P. Hobson, and M. Bridges, *MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics,* Monthly Notices of the Royal Astronomical Society **398**, 1601 (2009).

2

BCM: TOOLKIT FOR BAYESIAN ANALYSIS OF COMPUTATIONAL MODELS USING SAMPLERS

Bram THIJSSEN
Tjeerd M.H. DIJKSTRA
Tom HESKES
Lodewyk F.A. WESSELS

Parts of this chapter have been published in BMC Systems Biology 10:100 (2016).

ABSTRACT

COMPUTATIONAL models in biology are characterized by a large degree of uncertainty. This uncertainty can be analyzed with Bayesian statistics, however, the sampling algorithms that are frequently used for calculating Bayesian statistical estimates are computationally demanding, and each algorithm has unique advantages and disadvantages. It is typically unclear, before starting an analysis, which algorithm will perform well on a given computational model. Here, we present BCM, a toolkit for the Bayesian analysis of Computational Models using samplers. It provides efficient, multithreaded implementations of eleven algorithms for sampling from posterior probability distributions and for calculating marginal likelihoods. BCM includes tools to simplify the process of model specification and scripts for visualizing the results. The flexible architecture allows it to be used on diverse types of biological computational models. In an example inference task using a model of the cell cycle based on ordinary differential equations, BCM is significantly more efficient than existing software packages, allowing more challenging inference problems to be solved. BCM represents an efficient one-stop-shop for computational modelers wishing to use sampler-based Bayesian statistics.

2.1. BACKGROUND

There is an increasing interest in using Bayesian statistics for the analysis of computational models in biology [1–4]. With Bayesian statistics, the unknown parameters of a computational model are assigned a probability distribution describing their uncertainty. This distribution can be updated from prior information to give the posterior probability distribution, using Bayes' theorem:

$$P(\theta|X,\mathscr{M}) = \frac{P(X|\theta,\mathscr{M})P(\theta|\mathscr{M})}{P(X|\mathscr{M})} \tag{2.1}$$

where θ represents the parameters, X the measurement data and $\mathscr{M}$ the computational model. Furthermore, the marginal likelihood, or evidence, can be used todiscriminate between different computational models. It can be calculated by marginalizing the parameters:

$$P(X|\mathscr{M}) = \int P(X|\theta,\mathscr{M})P(\theta|\mathscr{M})d\theta \tag{2.2}$$

Typically, neither the posterior probability nor the marginal likelihood can be calculated directly, but sampling algorithms can be used to estimate them [5–16]. These sampling algorithms are computationally demanding, especially when the number of parameters is large and when the computational model is expensive to simulate. Typical models in systems biology indeed carry many parameters and are expensive to simulate [17] Additionally, the posterior probability distributions arising from such models are usually complex, containing multiple modes and ridges that are difficult to traverse [18]. Bayesian analysis of such systems biology models thus requires the use of advanced sampling algorithms. Since these sampling algorithms each have unique characteristics and can be more or less suitable for a particular task, it would be beneficial to have various algorithm easily available.

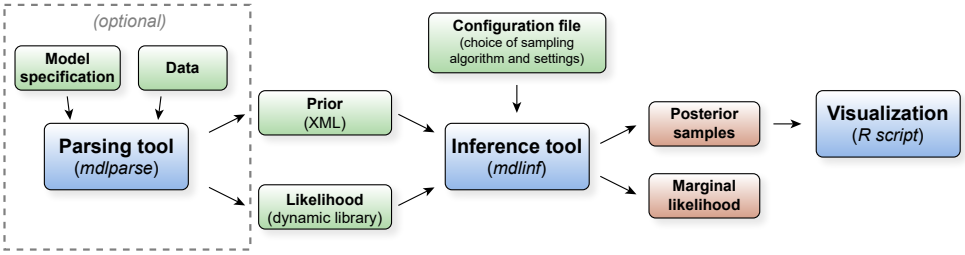


Figure 2.1: **Overview of BCM.** The inference tool is the main component of BCM, providing three classes of algorithms for generating samples from posterior probability distributions and calculating estimates of the marginal likelihood. The parsing tool can optionally be used to generate the prior and likelihood files from a model description file and data.

BCM, a toolkit for the Bayesian analysis of Computational Models using samplers, provides efficient, multithreaded implementations of eleven algorithms for calculating posterior probabilities and marginal likelihoods.

The BCM toolkit focuses on computational models that involve simulations or extensive calculations. Examples of such computational models are systems of ordinary differential equations describing biochemical reactions; or steady-state signaling networks, where the activity levels may be calculated in diverse ways. These computational models are in contrast to statistical models that can be specified in the BUGS or Stan languages. For such statistical models, excellent software packages already exist [19, 20]. For the computational models that are targeted by BCM, several alternative software packages also exist [16, 21–23]. However, each of these packages implements only a single type of sampling algorithm and most of them focus on one particular type of computational model. In contrast, with BCM the user can choose from eleven sampling algorithms and the plugin architecture allows diverse types of models. Thus, BCM represents a one-stopshop for Bayesian analysis of systems biology models, where the user has a high chance of finding a suitable algorithm for the analysis of the user-defined model.

2.2. IMPLEMENTATION

BCM consists of three components: an inference tool, a model parsing tool and an R script for further analysis and visualization (see Figure 2.1).

The inference tool (*mdlinf*) is the main component of BCM. It uses a specified sampling algorithm to generate samples from the posterior probability distribution and to calculate a marginal likelihood estimate. Error bounds for the marginal likelihood estimate are also provided, which are calculated directly from the samples using a method suitable for the particular algorithm used to calculate the marginal likelihood. As input, the inference tool requires three parts: a configuration file, an XML file specifying the prior, and a dynamic library that evaluates the likelihood function. For constructing the dynamic library that evaluates the likelihood function, BCM provides cross-platform boilerplate code, such that custom model simulation code can be easily adapted for use with BCM. Alternatively, the model parsing tool can be used as described further below.

The inference tool implements three classes of sampling algorithms: Markov chain

Monte Carlo (MCMC) [5, 6], sequential Monte Carlo (SMC) [7] and nested sampling [8]. For each class of sampling algorithms, BCM includes several options for proposal distributions, as well as extensions that can increase the sampling efficiency when dealing with complex inference problems, giving a total of eleven different sampling algorithms (Table 2.1).

Care has been taken to create efficient, multithreaded implementations of each algorithm. Firstly, the inference tool has been written in C++ and performance bottlenecks have been profiled and optimized. Secondly, each algorithm has been parallelized with a multithreading strategy suitable for that algorithm: for MCMC, multiple chains are distributed across threads, for SMC, particles are distributed in batches across threads, and for nested sampling, a batch of samples is generated at each iteration by all threads which are then re-used in subsequent nested sampling iterations.

The model parsing tool (*mdlparse*) is the second component of BCM. It can be used to generate the prior and likelihood files for the inference tool. The parsing tool reads a model description file that specifies the model, comprising the prior, likelihood and data references, and it outputs C++ source code for a dynamic library that evaluates the prior and likelihood function with the relevant data. This C++ code can then be used as a basis for further modification; or it can be directly compiled into a dynamic library. The input model description file uses a custom format with an easy-to-read syntax. An excerpt of a model description file is shown in Fig. 2.2. The use of the model parsing tool is optional and it is meant as an aid in model specification rather than as a comprehensive tool capable of fully specifying all types of models.

Finally, a script is provided to load the output of the inference tool into R for further analysis and for visualization of the results. This script can be used to display kernel density estimates of the posterior probability distribution of the sampled variables, as well as to make plots for visual posterior predictive checking; examples of both of these are shown in 2.4, 2.5 and 2.6. Basic functionality for convergence diagnostics is included as well, including autocorrelation functions and trace plots. Functions for conversion of the results to CODA objects [24] and to ggmcmc objects [25], two R packages for MCMC convergence diagnostics and output analysis, are also provided.

Sampling algorithm	Reference
Markov Chain Monte Carlo	[5, 6]
Parallel tempering	[9]
Adaptive proposals	[10]
Feedback-optimized temperatures	[11]
Thermodynamic integration	[12]
Automated parameter blocking	[13]
Sequential Monte Carlo	[7]
MCMC proposals	[7]
Kernel density estimate proposals	[7]
Automated temperature schedule	[14]
Nested sampling	[8]
MCMC proposals	[8]
Ellipsoid proposals	[15]
MultiNest	[16]

Table 2.1: **Sampling algorithms and extensions implemented in BCM.**

```
[Data]
some_data = "path/data.tsv"

[Variables]
initial_protein_activity = real[3] => { "lacI", "tetR", "cI" }
protein_activity         = real[3] => { "lacI", "tetR", "cI" }
a_kinetic_parameter      = real
data_sigma               = real

[Priors]
initial_protein_activity = uniform(lower=0.0, upper=1.0)
a_kinetic_parameter      = normal(mu=2, sigma=4)
data_sigma               = half_cauchy(scale=0.1)

[Likelihood]
protein_activity = Simulate(initial_protein_activity, a_kinetic_parameter)
logp = 0
for i = 0 to 4
    logp += studentt(data->some_data[i,"Column name"] |
        mu=protein_activity["lacI"], sigma=data_sigma, nu=3)
end for
```

Figure 2.2: **Excerpt of a model description file.** The model parsing tool can parse this file, load the relevant data, and output C++ source code for a dynamic library that evaluates the likelihood function. In this example, the "Simulate()" function still has to be implemented by the user with a desired simulation method.

2.3. RESULTS

2.3.1. ANALYTICALLY TRACTABLE EXAMPLE

To showcase BCM, and to explore how each class of algorithms deals with increasing dimensionality and complex distributions, we first analyzed a problem which is analytically tractable: the Gaussian shells problem described in [16, 26]. While this example is not directly relevant for systems biology, its likelihood function is multimodal and ridge-shaped, resembling the likelihoods often encountered in systems biology models. The likelihood function for this Gaussian shells problem is given by

$$P(\boldsymbol{\theta}) = \sum_{i=1}^{2} \frac{1}{\sqrt{2\pi w^2}} \exp\left(-\frac{(|\boldsymbol{\theta} - \mathbf{c_i}| - r)^2}{2w^2}\right), \tag{2.3}$$

where $r = 2$, $w = 0.1$, and $\boldsymbol{\theta}$ and $\mathbf{c_i}$ are n-dimensional vectors. $\boldsymbol{\theta}$ is the vector of variables which are to be sampled and $\mathbf{c_i}$ are two constant vectors describing the centers of the two peaks and are assigned the values $c_{1,x} = 3.5$, $c_{2,x} = -3.5$, and 0 in the other dimensions. This likelihood function is then composed of two narrow, well-separated, ring-shaped peaks (Figure 2.3A), which is a challenging sampling problem.

We tested three sampling algorithms on this problem, one from each class of sampling algorithms: feedback-optimized parallel-tempered Markov chain Monte Carlo (FO-PTMC) [11], sequential Monte Carlo (SMC) [7] with the automated temperature schedule selection of [14] but without using Approximate Bayesian Computation, and MultiNest [16].

As shown in Table 2.2, all three algorithms give the correct estimate for the marginal likelihood within the error bounds. When the number of dimensions is 10 or fewer, MultiNest is extremely efficient: it requires the fewest likelihood evaluations while achieving the tightest error bound. When the number of dimensions is increased beyond 10

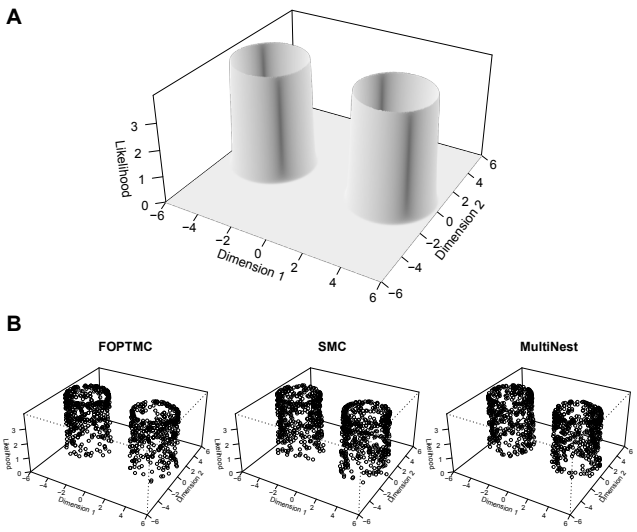


Figure 2.3: **Gaussian shells example.** (A) Likelihood of the Gaussian shells problem in the 2-dimensional case. (B) Samples generated from the likelihood by three sampling algorithms. In each case, the samples are well-distributed throughout each mode, and the two modes are sampled in approximately equal proportions.

however, MultiNest becomes very inefficient. At this point the exponential scaling of the algorithm becomes apparent. In the higher-dimensional setting, the SMC algorithm deals with this problem most efficiently. FOPTMC is least efficient: it requires the largest number of likelihood evaluations and has the largest error bound. FOPTMC can still effectively explore the posterior distribution (as shown in Figure 2.3B), however, the temperature schedule of the parallel chains in FOPTMC is optimized for exploration of the posterior rather than for estimation of the marginal likelihood and as a result there is an increasingly large error in the marginal likelihood estimate at higher dimensionality.

Dim.	Log marginal likelihood				Likelihood evaluations (x1000)		
	Analytical	FOPTMC	SMC	MultiNest	FOPTMC	SMC	MultiNest
2	-1.75	-1.80 ± 0.68	-1.74 ± 0.39	-1.73 ± 0.29	147	79	18
5	-5.67	-5.98 ± 1.65	-5.66 ± 0.47	-5.73 ± 0.38	287	281	28
10	-14.59	-14.92 ± 3.34	-14.64 ± 0.62	-14.13 ± 0.63	969	521	95
30	-60.13	-61.11 ± 9.10	-59.85 ± 0.97	*	6420	1511	*
100	-255.62	-257.7 ± 24.8	-255.8 ± 1.54	*	96,251	4271	*

Table 2.2: **Performance of three sampling algorithms in calculating marginal likelihoods.** The following algorithms were used: FOPTMC feedback-optimized parallel-tempered Markov Chain Monte Carlo [11], SMC automated-temperature sequential Monte Carlo but without ABC approximation [14], and MultiNest [16]. The column 'Analytical' gives the marginal likelihood value calculated analytically. (*) indicates that the computation time exceeded the maximal time of 1 h; the other calculations required at most 5 min.

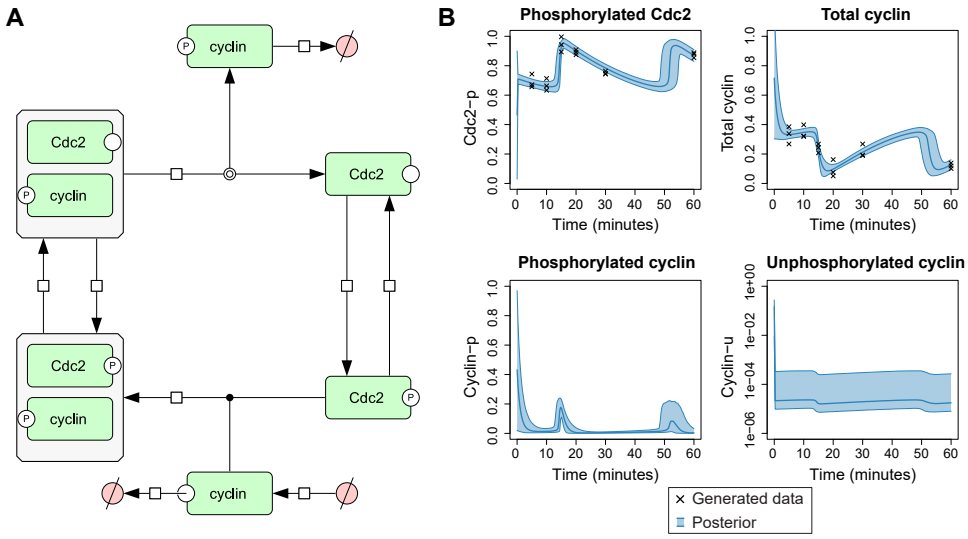


Figure 2.4: **ODE-based model of cell cycle regulation.** (A) Graphical representation of the cell cycle model of Tyson [27]. b Posterior distribution of the two observables; phosphorylated Cdc2 and the total amount of cyclin, and of two unobserved species, phosphorylated and unphosphorylated cyclin. The black crosses represent the generated data which are used for the inference. The shaded blue area represents the posterior 95% confidence interval of the mean of the observables.

2.3.2. KINETIC ORDINARY DIFFERENTIAL EQUATION MODEL

Having explored the behavior of several sampling algorithms in an analytically tractable example, we now illustrate the use of BCM for analyzing biological computational models. As an example of this, we investigated the inference of the parameters of a model based on a system of ordinary differential equations (ODEs). The 6-variable cell cycle model of Tyson [27] was used, as downloaded from BioModels [17]. A graphical representation of this model is shown in Figure 2.4A.

To recreate a typical setting in biology, data was generated from the model at six time points for two observables with three replicates (see Additional file 1). Then BCM was used to infer all 16 parameters of the model (10 kinetic parameters and 6 initial conditions) from these 36 data points. The priors for the kinetic parameters were set to a uniform distribution spanning an order of magnitude on either side of the parameter values that were used to generate the data, and the priors for the initial conditions were set to a uniform distribution between 0 and 1 (see blue curves in Figure 2.5C). The likelihood function was set equal to the one that generated the data, that is, a normal distribution with standard deviation 0.05.

Despite the small size of the model, this inference problem is challenging. Firstly, the ODE system is stiff, and even with the use of an implicit ODE solver it is costly to simulate. Secondly, there are multiple distinct ways in which the model can fit the data, leading to sub-optimal modes in the posterior distribution. Thus, a sampler must be able to escape these local optima, and it must be able to converge to the correct posterior distribution with a limited number of likelihood evaluations due to the computational

cost of the simulations.

Four sampling algorithms were tested on this problem: SMC, MultiNest, FOPTMC (now extended with automated parameter blocking [13]), and additionally nested sampling with MCMC proposals (Nested-MCMC) was added as an alternative nested sampling strategy. In this inference task, FOPTMC with automated parameter blocking was most efficient, requiring 14 h to generate 2000 samples from the posterior. SMC required 19 h, while Nested-MCMC required 30 h and MultiNest had to be discontinued as the acceptance rate quickly dropped to essentially zero. The tests were performed using 16 threads on an Intel Xeon E5-2680 processor.

The Bayesian estimates of the parameters and the trajectories of the species can be used to study the uncertainty in the model. Figure 2.4B shows the posterior distribution of the two observables, as well as of two inferred species for which no observable data was generated, as estimated by FOPTMC. We can see that the data are sufficient to constrain the trajectories of the observed species. For the unobserved species phosphorylated cyclin, the overall trajectory can also be inferred. Nevertheless, for this unobserved species, the second peak is more variable – here the data is insufficient to constrain the precise magnitude of the peak. For the other unobserved species, unphosphorylated cyclin, we see that there is greater uncertainty. The posterior distribution indicates only that the average levels are low, but the precise levels cannot be inferred from these data.

Figure 2.5C shows the marginal posterior probability distributions of the parameters. It can be seen that for all parameters, the values used to generate the data fall within areas of non-zero probability of the posterior. In most cases the data-generation values also have maximum posterior probability, but interestingly this is not true for all parameters, such as for the activation and deactivation of Cdc2. Furthermore, some parameters are not identifiable, for example the rates of phosphorylation and desphosphorylation of Cdc2 cannot be determined from the data. In general, such lack of identifiability could be for structural reasons, that is, the parameters cannot be inferred in theory given the observed species, due to a redundant parameterization. Alternatively, the parameters may be identifiable in theory, but the data may provide insufficient information to constrain the parameters in practice.

Overall, the Bayesian estimates provide useful measures of the uncertainty in parameter values, model fit and model predictions.

2.3.3. COMPARISON WITH EXISTING SOFTWARE PACKAGES

There are several software packages which can perform Bayesian inference of the parameters of ODE-based models: BioBayes [21], ABC-SysBio [22], SYSBIONS [23] and Stan [20]. BioBayes uses parallel-tempered Markov Chain Monte Carlo, ABC-SysBio uses sequential Monte Carlo sampling in combination with Approximate Bayesian Computation, SYSBIONS uses nested sampling, and Stan uses Hamiltonian Monte Carlo and the No-U-Turn sampler (NUTS).

To compare BCM with these software packages, a simplified version of the previous inference problem was used. Instead of inferring all 16 parameters, the initial conditions and 4 of the 10 kinetic parameters were fixed to the values used to generate the data, leaving 6 parameters to be inferred. Figure 2.6A shows the marginal posterior probability distributions of the simplified problem, as estimated by BCM using FOPTMC (see

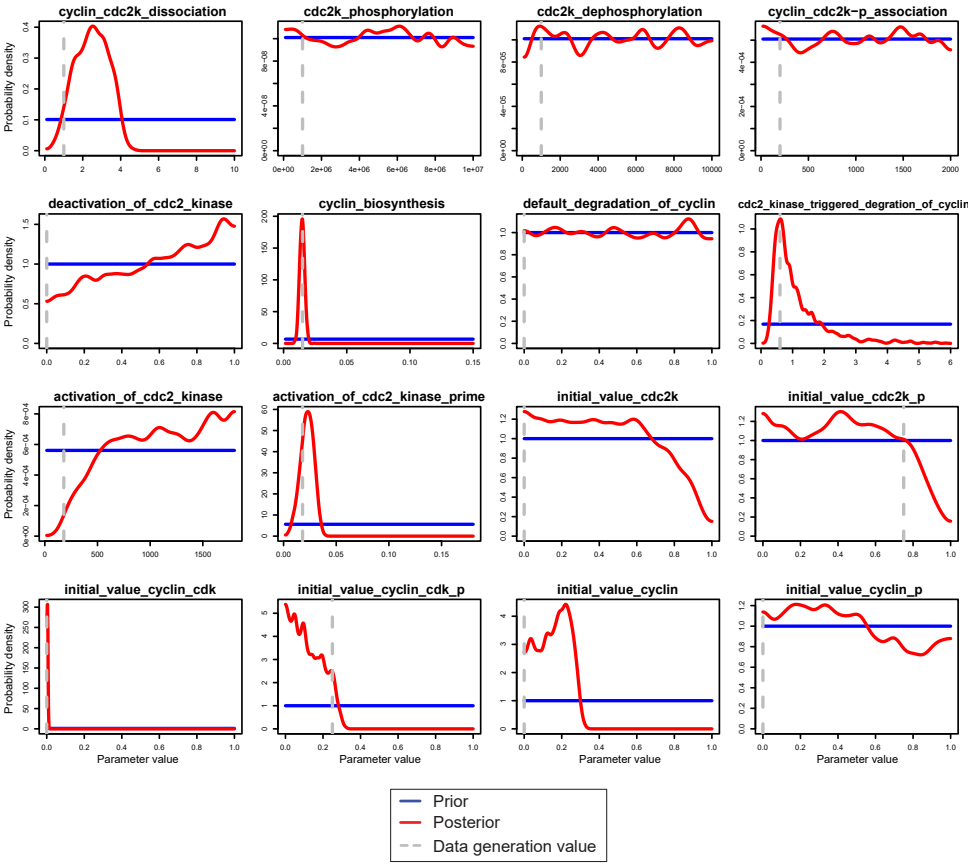


Figure 2.5: **Marginal posterior distributions of the cell cycle model parameters.** The blue lines indicate the prior, the red lines the estimated posterior, and the dashed grey lines indicate the values that were used to generate the data. The densities are estimated from the posterior samples using kernel density estimation with Sheather-Jones bandwidth selection.

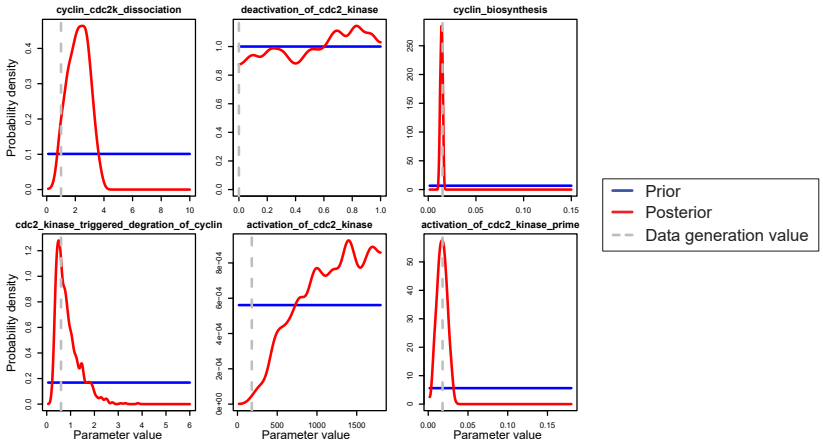


Figure 2.6: **Marginal posterior distributions of the cell cycle model parameters.** Prior and posterior probability distributions of the 6 parameters of the simplified inference problem.

Additional file 2: Figure S1 for the posteriors estimated by each algorithm/software package). The other software packages were optimized for this problem as much as possible to give a fair comparison (see Additional file 1).

Figure 2.7B shows the time required to generate 1000 samples from the posterior with each software package and algorithm, using eight threads on an Intel Xeon E5-2680 processor. It is clear that BCM is significantly faster than the other software packages. In particular the MultiNest algorithm in BCM is extremely efficient in this low-dimensional setting, requiring only 75 s. The other algorithms in BCM required between 25 and 50 min, except for ellipsoidal nested sampling which required three hours. From the other software packages, only SYSBIONS and Stan were able to solve this inference problem in a reasonable amount of time. SYSBIONS required five hours using Nested-MCMC, which is approximately six times longer than BCM with the same algorithm. For Stan, using the NUTS algorithm, the sampling with a chain does not always converge as the NUTS algorithm does not have a means to escape sub-optimal modes. This problem was addressed by starting eight separate chains in parallel, in which case most of the chains were sampling the correct, optimal mode. In this case, Stan required approximately six hours to generate the requested samples. BioBayes was able to reach apparent convergence in 4.5 days. For ABC-SysBio, and SYSBIONS using ellipsoidal sampling, the samplers did not reach convergence in 7 days (see Additional file 1).

2.4. CONCLUSION

The BCM toolkit provides efficient, multithreaded implementations of eleven sampling algorithms for generating posterior samples and calculating marginal likelihoods. Additional tools are included which facilitate the process of specifying models and visualizing the sampling output. This toolkit can be used for analyzing the uncertainty in the parameters and the predictions of computational models using Bayesian statistics.

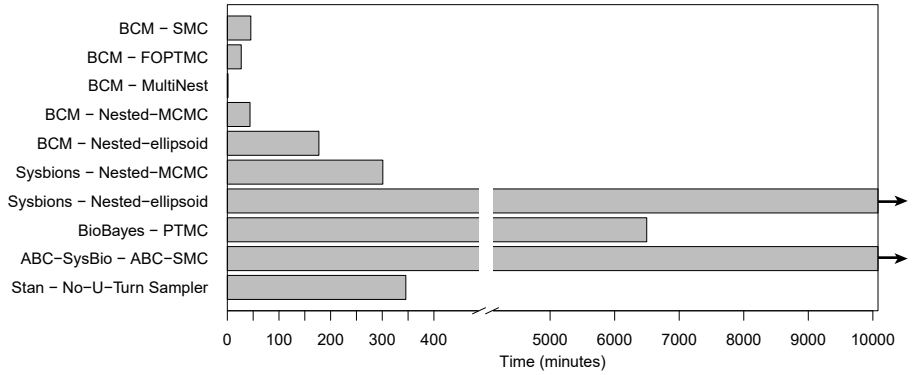


Figure 2.7: **Performance comparison of software packages.** Bars indicate the time required for generating 1,000 samples from the posterior using BCM, SYSBIONS, BioBayes, ABC-SysBio and Stan, with several different sampling algorithm. The sampling was terminated if it had not converged after 7 days.

The examples show that it depends on the problem which sampling algorithm will perform well. In the Gaussian shells example, where the focus was on marginal likelihood estimation, MultiNest performed best in a low-dimensional setting, and in the medium- to high dimensional setting sequential Monte Carlo was most efficient. In the cell cycle example, where the focus was on parameter inference, parallel-tempered Markov chain Monte Carlo was more efficient than sequential Monte Carlo. There are various aspects of the posterior probability distribution which affect the performance of the different algorithms; for example the number of modes, how well the shapes of the modes are approximated by the proposal distributions, and the location and volume of the posterior modes with respect to the prior. These features of the posterior probability distribution will typically not be known for the problem of interest before starting the analysis, and it is then unclear which algorithm might be most suitable. The availability of various algorithms in BCM will therefore be useful in the Bayesian analysis of diverse models.

In the second example, we have shown that BCM can be used to infer the parameters of an ODE-based model of the cell cycle. BCM is significantly more efficient in this task than existing software packages. This increase in efficiency was possible due to the parallelization of the sampling algorithms in combination with the use of optimized C++ as programming language. Due to the higher efficiency, BCM allows the analysis of larger or more challenging computational models than was previously feasible. In previous cases where Bayesian analysis of complex biological computational models was done, such as in [3, 4, 28], sampling algorithms were newly implemented for each project. The availability of BCM as an efficient, reusable software package can help in streamlining such projects in the future.

AVAILABILITY AND REQUIREMENTS

Project name: BCM – toolkit for Bayesian analysis of Computational Models using samplers

Project home page: http://ccb.nki.nl/software/bcm/

Operating systems: Windows, Linux, Mac

Programming language: C++ and R

Dependencies: Boost C++ libraries (tested with version 1.55.0), CMake (version 3.2 or later).

License: Mozilla Public License 2.0

REFERENCES

[1] D. J. Wilkinson, *Bayesian methods in bioinformatics and computational systems biology.* Briefings in bioinformatics **8**, 109 (2007).

[2] V. Vyshemirsky and M. Girolami, *Bayesian ranking of biochemical system models.* Bioinformatics (Oxford, England) **24**, 833 (2008).

[3] T.-R. Xu, V. Vyshemirsky, A. Gormand, A. von Kriegsheim, M. Girolami, *et al.*, *Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species.* Science signaling **3**, ra20 (2010).

[4] H. Eydgahi, W. W. Chen, J. L. Muhlich, D. Vitkup, J. N. Tsitsiklis, and P. K. Sorger, *Properties of cell death models calibrated and compared using Bayesian approaches.* Molecular systems biology **9**, 644 (2013).

[5] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller, *Equation of state calculations by fast computing machines,* The journal of chemical physics **21**, 1087 (1953).

[6] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications,* Biometrika **57**, 97 (1970).

[7] P. Del Moral, A. Doucet, and A. Jasra, *Sequential Monte Carlo samplers,* Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**, 411 (2006).

[8] J. Skilling, *Nested sampling for general Bayesian computation,* Bayesian Analysis **1**, 833 (2006).

[9] C. J. Geyer, *Markov Chain Monte Carlo Maximum Likelihood,* in *Proceedings of the 23rd Symposium Interface*, 1 (1991) pp. 156–163.

[10] H. Haario, E. Saksman, and J. Tamminen, *An Adaptive Metropolis Algorithm,* Bernoulli **7**, 223 (2001).

[11] H. G. Katzgraber, S. Trebst, D. A. Huse, and M. Troyer, *Feedback-optimized parallel tempering Monte Carlo,* Journal of Statistical Mechanics: Theory and Experiment **2006**, P03018 (2006), 0602085v3 .

[12] A. Gelman and X.-L. Meng, *Simulating normalizing constants: from importance sampling to bridge sampling to path sampling,* Statistical Science **13**, 163 (1998).

[13] D. Turek, P. de Valpine, C. J. Paciorek, and C. Anderson-Bergman, *Automated Parameter Blocking for Efficient Markov-Chain Monte Carlo Sampling,* Bayesian Analysis **12**, 465 (2017), 1503.05621 .

[14] P. Del Moral, A. Doucet, and A. Jasra, *An adaptive sequential Monte Carlo method for approximate Bayesian computation,* Statistics and Computing **22**, 1009 (2011).

[15] P. Mukherjee, D. Parkinson, and A. R. Liddle, *A Nested Sampling Algorithm for Cosmological Model Selection,* The Astrophysical Journal **638**, 51 (2006).

[16] F. Feroz, M. P. Hobson, and M. Bridges, *MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics,* Monthly Notices of the Royal Astronomical Society **398**, 1601 (2009).

[17] V. Chelliah, N. Juty, I. Ajmera, R. Ali, M. Dumousseau, *et al.*, *BioModels: Ten-year anniversary,* Nucleic Acids Research **43**, D542 (2015).

[18] M. Girolami, *Bayesian inference for differential equations,* Theoretical Computer Science **408**, 4 (2008).

[19] D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best, *The BUGS project: Evolution, critique and future directions,* Statistics in medicine **28**, 3049 (2009).

[20] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, *et al.*, *Stan: A Probabilistic Programming Language,* Journal of Statistical Software **76** (2017).

[21] V. Vyshemirsky and M. Girolami, *BioBayes: A software package for Bayesian inference in systems biology,* Bioinformatics **24**, 1933 (2008).

[22] J. Liepe, C. Barnes, E. Cule, K. Erguler, P. Kirk, T. Toni, and M. P. H. Stumpf, *ABC-SysBio—approximate bayesian computation in python with GPU support,* Bioinformatics **26**, 1797 (2010).

[23] R. Johnson, P. Kirk, and M. P. H. Stumpf, *SYSBIONS: nested sampling for systems biology,* Bioinformatics **31**, 604 (2014).

[24] M. Plummer, N. Best, K. Cowles, and K. Vines, *CODA: Convergence Diagnosis and Output Analysis for MCMC,* R News **6**, 7 (2006).

[25] X. Fernández-i Marín, *ggmcmc: Analysis of MCMC Samples and Bayesian Inference,* Journal of Statistical Software **70** (2016).

[26] B. C. Allanach and C. G. Lester, *Sampling using a 'bank' of clues,* Computer Physics Communications **179**, 256 (2008), 0705.0486 .

[27] J. J. Tyson, *Modeling the cell division cycle: cdc2 and cyclin interactions.* Proceedings of the National Academy of Sciences of the United States of America **88**, 7328 (1991).

[28] A. Milias-Argeitis, A. P. Oliveira, L. Gerosa, L. Falter, U. Sauer, and J. Lygeros, *Elucidation of Genetic Interactions in the Yeast GATA-Factor Network Using Bayesian Model Selection,* PLoS Comput Biol **12**, e1004784 (2016).

SUPPLEMENTARY MATERIAL

Supplementary Material is available online at
https://doi.org/10.1186/s12918-016-0339-3

3

BAYESIAN DATA INTEGRATION FOR QUANTIFYING THE CONTRIBUTION OF DIVERSE MEASUREMENTS TO PARAMETER ESTIMATES

Bram THIJSSEN
Tjeerd M.H. DIJKSTRA
Tom HESKES
Lodewyk F.A. WESSELS

Parts of this chapter have been published in Bioinformatics 34:803-811 (2017).

ABSTRACT

COMPUTATIONAL models in biology are frequently underdetermined, due to limits in our capacity to measure biological systems. In particular, mechanistic models often contain parameters whose values are not constrained by a single type of measurement. It may be possible to achieve better model determination by combining the information contained in different types of measurements. Bayesian statistics provides a convenient framework for this, allowing a quantification of the reduction in uncertainty with each additional measurement type. We wished to explore whether such integration is feasible and whether it can allow computational models to be more accurately determined. To this end, we created an ODE model of cell cycle regulation in budding yeast, and integrated data from thirteen different studies covering different experimental techniques. We found that for some parameters, a single type of measurement, relative time course mRNA expression, is sufficient to constrain them. Other parameters, however, were only constrained when two types of measurements were combined, namely relative time course and absolute transcript concentration. Comparing the estimates to measurements from three additional, independent studies, we found that the degradation and transcription rates indeed matched the model predictions in order of magnitude. The predicted translation rate was incorrect however, thus revealing a deficiency in the model. Since this parameter was not constrained by any of the measurement types separately, it was only possible to falsify the model when integrating multiple types of measurements. In conclusion, this study shows that integrating multiple measurement types can allow models to be more accurately determined.

3.1. INTRODUCTION

Computational models in biology are frequently underdetermined [1], which can limit their usefulness. This underdetermination is a result of our limited capacity to measure biological systems. A dynamic model of an intracellular regulatory network, for example, might contain several proteins of interest that carry out important functions in the system. We would then ideally like to know the concentrations of all these proteins in their various states and complexes, inside the cell, over time. But such direct measurements are currently not possible. Instead we are limited to indirect measurements such as relative protein levels compared to a control, reporter-based measurements, or averages over populations of cells. A compounding difficulty is that the measurements are often relatively noisy. It is thus challenging to accurately determine the unknown parameters of computational models of biological systems.

Intuitively, one would expect that multiple types of measurements, obtained using different experimental techniques, provide more information than a single type of measurement. The combined information would then be more likely to constrain the parameters in a computational model compared to using only a single measurement type. However, this need not be the case; perhaps one particular dataset, such as the most detailed measurements, already contains all relevant information, making additional datasets irrelevant.

The quantification of how much information a dataset brings to the parameter estimates, can be achieved using Bayesian statistics [2, 3]. For all unknown parameters

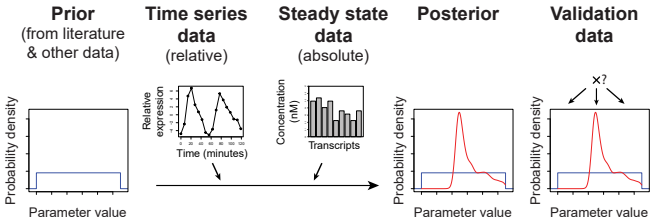


Figure 3.1: **Outline of the approach of data integration using Bayesian statistics.** Several initial datasets are assimilated into a prior probability distribution for all parameters in the model. Subsequently, multiple datasets are integrated to update the prior and obtain a posterior probability distribution for all parameters. Finally, this posterior probability distribution is compared to validation data.

in a model, a probability distribution is specified which quantifies the uncertainty in the parameters. This probability distribution can then be updated based on each of the different datasets, using Bayes' theorem. This provides a convenient framework for the integration of multiple datasets, as it allows a detailed comparison of the amount of information that can be extracted from each of the datasets.

Bayesian statistics has been applied to mechanistic computational models in biology in various settings and model types, including regulatory network models based on ordinary differential equations [3–7]. These applications have so far been limited to the use of a single dataset consisting of one type of measurement. It is thus unclear whether integration of multiple data types within the Bayesian formalism is feasible in practice and whether it is beneficial for achieving more accurate parameter estimates. The purpose of this study was to test the feasibility of this type of data integration, and to explore whether multiple data types can indeed provide more accurate parameter estimates.

We tested this approach using a model of a well-studied system, cell cycle regulation in budding yeast. Figure 3.1 shows the concept of data integration we used: various measurements are included as prior information, subsequently two types of data are incorporated during the inference, and finally the obtained parameter estimates are compared to measurements of these rates from independent studies.

3.2. APPROACH AND RESULTS

3.2.1. CONSTRUCTING AN INITIAL MODEL

We will use cell cycle regulation in budding yeast as test case, as this system is well studied and there is a host of data available. A central event in cell cycle regulation is the cyclic expression of the cyclin proteins [8]. We wished to model the cyclic expression pattern of four cyclins in particular: the G1-phase cyclin CLN3, the G1/Stransition cyclin CLN2, the S-phase cyclin CLB5 and the M-phase cyclin CLB2 (Figure 3.2A).

Although many models have been constructed of this system, for example [9, 10], we wished to obtain a simple, sparse model that is sufficient for explaining the cyclic expression of the cyclins. To this end we created a simple model that might be able to do this. The structure of this initial model is shown in Figure 3.2B and the reasoning behind it is as follows. Since the expression of the cyclins oscillate at the transcriptional level, we need to include the transcription factors that are responsible for regulating the

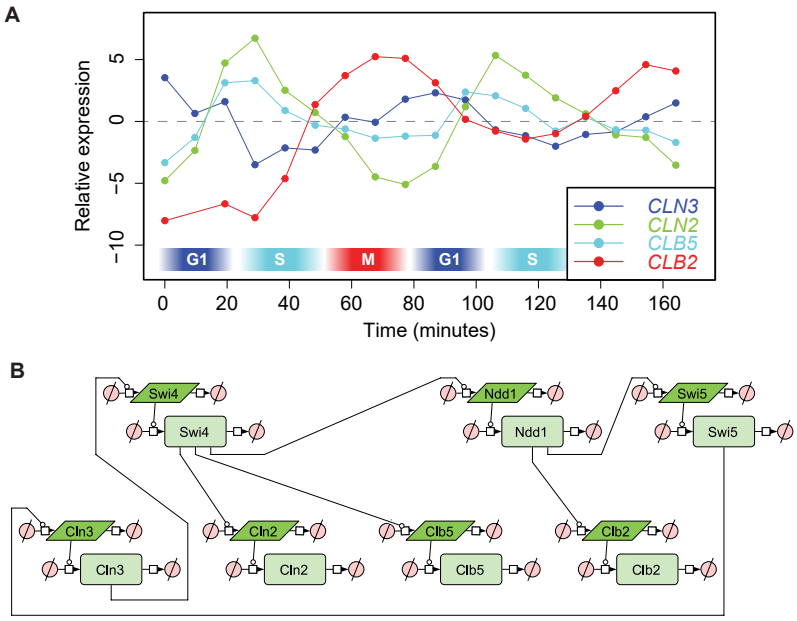


Figure 3.2: **Cyclins and model overview.** (A) The expression patterns of the four cyclins included in the model. The measurements are from (Spellman et al., 2003). The approximate cell cycle phase is indicated at the bottom. (B) Initial structure of the model in Systems Biology Graphical Notation.

transcription of the cyclins in the model. Thus, based on the overview of the cell cycle provided in [8], and especially Figure 3-34 therein, we included the three transcription factor complexes SBF, Mcm-Fkh and Swi/Snf. Each of these complexes is represented by one of their subunits: SBF is represented by the regulatory subunit SWI4; Mcm1-Fkh is represented by the coactivator NDD1; and Swi/Snf is represented by the subunit SWI5. We chose these subunits because they are regulatory factors and are transcriptionally oscillating [11]. As most data is available at the mRNA level, we explicitly included the mRNA transcripts as well as the proteins as species in the model. The dynamics are modeled by including rates for transcription, translation, and degradation of both mRNA and protein. To keep the model manageable, we did not explicitly include processes such as post-translational modifications, complex formation, and intracellular localization. While these processes are also clearly important for cell cycle regulation, the goal is not to create a detailed model but rather a simple model that is sufficient for explaining the cyclic expression of the cyclins. For the same reason, the model contains fewer signaling events than the more comprehensive model of Chen et al [10]. The starting model described here will likely require improvement, which we consider below. Starting from a simple model allows us to find a balance between model complexity and data fit. The resulting model can then be used for testing the integration of multiple datasets.

Another important modelling choice is that we specified the model entirely in physical units of concentration (micromolars) and time (seconds), rather than using dimensionless parameters and abundances. The physical units allow a comparison of the pa-

Measurement	Experimental technique	Used as	Reference
Protein concentration	2D-gel electrophoresis	Prior	[12]
mRNA concentration	Hybridization kinetics	Prior	[13]
Cell size	Electrical conductivity	Conversion	[14]
Transcript elongation rate	ChIP	Prior	[15]
RNA polymerase footprint	Nuclease digestion	Prior	[16]
Peptide elongation rate	Radioactive labeling	Prior	[17, 18]
Ribosome footprint	Nuclease digestion	Prior	[19]
mRNA time course (relative)	Microarray	Inference	[20–22]
mRNA concentration	SAGE	Inference	[23]
mRNA concentration	Microarray	Inference	[24]
Protein concentration	TAP tag; western blot	Inference	[25]
Protein concentration	GFP tag; flow cytometry	Inference	[26]
Protein concentration	2D-HPLC; MS/MS	Inference	[27]
mRNA degradation rate	Microarray	Validation	[28]
Transcription rate	GRO; ChIP-chip	Validation	[29]
Translation rate	Polysome profiling	Validation	[30]

Table 3.1: All datasets used in this study.

rameter estimates to measurements from independent studies. The model is specified in terms of ordinary differential equations, and the rate equations are based on mass action kinetics with the addition of a nonlinear term for modeling inhibitory effects. The model is described in more detail in the Methods section, and SBML versions of all models are included in Supplementary File 1.

3.2.2. CONSTRUCTING PRIORS FROM SEVERAL DATASETS

The Bayesian analysis required us to specify prior probabilities for the unknown parameters in the model. For each of the parameters, we specified priors based either on biochemical limits, or on published datasets providing information for a parameter. The prior probability distributions and how they were established are described in more detail in the Supplementary Methods. All datasets used throughout this study are listed in Table 3.1.

3.2.3. FITTING TIME COURSE MRNA MEASUREMENT DATA

As we wished to obtain a model for the cyclic expression of the cyclins, we first turned to measurements of mRNA gene expression over time [20–22], and tested whether the model can fit these datasets.

A complication with these datasets is that the measurements were taken under different growth conditions, with different synchronization methods and with slightly different yeast strains, resulting in different doubling times for the cells, ranging from 60 to 100 minutes. To make the datasets compatible, we used the time-normalized data provided by Cyclebase [11], and scaled the times back to an 80-minute cell cycle, which is a typical doubling time for yeast cells in rich (YEPD) medium [14].

We fitted the model to these three gene expression datasets simultaneously. The measurements are all made on synchronized cells relative to unsynchronized controls, and the likelihood function was specified such that it reflects this. Specifically, the likelihood of the observed values was centered on the log ratio of the modeled transcript con-

centration divided by the average modeled concentration over time (see the Methods section). We expected that the model would not exactly match the measurement data, and so a t-distribution was used as error model, such that occasional outlying measurements with respect to the model are not penalized too heavily.

The posterior probabilities were calculated using Markov chain Monte Carlo (MCMC) sampling. The relatively large number of dimensions, with the prior in each dimension spanning many orders of magnitude, makes this a challenging inference task. To be able to run the inference in reasonable time, the Bayesian inference software package BCM was used [31]. The posterior probability distribution contained sub-optimal modes, we therefore used parallel tempering [32] to have a means of escaping these. MCMC sampling relies on a proposal distribution; a distribution from which new candidate parameter values are drawn. For the efficiency of the sampler it is important that the proposal distribution reflects the shape and scale of the (unknown) posterior distribution. We therefore used automated blocking [33] and adaptively scaled the proposal distributions (see the Methods section). Traces and autocorrelation plots for the convergence analysis of all models are included in Supplementary File 2.

To test the goodness of fit, we first used graphical posterior predictive checking. The posterior predictive distribution describes a new, predicted dataset given the fitted model. Overlaying this posterior predictive distribution on the observed measurements provides a convenient way of identifying which data can and cannot be explained by the model. Figure 3.3 (top row) shows the posterior predictive distribution of the mean of the relative transcript levels in the fitted model overlaid on the observed measurements. It is immediately clear that the model cannot adequately explain the expression patterns of the four cyclins. The model can only fit the first peak of CLN3 expression, but not the subsequent oscillations or the activation of the other cyclins.

To further quantify the goodness of fit, coefficients of determination (R^2) were calculated for the four cyclins (Figure 3.3). We compared these values to the R^2 of a spline fit to the data. The spline fit gives a reference R^2 for the optimal fit that can be achieved. The median R^2 for the model fits range from 0.07 to 0.19, whereas a spline fit gives R^2 values ranging from 0.46 to 0.72, again showing that the initial model is insufficient to explain the expression patterns of the cyclins.

3.2.4. ITERATIVE MODEL REFINEMENT TO CREATE A WELL-FITTING MODEL FOR THE TIME COURSE MRNA MEASUREMENT DATA

As the simplest model could not adequately fit the transcription data, it was necessary to expand the model with additional explanatory factors. We thus searched the literature to find important mechanisms that were missing from the model. For each addition, we re-fitted the model to the data, and compared the posterior predictive distributions and R2 values for expression of the four cyclins. Note that we could not use the marginal likelihood for model selection here, because when we added additional species to the model we also included the expression data for those new species in the likelihood function. This affects the marginal likelihood; the marginal likelihood of two differing sets of data cannot be compared to each other.

The first addition to the model which we considered was the transcription factor HCM1. There is a significant delay between the transcriptional peak of SWI4 and NDD1,

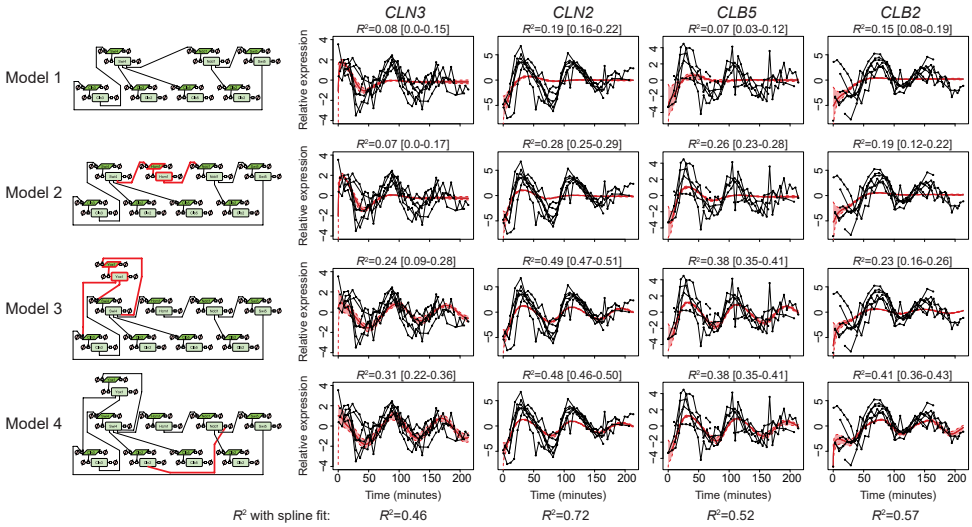


Figure 3.3: **Creating a model that can fit the time course mRNA data.** On the left the model structure is indicated in Systems Biology Graphical Notation, with the simplest model at the top. The changes with respect to the previous model are highlighted in red. On the right the mRNA time course measurement data of the cyclins is shown, overlaid with the posterior predictive of the mean of the data. The thick red line indicates the median and the shaded red area indicates the 90% confidence interval. Above each graph the median R2 is shown and the 90% confidence interval is given in brackets.

especially compared to the peaks of CLN2 and CLB5 which occur more rapidly after the expression of SWI4 (see Supplementary File 2 for the trajectories of all species). The transcription factor HCM1 has been found to be an important part of the transcriptional cell cycle regulation system [21], and the inclusion of this factor could introduce the necessary delay in the model. As shown in Figure 3.3 (second row), the addition of HCM1 indeed improved the fit of the model, particularly for the induction of the expression of CLN2 and CLB5 after SWI4 expression. However, the model was still not able to explain the oscillatory aspect of the expression of the four cyclins.

The lack of oscillatory behavior of the model suggested that a feedback loop might be required. We therefore considered the addition of the inhibitory transcription factor YOX1 [34]. This transcription factor provides a negative feedback loop from SWI4 back to both SWI4 and CLN3. As shown in Figure 3.3 (third row), with this addition the model could indeed recapitulate the oscillatory aspect of the expression pattern of the four cyclins.

As the magnitude of the oscillations in CLB2 was still greater in the data than could be explained by the model, we considered the addition of another regulatory mechanism, namely the degradation of NDD1 by the anaphase promoting complex [35]. This complex is normally active, unless it is inactivated by CLN2 [8]. Thus, NDD1 would be actively degraded until CLN2 signals the start of S-phase. As shown in Figure 3.3 (bottom row), with the addition of this mechanism the model can indeed better explain the expression pattern of the NDD1-target CLB2.

With these additions to the model, the expression patterns of the four cyclins are adequately explained ($R^2 > 0.3$ for all cyclins, and at least 65% of the R^2 achieved with a spline fit). Although further additions can be considered, we wished to keep the model as small as possible while achieving a reasonable fit. This was mainly done to keep the computational requirements manageable – to generate 1,000 posterior samples for the fourth, extended model required approximately 60 hours. The structure of the resulting model is similar to the Boolean network model of Orlando et al [36] in terms of the transcriptional regulatory network.

3.2.5. SIMULTANEOUS FITTING OF TIME COURSE AND STEADY STATE MEASUREMENT DATA

Now that the model is able to explain the relative time course measurements adequately, we can start including additional datasets to test whether the parameters of the model can be more tightly constrained with the integration of additional data. We turned to absolute measurements of the mRNA [23, 24] and protein [25–27] concentrations of the species in the model. These measurements were done at steady-state growth conditions in non-synchronized cells. We incorporated this in the likelihood by taking the time average of the modeled trajectories, and setting this time average as the modeled value of the steady state data, where the time average was taken over two cell cycles.

The addition of absolute concentration data to relative time series data may seem trivial, and it could potentially also be achieved by transforming the kinetic parameters and concentrations accordingly. However, keeping the model specified in physical dimensions (micromolars and seconds) is natural, and more importantly, it allows for a direct comparison of the kinetic rates with measurements of these rates later on.

Figure 3.4 shows the posterior trajectories of the transcripts of one of the cyclins, CLN3, after fitting the relative time course data alone, the absolute steady state data alone, or all data together (trajectories for all other species in the model are included in Supplementary File 2). Several observations can be made. First, it is apparent that the absolute concentrations can be quite high when only time course data is used. When the steady state data is included however, the concentrations are constrained to be much lower. Second, when only steady state data is used, the model displays various behaviors including stable expression, decay and oscillations (see the individual trajectories depicted in grey) – each of these behaviors would be consistent with the given average data over a period of two cell cycles. With all measurements types included, the model displays the correct oscillations at concentrations consistent with the steady state data. Finally, the fit to the relative time course data is not compromised by the inclusion of the absolute steady state data, and vice versa. The model is thus able to fit both types of data at the same time, and no modifications need to be made to the model structure to accommodate the steady state data.

Figure 3.5A shows the 90% posterior confidence intervals for several parameters in the model, for the relative time course data alone, the absolute steady state data alone, or all data together (confidence intervals and density plots for all parameters are included in Supplementary Figure 1 and 2). For several parameters, each data type separately provides some information, but the inclusion of the two types of data together provides significantly tighter confidence intervals, for example for the translation rate. There are

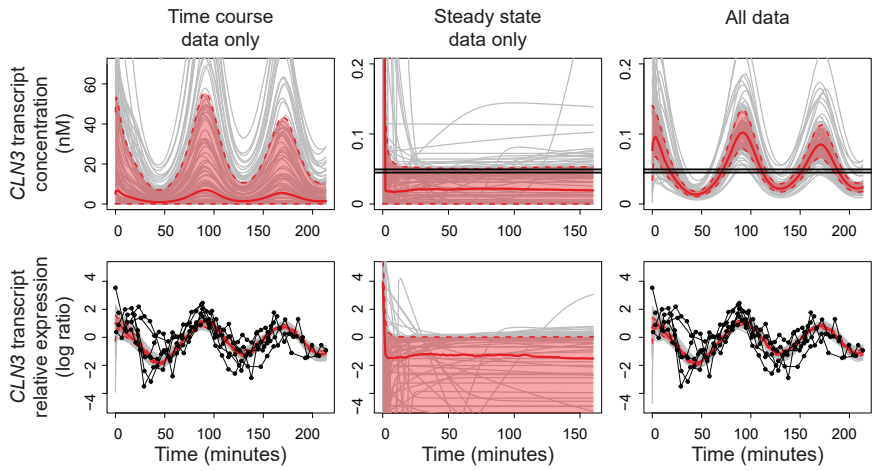


Figure 3.4: **Measurements and posterior predictive with the two types of data separately and together for the CLN3 transcript.** The top row shows the absolute concentrations and the bottom row shows the expression relative to the time average (by log ratio). In the left column, only time course data is used, in the middle column only steady state data is used and in the right column both types of data are used. The measurements are shown in black: horizontal line for absolute steady state data (note that each line is only a single measurement value), and connected dots for relative time course data (here each dot is a measurement value). The thick red line indicates the median of the posterior predictive and the shaded red area indicates the 90% confidence interval. Grey lines indicate individual trajectories for 100 posterior parameter samples.

also parameters that can already be inferred from the time course data alone; that is, for these parameters the addition of the steady state data does not reduce the confidence intervals, such as the degradation rate of CLN3. In many cases, the steady state data by itself provides little information for constraining the parameters, which is not surprising for a dynamic model. However, the addition of the steady state data to the time course data does reduce the uncertainty compared to the time course data alone. Examples of this are the degradation rate of SWI4 or the transcription rate of CLN2.

In general across all parameters, combining multiple types of measurements reduces the uncertainty in parameter estimates (Figure 3.5B). With all data types included, 45 out of 54 parameters have 90% confidence intervals of less than half of the prior range, whereas the steady state data by itself constrains only 1 parameter to this extent and the time course data 14 parameters. Comparing the added value of the absolute protein and transcript concentrations, we note that it is mainly the transcript concentrations which reduces the uncertainty (column 4 and 5 in Figure 3.5B). Nevertheless, adding the protein concentration data to the time course and transcript concentration data still further reduces the uncertainty for several parameters.

3.2.6. COMPARISON OF PARAMETER ESTIMATES WITH RATE MEASUREMENT DATA

To test whether the obtained parameter estimates are accurate, we compared them to measurements from three additional, independent datasets. In particular, the mRNA

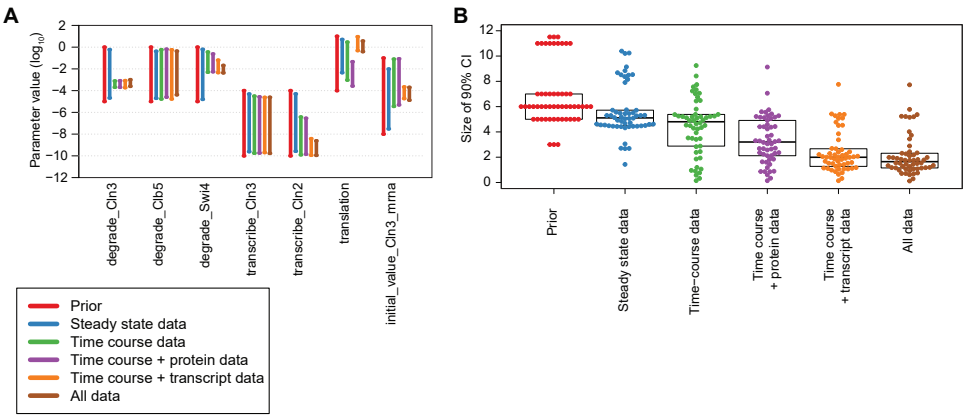


Figure 3.5: **Uncertainty reduction by the different datasets.** (A) Posterior 90% confidence intervals for several model parameters, as inferred using the absolute steady state data, relative time course data, the time course data together with absolute transcript or protein data, or all data. The confidence intervals and posterior probability densities for all parameters are given in Supplementary Figure 1 and 2. (B) Size of the 90% confidence intervals of all parameters on $\log_{10}$ scale. For the prior, the full range is given.

degradation rate [28]; transcription rates [29] and translation rates [30] have been measured for budding yeast. Figure 3.6A shows the measured values of the parameters compared to the posterior probability distributions of the parameter estimates.

For the mRNA degradation rate, the measurements are in close agreement with the predicted rates (Figure 3.6A, left panel). We assumed a common rate parameter for all species, while the measurements were done for each gene separately, and there is indeed some variability between the measurements for the genes that were included in the model. Nevertheless, the measured degradation rates of all genes are within the same order of magnitude as the estimated average degradation rate (the difference between the measurements and the maximum a posteriori estimate on $\log_{10}$ scale is less than 0.5), so the scale of the average mRNA degradation rate was predicted accurately.

For the transcription rates, these rates in the model are split into two parts: basal transcription and transcription factor induced transcription. The rate measurements are population averages, and as each cell would be in a different stage of the cell cycle, they will be expressing different levels of the transcription factors. To be able to compare the measurements of the transcription rates to the model's estimated rates, it is necessary to calculate the total, average transcription rate. This was obtained by averaging the transcription rate of each gene over time. This rate thus includes the effect of the time-varying expression of the transcription factors. When only time course data was used, the transcription rates were not constrained, but they do have non-zero probability at the measured values. However, when all data is included, the estimated transcription rates closely match the measured values for most genes (Figure 3.6A, middle panels; 7 of the 8 measured values lies within the 90% confidence interval, and the remaining gene is at least within the same order of magnitude). Thus, for the transcription rates, the addition of the absolute concentrations to the relative dynamic data constrained the parameter estimates to values close to or matching the measurements of these rates.

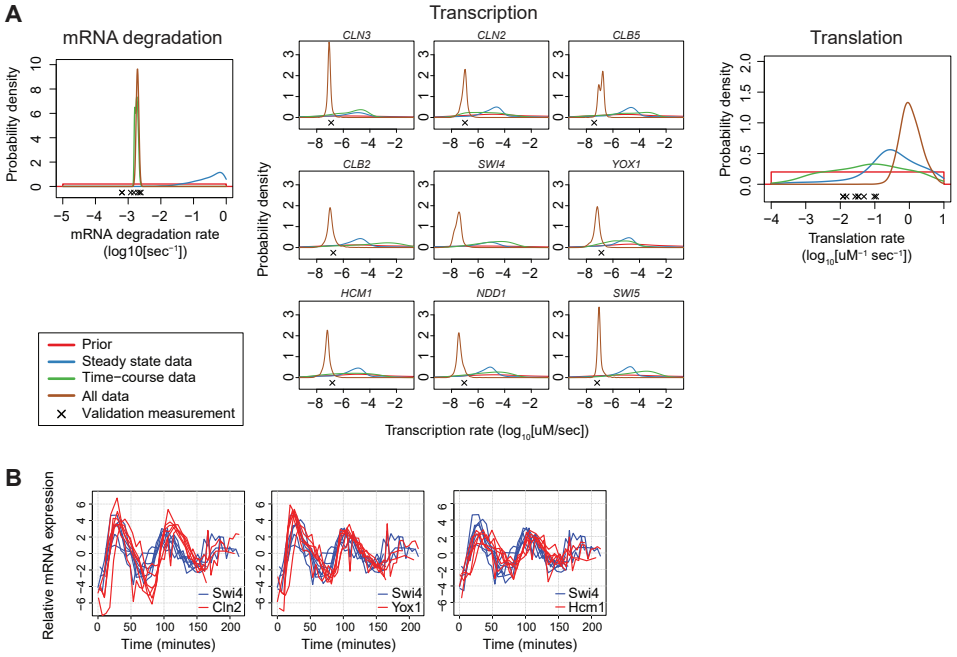


Figure 3.6: **Comparison of parameter estimates with validation data.** (A) Posterior probability distributions of the parameters, as inferred using the absolute steady state data, relative time course data or both. The validation measurements are marked with a black cross below the probability distributions. (B) Trajectories for the mRNA expression levels of the transcription factor subunit SWI4 and its target genes CLN2, YOX1 and HCM1.

The measured translation rates have been estimated from ribosome densities using polysome profiling, whereby a processing speed of 10 amino acids/s was assumed [30]. Note that the authors mention that their estimates should be used with caution. Nevertheless, assuming they are accurate, then the model estimate using all inference data are two orders of magnitude too high (Figure 3.6A, right panel). The model estimate is indeed quite high at around 1 protein/transcript/s. While this is feasible given the prior information, it would require that the transcripts are always essentially fully packed with ribosomes.

To find the reason for this high translation rate estimate, we investigated the trajectories of the transcription factors and their target genes. If we compare the mRNA expression trajectories for the transcription factor SWI4 and its targets CLN2, HCM1 and YOX1 (see Figure 3.6B), it makes sense that the model requires a high translation rate. The peaks in transcription of the target genes follow very closely after the peak in transcription of the transcription factor, especially in the first cell cycle. Given that this process of rapid induction of transcription in the model has to occur through the translation of the transcription factor, then there are two ways in which the model might fit the data: either the translation rate must be high, or the concentration of the transcript of the transcription factor must be high. When using only the relative data, it is not possible to distinguish between these scenarios; indeed in this case the translation rate is not constrained: the 90% confidence interval of the translation rate spans almost 3 orders of magnitude (Figure 3.5). However, when including both relative and absolute data, the inference can make use of the information that the concentration of the transcription factor is low. It can thus be inferred that the translation rate must be high, given this model.

It is known that other mechanisms are at play here as well, such as the regulation of SWI4 and the SBF transcription factor complex through phosphorylation by different cyclin/CDKs [37]. Indeed it has been shown that induction of G1-phase transcripts can occur in the absence of protein translation [38]. It is likely that a model with additional layers of SWI4 regulation would be able to fit all data with lower translation rates. Unfortunately we were not able to expand the model with such additional effects, as the parameter inference for these expanded models would involve a prohibitive amount of computation time. Regardless, these results show that the model can be identified as being incomplete by using the inference of parameters from multiple datasets. This model deficiency could not be deduced from any of the datasets alone.

3.3. METHODS

3.3.1. MODEL EQUATIONS

The computational model consists of two types of species: the proteins, and the mRNA transcripts. The rate equations for these species are based on mass action kinetics, with the addition of a nonlinear term for modeling inhibitory effects. For transcripts, the rate equation contains three terms: one for transcription, one for inhibition of transcription and one for degradation. The transcription rate is proportional to the concentration of the activating transcription factor for that gene. This transcription can be inhibited by an inhibitory transcription factor. Each transcript has exactly one activating transcription

factor and at most one inhibitory transcription factor. For proteins, the rate equation also contains three terms: one for translation, one for degradation, and one for inhibition of degradation. The translation rate is proportional to the concentration of the transcript for that protein, and the degradation rate is proportional to the concentration of the protein itself. See the Supplementary Methods for a more detailed description and the equations.

3.3.2. PRIOR DISTRIBUTIONS

For all parameters, we used uniform priors on a $\log_{10}$ scale. A log scale was chosen as we were interested in the orders of magnitude of the parameters rather than their precise values. The limits of the uniform distributions were chosen based on various data points and biochemical limits as described the Supplementary Methods.

3.3.3. LIKELIHOOD

Firstly, the time average of the concentration of a transcript was calculated by averaging over two full cell cycles. Then, for relative time course data measured using synchronized cells relative to unsynchronized cells, we modeled the relative value by dividing the modeled concentration by the time average and taking the log. As error model we used a t-distribution with three degrees of freedom, as a means of robust inference [39]. This distribution was centered on the log ratio of the relative expression. For the absolute concentration data, the time average value is $\log_{10}$ transformed, and again a t-distribution is used as error model. As for the prior, the likelihood is specified on a log scale as it is sufficient if the model captures the right order of magnitude. See the Supplementary Methods for the equations.

3.3.4. MODEL INFERENCE

The posterior probability distributions were calculated using parallel-tempered Markov chain Monte Carlo [32], using the Bayesian inference software package BCM [31]. For the initial model, we used 32 parallel chains with the temperatures of the chains distributed quadratically. The burn-in period was set to 1.25 million samples followed by a sampling period of 5 million posterior samples, which were subsampled at 1 in 2,500. At each step, a random choice was made between updating each chain with 5 Metropolis-Hastings steps, and swapping a random adjacent pair of chains. The probability of selecting a swap step was set to 0.99. For the proposal distribution in the Metropolis-Hastings steps, the parameters were blocked automatically [33] and we used a multivariate normal distribution for each block of parameters. The proposal covariance matrix for each block was set to the empirical covariance of the preceding samples and adaptively scaled to obtain an acceptance rate of 0.23 within each block. These settings produced sufficiently uncorrelated posterior samples (see Supplementary File 2 for traces and autocorrelation plots), and were sufficient to achieve at least 100 round trips from prior to posterior. The sampling period and subsampling was doubled for model 3 and quadrupled for model 4, such that the resulting posterior samples were still sufficiently uncorrelated and at least 100 round trips from prior to posterior were achieved.

All files required for running the inference in BCM, including the prior and likelihood specification, the models in SBML/CellDesigner format, as well as a NetCDF archive

containing the pre-processed data, are included in Supplementary File 1.

3.3.5. MODEL CHECKING

The model fit was investigated using the posterior predictive distribution and coefficients of determination. The posterior predictive distribution is the probability distribution of a new set of data, given the model and the observed data. This distribution was approximated using the posterior Monte Carlo samples. The coefficients of determination for the time course data were calculated relative to a null model which has a separate mean for each experiment. A reference R2 was calculated by fitting a cubic spline to the data with the smoothing parameter selected through cross-validation. See the Supplementary Methods for details and equations.

3.4. DISCUSSION

Model determination is an important aspect of computational modeling. Models in systems biology are frequently underdetermined, and as a result it is often not possible to confirm or falsify a particular model. There is thus a need for methods to determine models more accurately. With the increasing amount of data available for many biological systems, the use of multiple datasets to constrain the parameters from different angles is a promising avenue. Bayesian statistics provides a coherent and convenient framework to accomplish this. Here, we have shown that it is feasible to integrate diverse datasets during the Bayesian inference of parameters of an ODE-based model. The process as described here may be useful as a general recipe for integrating diverse measurement types also in other settings. More importantly, we have shown that this integration of diverse data types can provide tighter posterior estimates, at least in obtaining the right order of magnitude, thus achieving more accurate model determination. We noticed that even when a single dataset, taken by itself, provides little information, it can still significantly improve parameter estimates when used in conjunction with other datasets.

There are several challenges when using this type of data integration based on model simulation and Bayesian statistics. The biggest challenge is the scaling of the computational demands with respect to the size of the model. This is due to two reasons. First, the simulation of a computational model typically does not scale well with model size (cubically in the case of direct, implicit ODE solvers). Second, the parameter inference is increasingly challenging when the number of parameters increases. Although in theory Monte Carlo methods scale independently of the dimensionality, this requires that the samples are concentrated in regions of high posterior probability. The efficiency of generating a good set of samples critically depends on the proposal distribution that is used. Given the complex shape of the posterior probability distributions of biological computational models, in particular the presence of multiple modes and ridges [7, 40], proposal distributions typically become much less efficient with higher dimensionality. Both of these computational challenges apply more generally to any approach using model simulation and global parameter inference. Increases in computational capabilities, more efficient simulation methods, and sampling or optimization methods tailored for the inference of biological computational models, may allow larger models in the

future.

For budding yeast, and their cell cycle regulation in particular, many more measurements have been performed, such as mRNA quantification by qPCR [41] and RNA sequencing [42], protein-level time course data by mass spectrometry [43] and GFP-tagged time lapse microscopy [44]. In principle, these data can be integrated with the same approach as was done for the data in the present study, and it would be interesting to explore the contributions and concordance of these measurements. A challenge for further integration of time-course data is the synchronization of the timing, which is not straightforward when using different experimental setups. This synchronization can also directly affect kinetic rates, for example the alignment of transcript and protein time course data can directly influence the estimated translation rate.

To be able to compare the contribution of the different datatypes, it is necessary to quantify the uncertainty in the parameter estimates, which was achieved here using Bayesian statistics. The quantification of uncertainty has previously been achieved with different approaches as well (reviewed in [45]), including using the profile likelihood [46] and through bootstrapping [47]. The incorporation of multiple datasets in the likelihood function can in principle be translated to these formalisms as well. A unique advantage of the Bayesian approach is the ability to explicitly include data as prior information, which we have utilized to incorporate various datasets. Profile likelihoods may be computationally more efficient to calculate than posterior probabilities, although the calculations still involve the most challenging aspect, namely global optimization. The profile likelihood is limited in that it provides uncertainty estimates for each parameter separately rather than for all parameters jointly.

In conclusion, we have shown that diverse types of measurements can be successfully integrated during the inference of parameters of ODE systems using Bayesian statistics. This integration provided more tightly constrained parameter estimates, thereby achieving a more accurate model determination.

REFERENCES

[1] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, *Universally sloppy parameter sensitivities in systems biology models,* PLoS Computational Biology **3**, 1871 (2007), 0701039 .

[2] D. J. Wilkinson, *Bayesian methods in bioinformatics and computational systems biology.* Briefings in bioinformatics **8**, 109 (2007).

[3] V. Vyshemirsky and M. Girolami, *Bayesian ranking of biochemical system models.* Bioinformatics (Oxford, England) **24**, 833 (2008).

[4] T.-R. Xu, V. Vyshemirsky, A. Gormand, A. von Kriegsheim, M. Girolami, *et al.*, *Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species.* Science signaling **3**, ra20 (2010).

[5] T. Toni and M. P. H. Stumpf, *Simulation-based model selection for dynamical systems in systems and population biology,* Bioinformatics **26**, 104 (2009).

[6] H. Eydgahi, W. W. Chen, J. L. Muhlich, D. Vitkup, J. N. Tsitsiklis, and P. K. Sorger, *Properties of cell death models calibrated and compared using Bayesian approaches.* Molecular systems biology **9**, 644 (2013).

[7] S. Hug, A. Raue, J. Hasenauer, J. Bachmann, U. Klingmüller, J. Timmer, and F. Theis, *High-Dimensional Bayesian Parameter Estimation: Case Study for a Model of JAK2/STAT5 Signaling,* Mathematical Biosciences **246**, 293 (2013).

[8] D. O. Morgan, *The Cell Cycle - Principles of Control* (Oxford University Press, 2007).

[9] J. J. Tyson, *Modeling the cell division cycle: cdc2 and cyclin interactions.* Proceedings of the National Academy of Sciences of the United States of America **88**, 7328 (1991).

[10] K. C. Chen, L. Calzone, A. Csikasz-Nagy, F. R. Cross, B. Novak, and J. J. Tyson, *Integrative Analysis of Cell Cycle Control in Budding Yeast,* Molecular biology of the cell **15**, 3841 (2004).

[11] A. Santos, R. Wernersson, and L. J. Jensen, *Cyclebase 3.0: A multi-organism database on cell-cycle regulation and phenotypes,* Nucleic Acids Research **43**, D1140 (2015).

[12] B. Futcher, G. I. Latter, P. Monardo, C. S. McLaughlin, and J. I. Garrels, *A sampling of the yeast proteome.* Molecular and cellular biology **19**, 7357 (1999).

[13] L. M. Hereford and M. Rosbash, *Number and distribution of polyadenylated RNA sequences in yeast,* Cell **10**, 453 (1977).

[14] C. B. Tyson and P. G. Lord, *Dependency of size of Saccharomyces cerevisiae cells on growth rate,* Journal of Bacteriology **138**, 92 (1979).

[15] P. B. Mason and K. Struhl, *Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo,* Molecular Cell **17**, 831 (2005).

[16] C. P. Selby, R. Drapkin, D. Reinberg, and A. Sancar, *RNA polymerase II stalled at a thymine dimer: Footprint and effect on excision repair,* Nucleic Acids Research **25**, 787 (1997).

[17] K. W. Boehlke and J. D. Friesen, *Cellular content of ribonucleic acid and protein in Saccharomyces cerevisiae as a function of exponential growth rate: calculation of the apparent peptide chain elongation rate,* Journal of Bacteriology **121**, 429 (1975).

[18] C. Waldron, R. Jund, and F. Lacroute, *The elongation rate of proteins of different molecular weight classes in yeast,* FEBS Letters **46**, 11 (1974).

[19] S. L. Wolin and P. Walter, *Ribosome pausing and stacking during translation of a eukaryotic mRNA.* The EMBO journal **7**, 3559 (1988).

[20] P. T. Spellman, G. Sherlock, M. Q. Zhang, R. Vishwanath, K. Anders, M. B. Eisen, P. O. Brown, and B. Futcher, *Comprehensive Identification of Cell Cycle – regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization,* Molecular biology of the cell **9**, 3273 (2003).

[21] T. Pramila, W. Wu, S. Miles, W. S. Noble, and L. L. Breeden, *The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle,* Genes & Development **20**, 2266 (2006).

[22] M. V. Granovskaia, L. J. Jensen, M. E. Ritchie, J. Toedling, Y. Ning, P. Bork, W. Huber, and L. M. Steinmetz, *High-resolution transcription atlas of the mitotic cell cycle in budding yeast.* Genome biology **11**, R24 (2010).

[23] V. E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, *et al., Characterization of the yeast transcriptome,* Cell **88**, 243 (1997).

[24] F. C. Holstege, E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, *et al., Dissecting the regulatory circuitry of a eukaryotic genome.* Cell **95**, 717 (1998).

[25] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman, *Global analysis of protein expression in yeast.* Nature **425**, 737 (2003).

[26] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman, *Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise,* Nature **441**, 840 (2006).

[27] P. Lu, C. Vogel, R. Wang, X. Yao, and E. M. Marcotte, *Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation,* Nat Biotechnol **25**, 117 (2007).

[28] Y. Wang, C. L. Liu, J. D. Storey, R. J. Tibshirani, D. Herschlag, and P. O. Brown, *Precision and functional specificity in mRNA decay.* Proceedings of the National Academy of Sciences of the United States of America **99**, 5860 (2002).

[29] V. Pelechano, S. Chávez, and J. E. Pérez-Ortín, *A Complete Set of Nascent Transcription Rates for Yeast Genes,* Current Science **101**, 1435 (2011), 1203.2655 .

[30] Y. Arava, Y. Wang, J. D. Storey, C. L. Liu, P. O. Brown, and D. Herschlag, *Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae,* Proceedings of the National Academy of Sciences of the United States of America **100**, 3889 (2003).

[31] B. Thijssen, T. M. H. Dijkstra, T. Heskes, and L. F. A. Wessels, *BCM: toolkit for Bayesian analysis of Computational Models using samplers,* BMC Systems Biology **10**, 100 (2016).

[32] C. J. Geyer, *Markov Chain Monte Carlo Maximum Likelihood,* in *Proceedings of the 23rd Symposium Interface*, 1 (1991) pp. 156–163.

[33] D. Turek, P. de Valpine, C. J. Paciorek, and C. Anderson-Bergman, *Automated Parameter Blocking for Efficient Markov-Chain Monte Carlo Sampling,* Bayesian Analysis **12**, 465 (2017), 1503.05621 .

[34] T. Pramila, S. Miles, D. Guhathakurta, D. Jemiolo, and L. L. Breeden, *Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M / G1 phase of the cell cycle,* Genes & development **16**, 3034 (2002).

[35] J. Sajman, D. Zenvirth, M. Nitzan, H. Margalit, K. J. Simpson-Lavy, *et al.*, *Degradation of Ndd1 by APC/CCdh1 generates a feed forward loop that times mitotic protein accumulation,* Nature Communications **6**, 7075 (2015).

[36] D. A. Orlando, C. Y. Lin, A. Bernard, J. Y. Wang, J. E. S. Socolar, E. S. Iversen, A. J. Hartemink, and S. B. Haase, *Global control of cell-cycle transcription by coupled CDK and network oscillators.* Nature **453**, 944 (2008).

[37] R. F. Siegmund and K. A. Nasmyth, *The Saccharomyces cerevisiae Start-specific transcription factor Swi4 interacts through the ankyrin repeats with the mitotic Clb2/Cdc28 kinase and through its conserved carboxy terminus with Swi6.* Molecular and cellular biology **16**, 2647 (1996).

[38] N. J. Marini and S. I. Reed, *Direct induction of G1-specific transcripts following reactivation of the Cdc28 kinase in the absence of de novo protein synthesis,* Genes and Development **6**, 557 (1992).

[39] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Models for robust inference,* in *Bayesian Data Analysis, Third edition* (CRC Press, Boca Raton, FL, 2014) pp. 435–445.

[40] M. Girolami, *Bayesian inference for differential equations,* Theoretical Computer Science **408**, 4 (2008).

[41] F. Miura, N. Kawaguchi, M. Yoshida, C. Uematsu, K. Kito, Y. Sakaki, and T. Ito, *Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs.* BMC genomics **9**, 574 (2008).

[42] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, *The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing,* Science **320**, 1344 (2008).

[43] M. R. Flory, H. Lee, R. Bonneau, P. Mallick, K. Serikawa, D. R. Morris, and R. Aebersold, *Quantitative proteomic analysis of the budding yeast cell cycle using acid-cleavable isotope-coded affinity tag reagents,* Proteomics **6**, 6146 (2006).

[44] D. A. Ball, J. Marchand, M. Poulet, W. T. Baumann, K. C. Chen, J. J. Tyson, and J. Peccoud, *Oscillatory dynamics of cell cycle proteins in single yeast cells analyzed by imaging cytometry,* PLoS ONE **6** (2011).

[45] J. Vanlier, C. Tiemann, P. Hilbers, and N. van Riel, *Parameter uncertainty in biochemical models described by ordinary differential equations.* Mathematical biosciences **246**, 305 (2013).

[46] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer, *Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood,* Bioinformatics **25**, 1923 (2009).

[47] C. Brännmark, R. Palmér, S. T. Glad, G. Cedersund, and P. Strålfors, *Mass and information feedbacks through receptor endocytosis govern insulin signaling as revealed using a parameter-free modeling framework,* Journal of Biological Chemistry **285**, 20171 (2010).

SUPPLEMENTARY MATERIAL

3.4.1. MODEL EQUATIONS

The computational model consists of two types of species: the proteins p_i, and the mRNA transcripts m_i. The rate equations for these species are based on mass action kinetics, with the addition of a nonlinear term for modeling inhibitory effects.

For transcripts, the rate equation contains three terms: one for transcription, one for inhibition of transcription and one for degradation. The three terms are marked in the equation below, and they correspond to the following:

transcription The transcription rate is a sum of a constant rate and a rate that is proportional to the concentration of the activating transcription factor for that gene. Each transcript has exactly one activating transcription factor.

transcription inhibition The transcription can be inhibited by an inhibitory transcription factor, which is modeled with a two-parameter logistic function (including a 50% inhibitory concentration and a steepness parameter). Each transcript has at most one inhibitory transcription factor.

degradation Degradation is modeled as exponential decay.

The rate equation for transcripts is then defined as:

$$\frac{dm_i}{dt} = \underbrace{(r_{\text{base}} + r_{\text{induced},i,j}\, p_j)}_{\text{transcription}} \underbrace{\frac{1}{1 + e^{s_{k,i}(\log p_k - \log c_{k,i})}}}_{\text{transcription inhibition}} - \underbrace{d_{\text{mRNA}}\, m_i}_{\text{degradation}}, \tag{3.1}$$

where m_i is the concentration of mRNA transcript i in μM, r_{base} is the base transcription rate in μM/s, $r_{\text{induced},i,j}$ is the transcription rate of transcript i induced by transcription factor j in μM/μM/s, p_j is the concentration of the transcription factor j in μM, and d_{mRNA} is the degradation rate of all mRNA transcripts in 1/s. If an inhibitory transcription factor for gene i is included in the model, then this inhibitory transcription factor is indicated by index k, and $s_{k,i}$ is the steepness of the inhibition curve and $c_{k,i}$ is the 50%-inhibitory concentration in μM. If no inhibitory transcription factor is present for gene i, the transcription inhibition term is set to 1, such that the transcription is not inhibited.

For proteins, the rate equation also contains three terms: one for translation, one for degradation, and one for inhibition of degradation.

translation The translation rate is proportional to the concentration of the transcript (m_i) for that protein.

degradation The degradation is modeled as exponential decay, but it is split into two parts: a first part (d_i) which is constant and represents general decay/degradation of the protein, and a second part ($d_{\text{induced},k,i}$) which represents active, specific degradation, which can be inhibited by another protein. Both parts of the degradation rate are proportional to the concentration of the protein itself.

inhibited degradation The inhibition of degradation is modeled with a two-parameter logistic function (including a 50% inhibitory concentration and a steepness parameter). Each protein has at most one protein that can inhibit its degradation.

The rate equation for proteins is then defined as:

$$\frac{dp_i}{dt} = \underbrace{u m_i}_{\text{translation}} - \underbrace{d_i p_i}_{\text{degradation}} - \underbrace{\frac{d_{\text{induced},k,i}}{1 + e^{s_{k,j}(\log p_k - \log c_{k,i})}}\, p_i}_{\text{inhibited degradation}}, \tag{3.2}$$

where p_i is the concentration of protein i in μM, u is the translation rate in μM/μM/s, m_i is the concentration of mRNA transcript i in μM, and d_i is the degradation rate of protein i in 1/s. If a degradation-inhibiting protein is included in the model, then this protein is indicated by index k, and $d_{\text{induced},k,i}$ is the degradation rate of protein i that can be inhibited by protein k, $s_{k,i}$ is the steepness of the inhibition curve and $c_{k,i}$ is the 50%-inhibitory concentration in μM. If no degradation-inhibiting protein is present for protein i, the inhibited degradation term is set to 0.

3.4.2. PRIOR DISTRIBUTIONS

CELL SIZE

For various calculations and conversions, we need the cell size. Although the cell size varies between conditions and during the cell cycle, we assumed that the cell size is 37

μm^3 and constant. This is an average size for yeast cells growing in rich (YEPD) medium [14]. Combined with Avogadro's constant, this means that 1 molecule per cell corresponds to approximately $4.5 \cdot 10^{-5}$ μM.

CONCENTRATIONS AND INITIAL CONDITIONS

For setting a prior on the initial conditions of proteins, we used the dataset of Futcher et al. [12]. They sampled the yeast proteome and established protein copy number per cell for these proteins. Since the sampling across the proteome in this study was not uniform, but instead focused on the most abundant proteins, we used this dataset only to provide a reasonable upper limit. The most highly expressed proteins were present in the range of 1 million copies per cell, or approximately 45 μM. To provide some margin of error, we set the upper limit of protein concentrations at 100 μM. We assumed 0.1 nM as lower limit (approximately 2.2 molecules per cell).

For setting a prior on the initial conditions of transcripts, we used the measurements of Hereford and Rosbash [13]. They estimated the transcript copy number per cell to vary between 1 and 200 copies per cell. To provide some margin of error, we set the upper limit of transcript concentration at 0.1 μM (approximately 2,200 transcripts per cell). We assumed 10^{-8} μM as lower limit, to allow transcripts to be practically absent as well ($\ll$1 transcript per cell).

TRANSCRIPTION RATES

The model contains two classes of transcription rate parameters: basal transcription and transcription factor-induced transcription. To allow either of the two types of transcription to be practically absent, we set the lower limit to 10^{-10} μM/s and 10^{-10} μM/μM/s respectively. For the upper limit, we consider the case where transcription initiation is not rate limiting; the transcription rate is then bound by the transcript elongation rate and the number of polymerases transcribing the gene. We assume that the elongation rate is constant; this rate has been estimated at 2 kb/min [15]. The footprint of RNA polymerase has been estimated at 40 nucleotides [16]. If a gene is fully packed with polymerases, this gives a transcription rate of approximately 0.8 transcripts/s. To allow for some margin of error, we take the upper limit as 10^{-4} μM/s (approximately 2.2 transcripts/s).

Recall that a priori we expect that proteins are in the concentration range of 0.1 nM to 100 μM. To allow for transcription factors at the lower concentration limit, 0.1 nM, to already fully induce transcription of their target genes, the upper limit for transcription factor induced transcription rates was set at $10^{-4}/10^{-4}$=1 μM/ μM/s.

TRANSLATION RATES

For translation rates we use a similar logic as for the transcription rate: as upper limit we take the case where translation initiation is not limiting and the translation speed is bound by the ribosome progression and how much space the ribosome occupies on the transcript. We assume that the peptide elongation rate is constant, and it has been estimated at 10 amino acids/s [17] and 10.5 amino acids/s [18]. It has been reported that ribosomes can stack together along a transcript as closely as 27 nucleotides apart [19]. Together, this gives a translation rate of 1.2 proteins/transcript/s when initiation is not limiting and all ribosomes progress unimpeded over the transcript. To allow for some margin of error, such as faster elongation of the specific proteins studied here, we set

the maximal translation rate to 10 proteins/transcript/s. The lower limit is set at 10^{-4} proteins/transcript/s.

DEGRADATION RATES

For degradation rates, we withheld the available studies for validation purposes and so these could not be used for setting a prior distribution. We set the prior distribution for degradation rates to a wide range: between 10^{-5} and 1 s^{-1} for both proteins and transcripts, corresponding to a half-life between 19 hours and 0.7 seconds.

INHIBITION RATES

Inhibition was modeled by a non-linear function containing two parameters: the steepness and the 50%-inhibition concentration. The steepness is allowed to vary between 0.1 and 100. For the 50%-inhibition concentration, we used the same prior as for protein concentrations: between 0.1 nM and 100 µM.

MEASUREMENT VARIANCES

The prior for the measurement variance was set to an exponential distribution with rate $\lambda = 0.5$ for time course data and $\lambda = 1.0$ for steady state data.

3.4.3. LIKELIHOOD

The time average of the concentration of a transcript was calculated as

$$\bar{m_i} = \frac{1}{2t_{\text{cell-cycle}}} \int_0^{2t_{\text{cell-cycle}}} m_i(t)\,dt, \tag{3.3}$$

where $t_{\text{cell-cycle}}$ is the duration of the cell cycle (4800 seconds). Two full cell cycles were used, as the start of the first cell cycle can be affected by the method used to synchronize the cells and this effect can be alleviated by including a second cell cycle. Beyond the second cell cycle the cells typically start to diverge and are no longer synchronized. The average concentration of proteins is calculated in the same way with p_i instead of m_i.

For relative time course data measured using synchronized cells relative to unsynchronized cells, we modeled the relative value by dividing the modeled concentration at time t_n by the time average and taking the log:

$$x_i(t_n) = \log_2\left(\frac{m_i(t_n)}{\bar{m_i}}\right). \tag{3.4}$$

The likelihood function for the relative time course data is then defined as

$$P(y_{i,t_n}|x_i(t_n)) = t(y_{i,t_n}|\mu = x_i(t_n), \sigma = \sigma_i, \nu = 3), \tag{3.5}$$

where y_{i,t_n} is the measurement of gene i at time t_n and σ_i is the measurement variance for gene i. A t-distribution with three degrees of freedom is used as error model as a means of robust inference. The model cannot precisely represent the trajectories and the t-distribution can better accommodate the outlying measurements with respect to the model trajectories than the normal distribution.

For the absolute concentration data, the likelihood function is defined as

$$P(y_i|m\bar{n}_i) = t(y_i|\mu = \log_{10}(m\bar{n}_i), \sigma = \sigma_j, \nu = 3) \tag{3.6}$$

for transcripts, where y_i is the $\log_{10}$-transformed measurement of the concentration of transcript i and σ_j is the measurement variance for all transcripts in dataset j. For proteins the equation is identical but with m_i replaced by p_i. The likelihood is specified on a log scale as it is sufficient if the model captures the right order of magnitude of the measurement, rather than the precise concentration.

3.4.4. MODEL CHECKING

The model fit was investigated using the posterior predictive distribution and coefficients of determination.

The posterior predictive distribution is the probability distribution of a new set of data, given the model and the observed data. This distribution was approximated with the posterior Monte Carlo samples:

$$\begin{aligned} P(y^{\text{pred}}|y, \mathscr{M}) &= \int P(y^{\text{pred}}|\theta, \mathscr{M}) P(\theta|y, \mathscr{M}) d\theta \\ &= \frac{1}{N} \sum_{i=1}^{N} P(y^{\text{pred}}|\theta_i, \mathscr{M}) \end{aligned} \tag{3.7}$$

where $\theta_{1..N}$ are the Monte Carlo samples from the posterior distribution with each θ being a vector containing all model parameters, and $\mathscr{M}$ indicates the model.

The coefficients of determination for the time course data were calculated as

$$R^2 = 1 - \frac{\sum_{j=1}^{M} \sum_{i=1}^{N_i} (y_{j,i} - x(t_i))^2}{\sum_{j=1}^{M} \sum_{i=1}^{N_i} (y_{j,i} - \bar{y}_j)^2}, \tag{3.8}$$

where j indexes the time course gene expression datasets, i indexes the time points within that dataset, x is the modeled value at time t_i, $y_{i,j}$ is the data value and $\bar{y}_j$ the mean of the data in that experiment. This equation corresponds to a null model which has a separate mean for each experiment.

The reference R^2 was calculated by fitting a cubic spline to the data with the smoothing parameter selected through cross-validation, and then setting $x(t_i)$ equal to the resulting spline value at t_i. The smoothing spline was fitted using the R function smooth.spline with default settings.

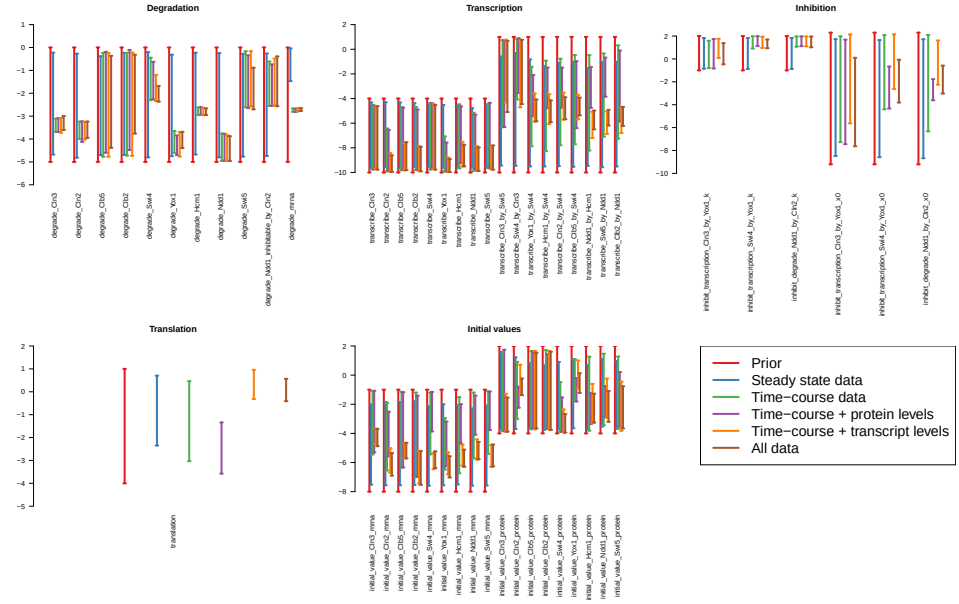


Figure 3.7: **Supplementary Figure 1**: Posterior 90% confidence intervals for all model parameters, as inferred using the absolute steady state data, relative time course data or both.

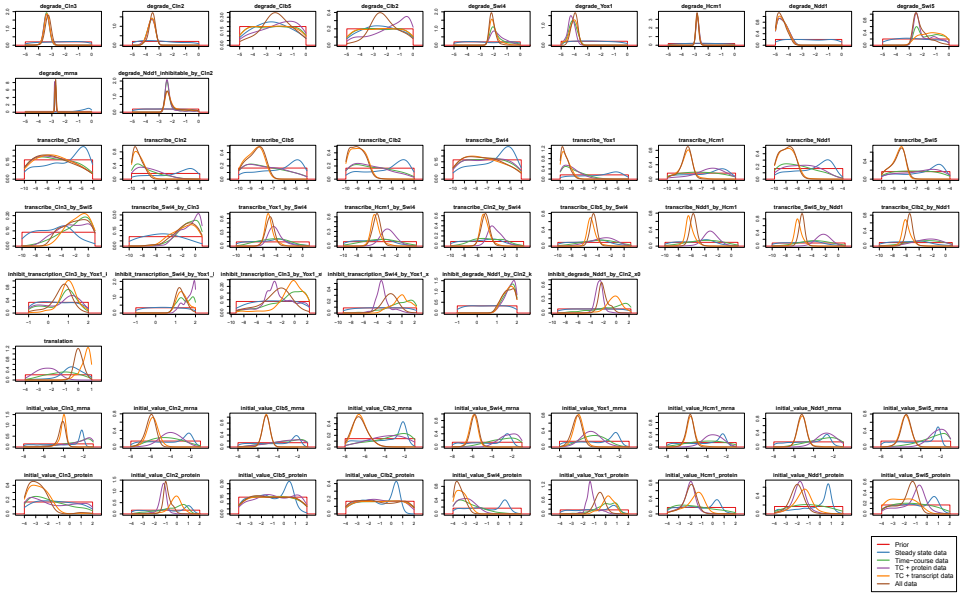


Figure 3.8: **Supplementary Figure 2**: Marginal probability distribution density estimates for the parameters estimated from the different datasets. Bandwidths for the kernel density estimates were selected using Sheather-Jones bandwidth selection.

4

INTEGRATIVE MODELING IDENTIFIES KEY DETERMINANTS OF INHIBITOR SENSITIVITY IN BREAST CANCER CELL LINES

Katarzyna JASTRZEBSKI
Bram THIJSSEN
Roelof J.C. KLUIN
Klaas DE LINT
Ian J. MAJEWSKI
Roderick L. BEIERSBERGEN
Lodewyk F.A. WESSELS

Parts of this chapter have been accepted for publication in Cancer Research.

ABSTRACT

CANCER cell lines differ greatly in their sensitivity to anticancer drugs. This variability arises due to different oncogenic drivers and drug resistance mechanisms operating in each cell line. Although many of these mechanisms have been discovered, it remains a challenge to understand how they interact to render an individual cell line sensitive or resistant to a particular drug. To better understand this variability, we first profiled a panel of thirty breast cancer cell lines, in the absence of drugs, for their mutations, copy number aberrations, mRNA and protein expression, protein phosphorylation, and response to seven different kinase inhibitors. We then constructed a knowledge-based, Bayesian computational model that integrates these data types and estimates the relative contribution of the various drug sensitivity mechanisms. The resulting model of regulatory signaling can explain the majority of the variability observed in drug response. The model also identifies cell lines with an unexplained response, and for these we then searched for novel explanatory factors. Among others, we found that the 4E-BP1 protein expression level – and not just the extent of phosphorylation – is a determinant of mTOR inhibitor sensitivity. We further investigated and validated this finding experimentally. We found that overexpression of 4E-BP1 in cell lines that normally possess low levels of this protein, is sufficient to increase mTOR inhibitor sensitivity. Taken together, our work demonstrates that combining experimental characterization with integrative modeling can be used to systematically test and extend our understanding of the variability in anticancer drug response.

4.1. INTRODUCTION

Breast cancer is the most commonly diagnosed cancer in women and the second leading cause of cancer-related death [1]. Decades of research have increased our knowledge of the molecular basis of this disease while recent large scale genomics studies have provided detailed information on mutations, copy number aberrations and gene expression across different tumor samples, including breast cancer [2, 3]. However, despite this increasing amount of knowledge, response rates for cancer treatment remain very low for many cancer subtypes, including breast cancer.

One of the challenges of cancer treatment is the genetic complexity of the disease, involving different oncogenic drivers or combinations thereof that allow the tumor to grow and proliferate. In some subtypes of breast cancer, the major oncogenic drivers are known, as is the case in HER2-amplified breast cancer, associated with the overexpression and aberrant activation of HER2-receptor signaling. Targeted treatment against HER2 indeed provides clinical benefit [4, 5]. However, intrinsic resistance is frequently encountered, and acquired resistance often develops in tumors which are initially sensitive [6]. A variety of drug resistance mechanisms to HER2-targeted therapies have been discovered in cell lines. For example, activation of the PI3K pathway resulting from mutations in PI3K [7], the loss of PTEN [8] or autocrine HGF signaling [9] have all been reported as mechanisms of drug resistance. However, due to the multitude of resistance mechanisms, which is further complicated by the cross-talk in downstream signaling, it is unclear to what extent each of these mechanisms is important for determining the sensitivity of a particular cell line or tumor. In other types of breast cancer, in particular

triple-negative breast cancer, the regulatory signaling which drives the growth of the tumor is even less clear, although PI3K/AKT pathway deregulation has been identified as a recurrent event [2].

Similar to patients, cell lines show a large degree of variability in drug response [10–12]. Understanding the heterogeneous response in cell lines is an important starting point for understanding patient response. But despite many efforts, fully explaining drug response in vitro remains a challenge. Computational modelling can be used as a tool capable of untangling the complexity of drug sensitivity and resistance across different cell lines. There have been various approaches to computational modelling of drug sensitivity. These approaches can be broadly divided into two categories: approaches using linear or black box statistical models [11, 13, 14], and approaches using more detailed mechanistic computational models [15, 16]. While both approaches have recovered known drug sensitivity mechanisms and identified several novel associations, they have several limitations. For the black box statistical models, such as elastic net regression [10, 11], random forests [13], support vector machines [13] or a clustering-based method named ACME [14], the models are not sufficiently detailed to capture how molecular characteristics affect drug sensitivity. For example, the interactions between molecular aberrations are not explicitly modelled. This precludes finding all but the strongest associations, despite the very large number of cell lines that were profiled. In addition, available knowledge of signaling pathways is not employed, which could increase the statistical power to find molecular mechanism that associate with drug sensitivity. In the more detailed, mechanistic computational models [15, 16] such background knowledge is used, however, in these cases the number of cell lines studied has been limited, and it is thus unclear to what extent the particular mechanisms are important for explaining the variability across a larger set of cell lines. In addition, these mechanistic modelling studies used only a single data type, for example (phospho)protein expression, limiting the insight into the impact of other molecular aberrations present in the cell lines examined.

To address these limitations, we set out to combine detailed computational modelling of drug sensitivity mechanisms with extensive measurements of multiple data types derived from a breast cancer cell line panel. We developed a combined experimental/computational modelling approach which can utilize background knowledge from the literature and integrate diverse types of data, including DNA sequencing, RNA sequencing, protein expression and protein phosphorylation, with drug response data. We subsequently employed the computational model to analyze how the regulatory signaling in each cell line influences response to each drug. The computational model can also be used to identify cases where drug response cannot be explained fully by the existing knowledge. In one case, this lead us to identify and confirm the level of expression of eukaryotic translation initiation factor 4E-binding protein 1 (4E-BP1) as a determinant of response to mTOR inhibitors in breast cancer cell lines.

Together, we show the utility of employing integrative computational modelling to combine prior knowledge with measurements of multiple molecular data types to systematically test and extend our understanding of drug response to kinase inhibitors.

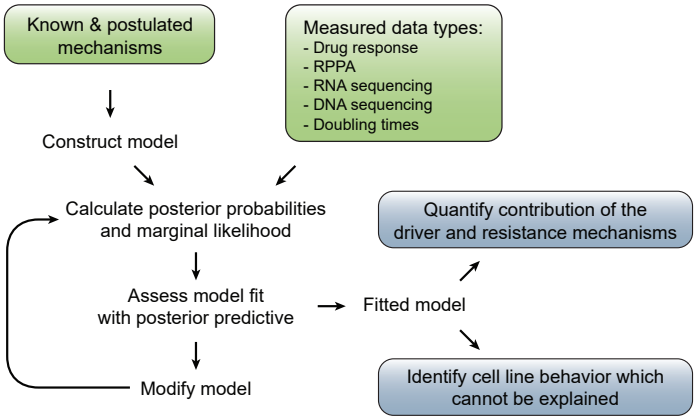


Figure 4.1: **Procedure used to construct the computational models.** We used the literature to construct and iteratively update the model until a good fit for the data was obtained.

4.2. QUICK GUIDE TO EQUATIONS AND ASSUMPTIONS

To integrate the different data types with knowledge of the regulatory signaling pathways, we created an integrative computational model using a modeling framework we call Inference of Signaling Activity (ISA). A challenge in constructing such a model is deciding which aspects of cell biology should be included and which can be omitted. To tackle this challenge, we developed the model iteratively using the procedure shown in Figure 4.1. We started out with a small, simple network including only the signaling nodes EGFR, ERK, AKT and a node depicting proliferation. We then surveyed the literature for signaling events, molecular mechanisms and recurrent mutations and copy number aberrations associated with breast cancer and known to be involved in determining drug sensitivity. We iteratively added more of these relevant mechanisms to the model. Specifically, for every mechanism we created a new model with additional nodes to represent the mutation, amplification or signaling molecule. At each iteration we tested the goodness of fit with the posterior predictive distribution (see section on testing the goodness of fit) and used the marginal likelihood to decide whether the newly added mechanisms should be retained. Finally, we stopped the process of model refinement when further additions no longer increased the marginal likelihood or when computation time grew impractically long.

The resulting model is shown in Figure 4.2, and includes growth factors, surface receptors, the MAPK and PI3K pathways, mutations and copy number aberrations which occur regularly in breast cancer, the kinase inhibitors and their targets and finally the proliferation of the cells. The signaling molecules in the model (the nodes) are linked using activation functions (the arrows), which describe how the signal is propagated between molecules.

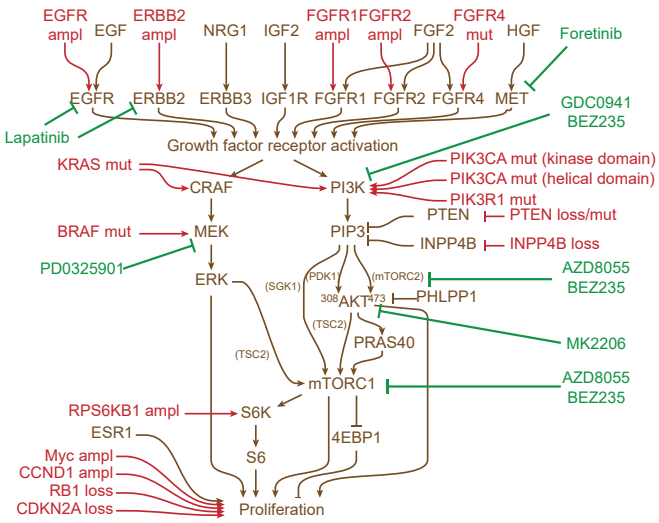


Figure 4.2: **Simplified overview of the computational model.** The graph shows the signaling nodes in brown, the mutations and gene losses and gains in red and the kinase inhibitors in green

The activity of a signaling molecule i in cell line j, $A_{i,j}$, is calculated as follows:

$$A^*_{i,j} = E_{i,j}(b_i + \sum_{k \in \text{parents}_A(i)} s_{k,i} A_{k,j} + \sum_{k \in \text{parents}_M(i)} s_{\text{mut},k,i} M_{k,j} \tag{4.1}$$

$$A_{i,j} = \max(\min(A^*_{i,j}, 1), 0). \tag{4.2}$$

The activity $A^*_{i,j}$ is a linear combination of a base activity b_i, the upstream signaling molecules ($A_{k \in \text{parents}(i),j}$) with signaling strength $s_{k,i}$ from molecule k to molecule i, and the upstream mutations ($M_{k \in \text{parents}(i),j}$) with signaling strength $s_{\text{mut},k,i}$, which is then multiplied by the expression of the signaling molecule itself ($E_{i,j}$). The resulting value is clamped between 0 and 1 to give interpretable values that are comparable throughout the network. For a full description, see the Supplementary Materials and Methods.

When an inhibitor is applied, the activities of the targets of the inhibitor are multiplied by the drug effect, a value between 0 and 1 which is calculated with a three-parameter logistic function (see Supplementary Material).

Figure 4.3 shows the structure of the model in template notation for a small part of the network. To illustrate signal propagation between molecules, consider, for example, S6K. The activity of S6K, represented in the figure by the variable *S6K signal*, is a function of the activity of the upstream kinase mTORC1 (*mTORC1 signal*), as well as of the total amount of S6K in the cell (*S6K expression*). The activation function for calculating S6K signal has several parameters, namely the basal activity (*S6K base signal*), the strength of the link between mTORC1 and S6K (*mTORC1->S6K strength*). Importantly, the parameters, represented by dashed circles, are shared by all cell lines. That is, the values of the parameters are the same for all cell lines. So, while each cell line can have a different amount of mTORC1 activity, a given amount of mTORC1 signal always gives rise to the same amount of input signal to S6K. Note that the model is not intended to be a

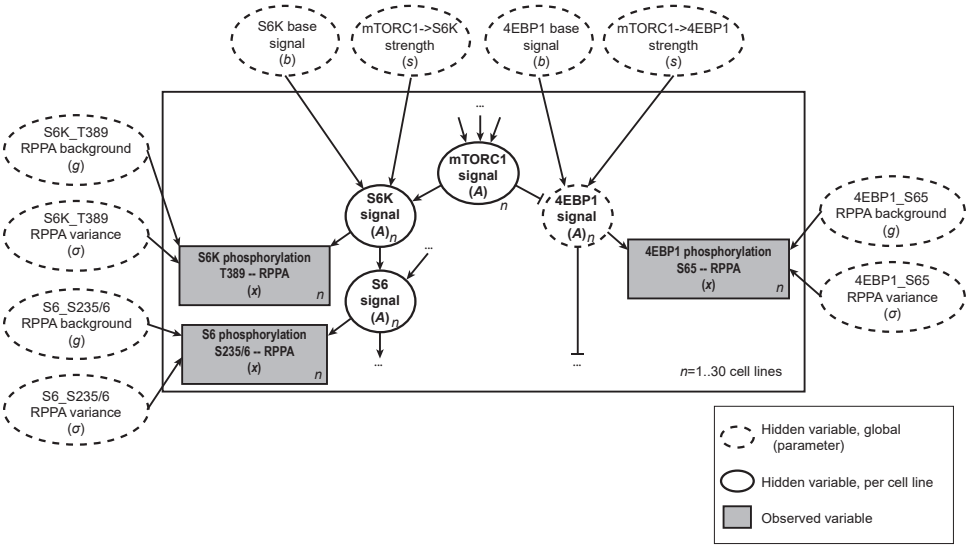


Figure 4.3: **Part of the computational model in template notation.**

precise description of all chemical reactions within the network, but rather an abstract representation of relevant regulatory signaling.

We do not explicitly include feedback signaling events in the model. Although feedback signaling is an important aspect of cellular regulatory networks, we find that even without feedback signals the relative viability after drug treatment can still be described well. This is most likely due to the fact that the activity of feedback loops may still be indirectly reflected in the steady state signal strengths.

The model provides a framework that allows the integration of all data types to infer the parameter values and signaling activities. Some variables are observed directly, for example the presence of a mutation, and in this case the value of the node is set to the observed value directly. Other variables, namely the protein activity, the untreated growth rate and the viability after drug treatment, are only observed indirectly. For example, the amount of S6K phosphorylation only indirectly reflects S6K activity. In these cases, we add a random variable, $x_{i,j}$, which models the measurement value and which is dependent on the hidden model variables. We then use a likelihood function for this random variable $x_{i,j}$ to infer the parameters. We use a Student's t-distribution and center the distribution on $x_{i,j}$. The likelihood can be expressed as

$$P(y_{i,j,k}|\boldsymbol{\theta}) = t(y_{i,j,k}|\mu = x_{i,j}(\boldsymbol{\theta}), \sigma = \sigma_i, \nu = 3) \tag{4.3}$$

$$x_{i,j}(\boldsymbol{\theta}) \begin{cases} g_i + (1 - g_i) A_{i,j}(\boldsymbol{\theta}) & \text{for RPPA data} \\ r_j(\boldsymbol{\theta}) & \text{for growth data} \\ D_j(\boldsymbol{\theta}) & \text{for drug response data,} \end{cases} \tag{4.4}$$

where $x_{i,j}$ models the data as described above, $y_{i,j,k}$ is the measurement data, k indexes the biological replicate measurements, $A_{i,j}$ is the signaling activity defined above

in Equation 1.1, r is the untreated growth rate, and D is the relative viability (the variables r and D are defined in the Supplementary Material and Methods). $\boldsymbol{\theta}$ represents a vector of all model parameters in Equation 1.1.

To infer the parameter values and signaling activities, we used Bayesian statistics. For all parameters, prior probability distributions were specified to describe what values the parameters might assume a priori. Subsequently, two variants of Monte Carlo sampling were used to calculate the posterior probability distribution of all parameters.

The major assumptions are thus as follows. First, we assume that the signaling activity is a linear combination of its upstream inputs. Similarly, proliferation is a linear combination of the activity of several effector signaling molecules. Second, we do not explicitly model feedback events, but assume that their effect is indirectly reflected in the steady state signal strengths. Third, we assume that a signaling molecule with a certain amount of activity always gives rise to the same amount of input activity to its downstream nodes in every cell line, as the signaling strength $s_{k,i}$ is constant across cell lines. However, note that cell lines can still have widely different signaling activity due to variation in which mutations are present and having different gene expression levels.

4.3. MATERIALS AND METHODS

Detailed materials and methods are provided in the Supplementary Materials and Methods.

4.3.1. CELL LINE PANEL

A panel of 30 breast cancer cell lines was assembled from various sources, the details and growth conditions of which are listed in Supplementary Table 1. Cell lines were authenticated by suppliers using Short Tandem Repeat profiling. The SK-BR-7 cell line was obtained from an internal NKI cell bank and authenticated by STR profiling. We further confirmed the identity of the lines by comparing their mutation profiles (found by DNA sequencing) with those reported in COSMIC [17]. Upon receipt, all lines were first expanded and early passage stocks frozen in liquid nitrogen. Lines were kept in culture for no more than 3 months, after which a new cell aliquot was obtained from frozen stock if needed. All cell lines were tested in-house and found to be negative for mycoplasma. Doubling times for each cell line were estimated by fitting exponential curves to confluence measurements obtained using an IncuCyte FLR/ZOOM instrument (Essen Bioscience).

4.3.2. DRUG RESPONSE ASSAYS

Prior to carrying out drug response assays, cell line seeding densities were optimized. Cells seeded as for the cell doubling time experiments were assessed at the 96 h endpoint for percentage confluence, and incubated with CellTiter-Blue (CTB; Promega) for a measure of metabolic activity. This was to ensure that cell lines did not exceed 90% confluence at assay endpoint and that the CTB signal at this density was not saturated. Seeding densities used for each cell line are listed in Supplementary Table 1.

For drug response assays, cells were seeded at the optimized density and 24 h later treated with a 10-point 1:3 dilution series of a number of inhibitors using a Microlab

STAR workstation fitted with 8 x 1000µl channels and 96-probe head (Hamilton): AZD8055, top dose 3 x 10^{-5} M; BEZ235, 1 x 10^{-5} M; GDC0941, 3 x 10^{-5} M; MK2206, 3 x 10^{-5} M; PD0325901, 3 x 10^{-5} M; Lapatinib, 3 x 10^{-5} M; Foretinib, 3 x 10^{-5} M (all from Selleckchem). Each condition, including an untreated negative control and a phenyl arsine oxide (1 x 10^{-6} M) treated positive control, were set up in technical quadruplicate. Following 72 h incubation, cells were stained with CTB (1:30 dilution) for 4 h and the signal measured using an Envision spectrophotometer (Perkin Elmer). In the case of the validation experiments with HCC1806 and HCC1937 cell lines expressing 4E-BP1 or GFP constructs, cells were treated in a 9-point 1:3 dilution series of AZD8055, BEZ235 or GDC0941 using a HP D300 Digital Dispenser (Hewlett-Packard), while all other experimental conditions remained the same. Each assay was carried out in biological triplicate. Each replicate of a dose response experiment was further analyzed by normalization to the negative and positive control (the normalized data are provided in Supplementary Table 6) and fitting to a four-parameter sigmoid function that allowed for the calculation of the 50% inhibitory concentration (IC50, dose at which viability is 50% of the untreated control). The IC_{50} estimates are provided in Supplementary Table 7. For model inference, full dose response curve data were used.

4.3.3. LONG-TERM DRUG RESPONSE ASSAYS

HCC1806 parental, GFP- and 4E-BP1-expressing cells were seeded at 600 cells/well, while the HCC1937 panel was seeded at 1200 cells/well, in 96 well plates. Cells were treated, 24 h after seeding, with a 9-point 1:3 dilution series of AZD8055 (top dose 3.3 x 10^{-6} M) or BEZ235 (1.1 x 10^{-6} M) using a HP D300 Digital Dispenser (Hewlett-Packard). Each condition, including an untreated negative control and a phenyl arsine oxide (1 x 10^{-6} M) treated positive control, were set up in technical duplicate. Media and drugs were changed every 3-4 days over a period of 10-11 days of treatment. Cells were then washed with PBS, fixed with 3.7% formaldehyde/PBS and stained in 0.1% crystal violet solution. Images of dried, stained cells were digitized on a Perfection V750 PRO scanner (Epson).

4.3.4. MOLECULAR CHARACTERIZATION

Steady state RNA and protein expression was determined from cells seeded in 60 mm dishes and grown for 48 h (seeding densities are listed in Supplementary Table 1). RNA expression was determined using RNA sequencing by the NKI Genomics Facility and protein expression was determined in biological triplicate using RPPA analysis by the MD Anderson Cancer Center RPPA Facility. Genomic DNA samples were obtained from pellets of 0.5 x 10^6 cells. RNA sequencing data is available at ArrayExpress, reference E-MTAB-4801, and the normalized read counts are provided in Supplementary Table 8. DNA sequencing data is available at the European Nucleotide Archive, reference PRJEB14120. The RPPA data is included as Supplementary Table 9.

4.3.5. CAP-BINDING PULL DOWN ASSAYS

Cells were seeded in 100 mm dishes (BT549 and CAL-120 at 2.5 x 10^5; Hs 578T at 3 x 10^5; HCC1806 at 4 x 10^5; HCC1937 at 6.25 x 10^5) and cultured for 48 h, then treated with AZD8055 (1.11 x 10^{-7} M), BEZ235 (3.7 x 10^{-8} M) or vehicle (DMSO) for a further 24 h. Cells were washed once with ice-cold PBS and lysed in lysis buffer (25 mM Tris-

HCl, pH 7.6, 1% Triton X-100, 1 mM DTT) supplemented with cOmplete protease and phosSTOP phosphatase inhibitor cocktails (Roche). Lysates were cleared and assayed for protein concentration, then total protein samples were prepared using 20 µg of protein lysate. Cap pull-down samples were prepared by combining 50 µg of total lysate with 20 µl pre-washed m7GTP-agarose (Jena Bioscience), made up to a total volume of 500 µl with lysis buffer and tumbled at 4°C overnight. The following day, cap pull-downs were washed 3 x in ice-cold lysis buffer, then heated at 70°C for 10 min in 20 µl 1x Novex® LDS Sample Buffer and Sample Reducing Agent. The eluate from the cap pull-downs as well as the total protein control samples were then immediately separated on Novex® 4-12% gradient gels and immunoblotted using primary antibodies to 4E-BP1, eIF4G and HSP90 (for total lysates samples only), then reprobed to detect eIF4E protein.

4.3.6. GENERATION OF 4E-BP1 OVEREXPRESSING CELL LINES

pLX304-4E-BP1 was obtained from the CCSB-Broad Lentiviral Expression Collection, while the pLX304-GFP control construct was generated as outlined previously [18]. To produce lentiviral particles, HEK293T cells were co-transfected with the pLX304-4E-BP1 or -GFP bearing construct and a lentiviral packaging mix (pRSV-Rev, pMDLg/pRRE, pCMV-VSV-G; Addgene) using Polyethylenimine (PEI, Linear MW 25,000; Polysciences Inc.). Media was changed 24 h after transfection. After a further 24 h, viral supernatant was collected and 0.45 µm-filtered. HCC1806 and HCC1937 cells were transduced in the presence of hexadimethrine bromide (Sigma-Aldrich) and following 48 h selected using blasticidin.

4.3.7. PROLIFERATION OF 4E-BP1 OVEREXPRESSING CELL LINES

HCC1806 parental, GFP- and 4E-BP1-expressing cells were seeded at 800 cells/well, while the HCC1937 panel was seeded at 1000 cells/well, in 384 well plates with 4-6 replicates per condition. Proliferation was monitored using the IncuCyte ZOOM instrument (Essen Biosciences).

4.3.8. MODEL, DATA INTEGRATION AND INFERENCE

An overview of the computational model is given in the Quick Guide to Equations and Assumptions. The model inference was done using BCM [19]. A detailed description of all equations, data preprocessing and inference algorithms is given in the Supplementary Materials and Methods.

4.4. RESULTS

4.4.1. ESTABLISHING AND CHARACTERIZING A BREAST CANCER CELL LINE PANEL

We set out to establish an integrative computational model capable of explaining observed therapeutic responses based on molecular measurements. To this end, we sourced and comprehensively characterized thirty breast cancer cell lines (Figure 4.4 and Supplementary Table 1). Given the need for targeted treatment options for the triple negative breast cancer subtype, the panel was enriched for triple negative cell lines (eighteen), with four ER+, four HER2+ and four ER+/HER2+ cell lines included to represent the other

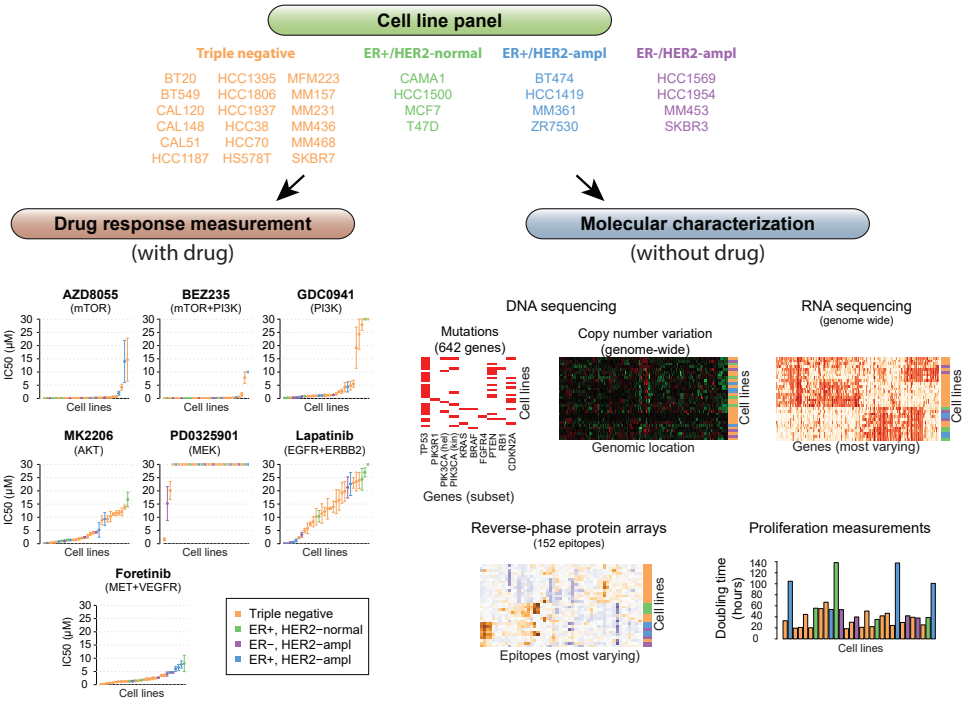


Figure 4.4: **Schematic of the composition and characterization of the panel of breast cancer cell lines.** Thirty breast cancer cell lines were sourced and expanded, representing four major classes of breast cancer subtypes - eighteen triple negative, four ER+, four HER2+ and four HER2+/ER+ cell lines. These cell lines were then assayed for their response to seven kinase inhibitors (bottom left panel - summary of response data with respect to the IC50 metrics per inhibitor and cell line) as well as characterized on a molecular level using DNA capture and mutation sequencing, RNA sequencing, proteomics (RPPA analysis) and growth rate assays. In the figures and all supplementary data, abbreviated cell line names are used; in particular MM stands for MDA-MB. Expanded views of the data plots are included in Supplementary Figures 1-6.

major subtypes.

The panel was characterized for response to seven kinase inhibitors, including AZD8055 (mTOR inhibitor), BEZ235 (dual mTOR/PI3K inhibitor), GDC0941 (PI3K inhibitor), MK2206 (AKT inhibitor), PD0325901 (MEK inhibitor), lapatinib (dual EGFR/HER2 inhibitor) and foretinib (cMET/VEGFR2 inhibitor). The sensitivity of each cell line to these inhibitors was determined in 10-point, 72 h dose response assays, in biological triplicate and technical quadruplicate (summarized by IC_{50} values in Figure 4.4 and Supplementary Figure 4.4). The drug sensitivity measurements largely agree with those obtained in the Genomics of Drug Sensitivity in Cancer screen [10, 17, 20] (Supplementary Figure 15). However, our use of a more focused panel of thirty cell lines and seven drugs allowed us to obtain more precise measurements (Supplementary Figure 16).

In addition to response data, we profiled the panel for mutation and copy number by DNA-seq, RNA expression by RNA-seq, protein expression and phosphorylation by reverse-phase protein array (RPPA) as well as proliferation rate under untreated, steady-

state growth conditions (Figure 4.4, bottom right panel; see Supplementary Figures 2 - 6 for enlarged versions of the graphs). This molecular characterization is done in the absence of drug treatment. The cell lines harbor a range of genetic events which occur in breast tumors and are present at comparable frequencies (see Supplementary Table 13 for a comparison of mutation frequencies with tumors from The Cancer Genome Atlas [2]). This cell line panel thus represents a relevant model system for the genetic diversity in breast tumors.

4.4.2. FITTED MODEL PROVIDES ESTIMATES OF REGULATORY SIGNALING BASED ON ALL AVAILABLE DATA TYPES

To first understand which signaling is relevant for each drug, we developed a modeling framework, Inference of Signaling Activity (ISA), to infer the signal strengths and signaling activities from all available data, as described in the Quick Guide to Equations and Assumptions and further detailed in the Supplementary Materials & Methods. We constructed a literature-based model and first fitted this to the response data for each drug separately, in conjunction with the molecular data measured in untreated cells. In other words, we first searched for values of the signaling strengths (as well as the other parameters) that can explain the variability of response across all cell lines, but for each drug separately. Although all of the interactions included in the model are well documented (see Supplementary Table 4 and 5), their relative contribution or significance is not known. For example, activating mutations in PIK3CA, the loss of PTEN or the expression of growth factors can all lead to activation of the PI3K pathway. However, it is unclear whether their effects are equally important, and if not, which of them has a stronger effect in a particular context.

Figure 4.5A illustrates the model estimates of signaling strengths (the links between signaling molecules) for lapatinib treatment. Values of the strength parameters indicate which signaling connections are important for propagating an oncogenic signal down to the proliferation node. For example, the link between ERBB2 amplification and ERBB2 activation has a strong peak at non-zero values (the density plot of *ERBB2amp->ERBB2*), thus indicating that the ERBB2 amplification gives rise to a proliferative signal. It is well known that amplification of ERBB2 and the resulting overexpression and auto-activation of this receptor provides a strong proliferation signal [21], and that this signal can be inhibited by lapatinib [22]. The model provides estimates for the downstream signaling (e.g. indicating that ERBB2 signals more to PI3K than to CRAF) and for the contribution of each of the resistance mechanisms. PIK3CA mutations indeed contribute to the proliferative signal (*PIK3CAhelical->PI3K*), and the model predicts that PIK3R1 mutations may have a similar effect, as we see that the parameter describing how much a PIK3R1 mutation activates PI3K tends towards higher values, and is most likely non-zero (*PIK3R1->PI3K*). However, the uncertainty in this parameter is large, because there is only one cell line that carries such a mutation, and consequently this parameter is only weakly constrained. As a last example, the contribution of HGF autocrine signaling(9) is represented by the parameter controlling how strongly expression of HGF leads to activation of the MET receptor (*HGF->MET*). The posterior probability distribution of this parameter closely follows the prior, indicating that this parameter, and thus the importance of this potential resistance mechanism, cannot be determined from the cur-

rent data. Together, this shows that the ISA modelling approach can be used to infer the contribution of different components driving sensitivity and resistance from all available data, while taking into account whether the parameters are identifiable.

We can use the estimates of the signaling activities (values of the nodes) to further explore the difference in signaling flow and drug response between cell lines. Figure 4.5B shows the estimates for ERBB2 activity and PIP3 activity in the lapatinib-treated condition scattered against the untreated condition. From the left panel, it can be seen that only the eight ERBB2-amplified cell lines show ERBB2 signaling activity, and that this activity is reduced upon lapatinib treatment in all these cell lines. In the right panel, we can see that in the lapatinib-sensitive cell lines, especially the most sensitive ones BT-474, SK-BR-3 and ZR-75-30, the reduction in ERBB2 activity also leads to a strong reduction in the PIP3 signal, whereas in the other cell lines, PIP3 signal persists, especially for the lapatinib-resistant cell line HCC1569 (see Supplementary Figure 1 for the drug sensitivity estimates). Two non-ERBB2-amplified cell lines also have a reduced PIP3 signal upon lapatinib treatment, including T-47-D and HCC1806, which stems from their inferred EGFR activity. This illustrates the utility of the model, given that there were no molecular measurements collected in the treated conditions, and these signaling estimates are inferred from the untreated molecular data combined with the relative viability data in the treated condition. A comparison of the inferred model estimates with molecular measurements in treated condition is described later, after all model adaptations have been considered.

As a second example, for the mTOR inhibitor AZD8055, we find several factors that are associated with response. As expected, PIK3CA mutations are strongly activating in this context and cell lines are apparently dependent on this activation (see Figure 4.5C, *PIK3CAkinase/helical->PI3K* and *mTORC1->proliferation*), which has previously been shown [23]. Additionally, we find that MYC activation, as a result of gene amplification, can provide a resistance mechanism to this mTOR inhibitor (*MYCamp->proliferation*), providing another validation of our approach [24].

To facilitate the further exploration of all the signaling estimates, we generated an interface which is available at http://ccb.nki.nl/software/BCCL_KI_response_model/. This website displays the signaling strengths in each cell line upon exposure to drugs. Figure 4.5D shows an example of BT-474 treated with lapatinib. In this case the model indicates that, for example, the MAPK pathway is barely involved, that there is no drug resistance provided by this cell line's RPS6KB1 amplification, and that instead, lapatinib mainly inhibits the PI3K pathway.

4.4.3. USING THE POSTERIOR PREDICTIVE DISTRIBUTION TO TEST THE GOODNESS OF FIT

While the signaling estimates appear to be reasonable, it is useful to have a systematic test of how well the fitted model describes the data. For this we used the posterior predictive distribution, which describes a new, predicted dataset based on the fitted model. We can overlay this predicted dataset on the observed measurements to have a convenient way of identifying which measurements can and cannot be explained by the model. Note that the posterior predictive distribution is not used as a measure of out-of-sample prediction, that is, it does not test how well the model predicts the behavior

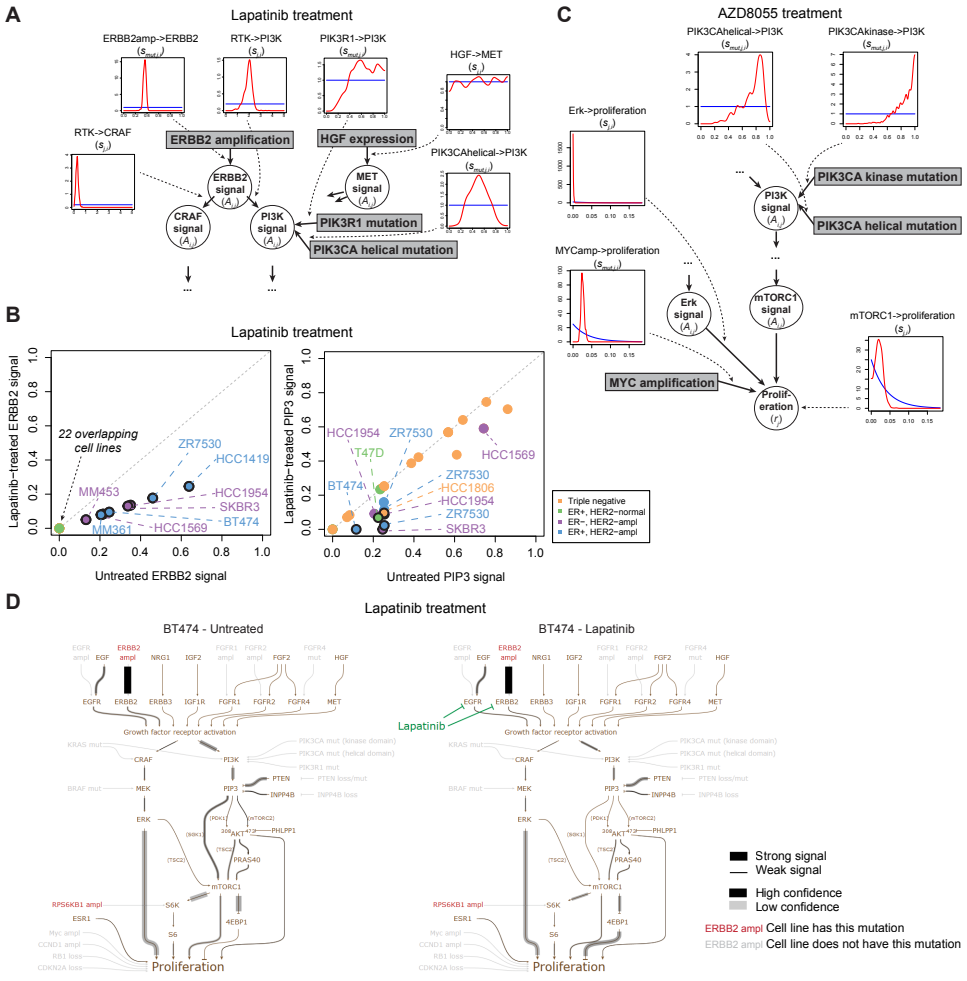


Figure 4.5: **Model estimates of signaling in the context of treatment with either lapatinib or AZD8055.** (A) Marginal posterior probability densities for several of the model parameters in the context of lapatinib treatment. Only the relevant parts of the model are shown. The densities are estimated using kernel density estimates with Sheather-Jones bandwidth selection. (B) Estimates of the activity of two signaling molecules, ERBB2 and PIP3, in untreated and lapatinib-treated conditions. A black circle around a point indicates significant difference (posterior probability > 0.975 for the lapatinib-treated signal being less than the untreated signal). Error bars are not shown here; a version with error bars indicating the 90% confidence intervals is included as Supplementary Figure 11. (C) Like (A), but now in the context of AZD8055 treatment. (D) Overview of all signaling activity estimates in the BT-474 cell line, in untreated and lapatinib-treated conditions. Thickness of the signaling bars indicates the signal strength multiplied by the parent activity (i.e. $s_{k,i} A_{k,j}$), and the scale of grey indicates uncertainty.

of unseen cell lines. Instead it is used as a measure of goodness of fit, allowing an exploration of whether the model can describe the behavior of the cell lines at hand.

Figure 4.6A shows the posterior predictive distribution overlaid on the measurement data for lapatinib. It is clear that the present model can accurately describe the relative proliferation of the cell lines as a function of drug concentration for almost all cell lines. For example, the sensitivity of the ERBB2-amplified line BT-474 and the resistance of the PIK3CA-mutated and ERBB2-amplified line MDA-MB-361 (MM361) can both be recapitulated by the model. Overviews of all posterior predictive checking for the drug response and phosphorylation data is supplied in Supplementary Data 1.

4.4.4. SEARCHING FOR ADDITIONAL EXPLANATORY FACTORS OF DRUG SENSITIVITY REVEALS NOVEL ASSOCIATIONS

While the model explained most of the drug response variability for lapatinib, we noticed that for some drug-cell line combinations, the fit was not as precise. For example, for foretinib, an inhibitor of c-Met and VEGFR2, we noticed that one cell line in particular, MFM-223, was much more sensitive than the model could describe (Figure 4.6B). We therefore investigated the experimental data to find a possible reason for this discrepancy. A discrepancy in a single cell line is not sufficient to apply statistical tests, but we did note that this cell line has a strong FGFR2 amplification. We therefore searched the literature to see whether there is a connection and found that foretinib has in fact been reported to inhibit FGFR2 in addition to its original design targets [25]. When this additional target of foretinib is added to the model, we indeed obtain a significantly improved fit (Figure 4.6B). The sensitivity of other cell lines like CAL-51 and HCC-1187 to foretinib is still not explained exactly, but we have not found other potential explanations for this in the data or literature, and therefore further studies would be needed to address this.

We also noticed that the model was not able to explain the response of some cell lines to mTOR inhibitors (see Figure 4.6C), but were unable to find additional mechanisms in the literature which could explain these discrepancies. Several cell lines are sensitive to these inhibitors even though they do not possess any of the factors known to cause sensitivity, and conversely some cell lines are resistant despite having such sensitizing factors. For example, while BT-549, HCC1395 and HCC1937 have all lost PTEN expression, only BT-549 is sensitive to BEZ235 treatment.

We therefore further interrogated the dataset to find additional drug sensitivity mechanisms with which we could extend the model to better explain the experimental observations. With multiple sensitive cell lines we can use statistical tests. We divided the cell line panel in groups of sensitive and resistant lines using Gaussian mixture modelling, and tested whether any genes were differentially expressed between these groups at either the RNA or protein level. The full lists of differentially expressed genes for all drugs are given in Supplementary Table 14. To further filter the differentially expressed genes, we also calculated their distance to the signaling molecules included in the model using protein-protein interaction networks [26]. This provided potential candidate regulators that are not only differentially expressed, but are also functionally closely related to the signaling molecules in the model (listed in Supplementary Table 14). For both mTOR inhibitors (AZD8055 and BEZ235) the protein expression level of 4E-BP1, in addition to 4E-BP1 phosphorylation level, showed the strongest differential expression (Figure 4.6D). At

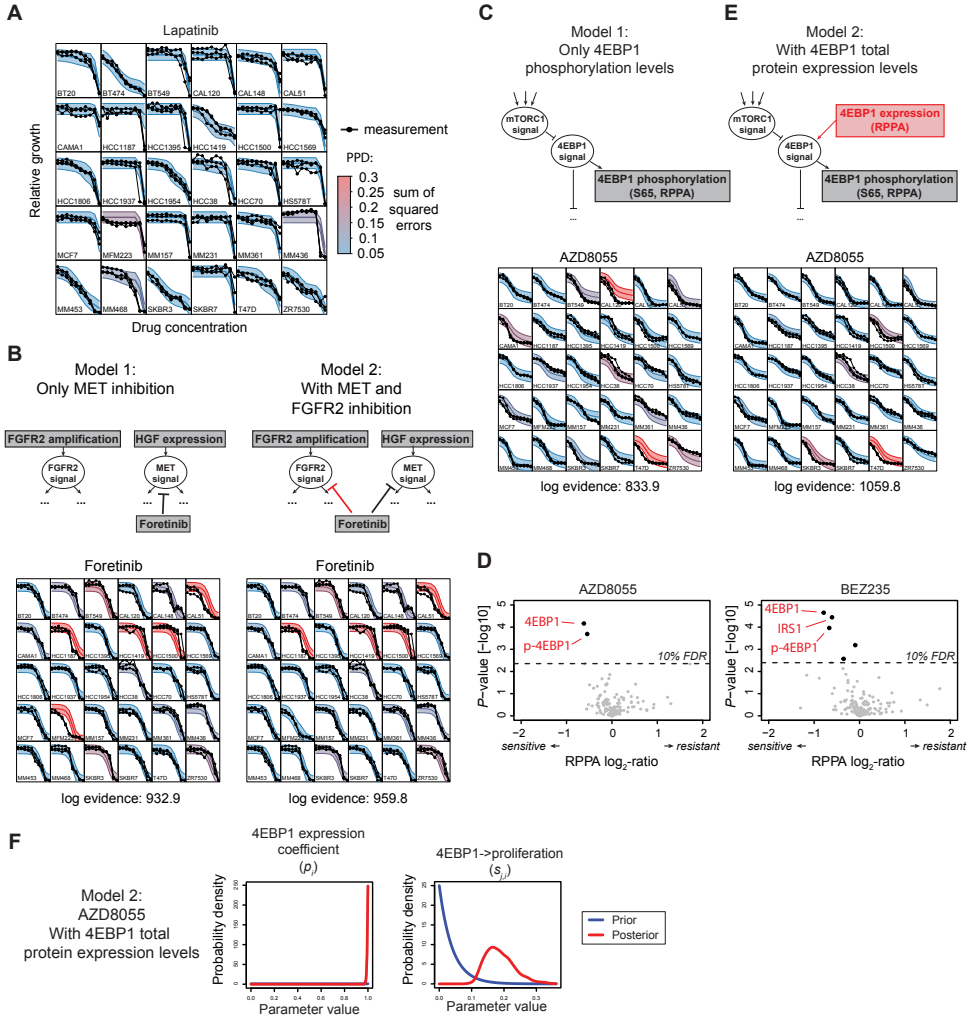


Figure 4.6: **Goodness-of-fit testing and model expansions.** (A) The 90% confidence interval of the posterior predictive distribution for lapatinib drug response (shaded area) is overlaid on the measurement data (in black). The posterior predictive distribution for each cell line is colored by the discrepancy with the data, quantified using the sum of squared errors over the 8 lowest concentrations (the 2 highest concentrations were excluded because discrepancies at such high concentrations are likely due to various off-target effects) (B) Comparison of two iterations of the model, when fitting the foretinib drug response. (C) Iterations of the model without protein expression levels as explanatory factor, when fitting the AZD8055 drug response. (D) Volcano plot for the association of the reverse-phase protein array data with sensitivity estimates to the two mTOR-inhibitors. For both inhibitors, 4E-BP1 protein expression levels correlate significantly with sensitivity. (E) New model iteration with protein expression included as explanatory factor (F) Posterior probability of two parameters controlling 4E-BP1 signaling in the expanded model. The 4E-BP1 expression coefficient, shown in the top panel, controls how strongly the 4E-BP1 protein expression level limits the signal through 4E-BP1. The 4E-BP1->proliferation parameter, shown in the bottom panel, controls how strongly the 4E-BP1 signal affects proliferation.

the RNA level, differential expression of EIF4EBP1 was also associated with BEZ235 response. Together, these data indicated that cell lines with high expression of this protein are more sensitive to mTOR inhibitors (see Supplementary Figure 12).

To test whether the inclusion of this factor provides a better explanation for the sensitivity of some of the lines to mTOR inhibitors, we expanded the model to include the protein expression levels of 4E-BP1. Although 4E-BP1 as a node was already included as a downstream target of mTORC1 in our previous models, only its phosphorylation state, not its protein expression level, was taken into account. Specifically, in Equation 1.1, the variable $E_{i,j}$ was previously based only on the binarized RNAseq expression data, and we modified this to include the RPPA protein expression levels (see Supplementary Materials and Methods for details). Figure 4.6E shows the model with the protein expression levels of 4E-BP1 included, while Figure 4.6C shows the results without 4E-BP1 included. The posterior predictive checking (bottom panel) clearly shows that the expanded model provides an improved fit to the data for multiple cell lines (especially for CAL-120, BT-549, CAMA-1 and ZR-75-30), while not compromising the fit of other cell lines. The log Bayes factor between the two models (the difference in log evidence) is 226 in favor of the expanded model, where a log Bayes factor greater than 5 indicates very strong evidence [27]. This indicates that the difference is highly significant and that the improvement in fit is not merely the result of adding more free parameters. Figure 4.6F shows the posterior probability of two parameters of 4E-BP1 signaling in the expanded model. The 4E-BP1 expression coefficient, shown in the top panel, describes how strongly the protein expression level of 4E-BP1 affects the amount of signal transmitted. Since the value of this parameter is very high, it indicates that the total protein expression level is a strong limiting factor for 4E-BP1 signaling in response to mTOR inhibitors. The bottom panel shows the parameter controlling the strength of 4E-BP1 signaling to proliferation, and as this is also found to be non-zero with high certainty, it implies that the 4E-BP1 signal is important for determining proliferation rate under mTOR inhibitor treatment.

The computational analysis therefore predicts that the protein expression level of 4E-BP1 is an important factor in explaining mTOR inhibitor sensitivity, in addition to the already known factors determining sensitivity and resistance.

4.4.5. EXTERNAL VALIDATION AND JOINT-DRUG MODEL

The model uses pre-treatment molecular data and measurements of relative viability after drug treatment to infer signaling activities after drug treatment. To gain more confidence in the signaling estimates, we compared the estimates with measurements of protein phosphorylation of cells while on treatment [28], and found that they generally agreed (see Supplementary Note 1 in section 4.5.22). We also constructed a reduced model, which could be fitted to the response data of all seven inhibitors at the same time (see Supplementary Note 2 in section 4.5.23). Finally, using an extended version of the modeling formalism [29], we confirmed that even though feedback signaling is likely to be active, the inclusion of such feedback loops does not affect how well the variability in drug response can be described (see Supplementary Note 3 in section 4.5.24).

4.4.6. THE PROTEIN EXPRESSION LEVEL OF 4E-BP1 IS A DETERMINANT OF MTOR INHIBITOR SENSITIVITY

Intrigued by the model prediction that 4E-BP1 protein expression is associated with mTOR inhibitor response, we investigated the biological effect of 4E-BP1 expression directly. For this, we turned to a subset of our panel of breast cancer cell lines that showed differences in response to AZD8055 and BEZ235 (Figure 4.7A). These included three of the most mTOR inhibitor sensitive cell lines (BT-549, CAL-120 and Hs 578T) all bearing a gain in the EIF4EBP1 gene-containing genomic region (Supplementary Figure 7) which also express high levels of 4E-BP1 protein (Figure 4.7B and Supplementary Figure 8), and two insensitive cell lines (HCC1806 and HCC1937) that do not harbor a gain of the EIF4EBP1 locus and express low levels of 4E-BP1 protein. Given that high 4E-BP1 expression may drive cells to recalibrate signaling in the pathway by increasing the expression and/or activity of mTOR, we investigated this possibility further. We first checked whether expression of 4E-BP1 and mTOR were correlated (Supplementary figure 19A). While three of the most highly 4E-BP1 expressing cell lines do show an increase in mTOR expression at the protein level, in the lines chosen for our functional studies, only CAL-120 shows elevated mTOR expression. The remaining four lines (BT549, HS578T, HCC1806, HCC1937) show comparable expression of mTOR, despite vast differences in both 4E-BP1 expression and response to mTOR inhibitors. At the RNA level (Supplementary Figure 19B), these associations were lost, suggesting that a concurrent post-transcriptional upregulation of mTOR is not a general mechanism of mTOR inhibitor sensitivity in 4E-BP1 overexpressing cells. We then investigated mTOR signaling in more detail by analyzing the phosphorylation and protein expression levels of several members of the PI3K, mTOR and MAPK pathways following 24 h treatment with two mTOR inhibitors AZD8055 and BEZ235 in these five cell lines. This showed that both compounds had effective on-target activity, leading to reduced phosphorylation of AKT (S473), S6 (S235/236) and 4E-BP1 (S65) across all five lines, at similar compound concentrations (Figure 4.7B). A minor compensatory increase in ERK phosphorylation was detected following inhibitor treatment. These data suggest that the difference in mTOR inhibitor sensitivity between these five cell lines is not caused by a difference in the compounds' ability to inhibit mTOR signaling in these lines.

To uncover the mechanism via which 4E-BP1 protein expression levels could affect response to mTOR inhibitors, we investigated the effect of inhibitor treatment on the formation of the eIF4F translation initiation complex (extensively reviewed in [30, 31]). As illustrated in the schematic in Supplementary Figure 14A, the eIF4F translation complex is composed of the eIF4E and eIF4G proteins, among others. 4E-BP1 is known to negatively regulate this complex by binding and sequestering the eIF4E subunit. This displaces eIF4G from binding to eIF4E and as a result the eIF4F complex cannot initiate cap-dependent translation. The sequestering of eIF4E by 4E-BP1 is, however, inhibited when 4E-BP1 is phosphorylated by mTORC1 on several sites, which is the case when nutrients and growth factors are not limiting. Under nutrient or growth factor depletion, or alternatively following treatment with mTOR inhibitors, 4E-BP1 becomes dephosphorylated, binds to eIF4E and thus eIF4F complex activity is repressed (Supplementary Figure 14B). This leads to the inhibition of translation, most acutely for mRNAs with complex 5' untranslated regions (UTRs) that include proliferation, survival and tumor promoting

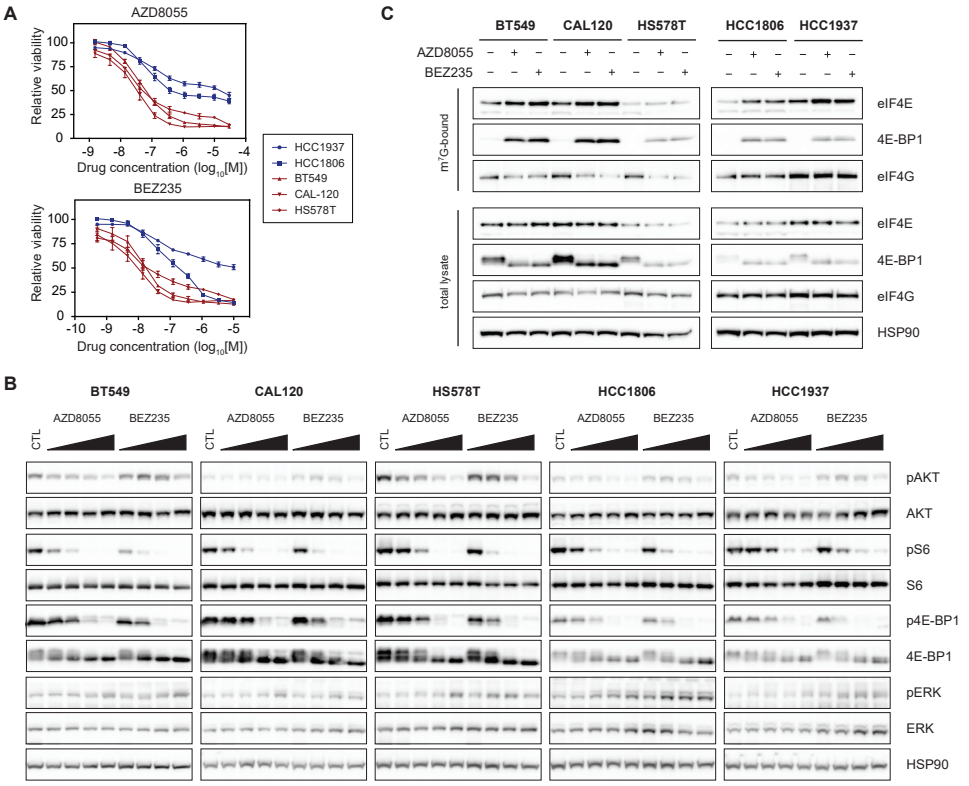


Figure 4.7: **Cell lines overexpressing the 4E-BP1 protein are more sensitive to mTOR inhibitors despite similar extent of target inhibition.** (A) Dose response assays (72 hr) to AZD8055 (mTOR inhibitor) and BEZ235 (dual mTOR/PI3K inhibitor) in cell lines which overexpress the 4E-BP1 protein – BT-549, CAL-120 and Hs 578T (red lines) - and those that do not - HCC1806 and HCC1937 (blue lines). Data represents three independent replicates +/- SEM. (B) Western blotting results of a number of PI3K/MAPK pathway components examined across the subpanel of high- and low-4E-BP1 expressing cell lines. Twenty four hours after seeding, cells were treated with increasing doses of AZD8055 (12 nM, 37 nM, 111 nM, 333 nM) and BEZ235 (4 nM, 12 nM, 37 nM, 111 nM) for a further 24 h, while untreated samples served as controls (CTL). Representative of three independent experiments is shown. (C) Cap pull down assays to assess the effect of mTOR inhibitor treatment on the formation of the translation initiation complex, eIF4F. Twenty four hours after seeding cells treated for a further 24 hr with 111 nM AZD8055, 37 nM BEZ235, or left untreated as control. Following lysis, cells were analyzed by m7G-cap pull down assays to determine effects on eIF4F complex formation. Protein expression of the components in the total lysates were also determined. Representative of three independent experiments is shown.

genes.

We investigated the dynamics of these interactions in the three mTOR inhibitor sensitive and two insensitive cell lines using the m7G-cap pull down assay, which allows the visualization of the changes in eIF4G or 4E-BP1 binding to eIF4E following treatment, as compared to their expression in the total protein lysate (Figure 4.7C). Our results show that mTOR inhibitor treatment leads to an increase in the binding of 4E-BP1 to eIF4E in all five cell lines, irrespective of their mTOR inhibitor response profile. In the sensitive cell lines, this increase in 4E-BP1 binding was sufficient to decrease eIF4G binding to eIF4E, as expected. In the insensitive cell lines though, the binding of 4E-BP1 to eIF4E was unable to displace eIF4G from eIF4E. This suggests that 4E-BP1 protein expression in the insensitive cell lines is below a critical threshold needed to effectively inhibit eIF4F complex formation following mTOR inhibitor treatment (see Supplementary Figure 14C), and likely explains the difference in mTOR sensitivity between these two sets of cell lines.

To further investigate whether an increase in 4E-BP1 protein expression is sufficient to increase mTOR inhibitor response, we used a lentiviral vector to overexpress 4E-BP1 in the two insensitive cell lines, HCC1806 and HCC1937 (Figure 4.8A). The 4E-BP1 protein, as well as a GFP control, was effectively overexpressed in both lines and the former was detectably phosphorylated. The expression of either protein did not affect the proliferation of these cell lines, suggesting that the activity of overexpressed 4E-BP1 was efficiently inhibited by mTORC1-mediated phosphorylation (Figure 4.8B). We next tested the impact of increased 4E-BP1 protein expression on the sensitivity of the cell lines to the mTOR inhibitor AZD8055, the dual mTOR/PI3K inhibitor BEZ235, as well as the PI3K inhibitor GDC0941. As shown in Figure 4.8C, in short-term 72 h drug treatment assays, the 4E-BP1-overexpressing cell lines were markedly more sensitive to mTOR inhibitors, responding at lower drug concentrations and with a decreased overall survival at higher drug concentrations. In contrast to the mTOR inhibitors, sensitivity of the 4E-BP1-overexpressing cell lines to the PI3K inhibitor GDC0941 was not increased, implying that it is specifically the inhibition of mTOR activity that is beneficial in improving response of highly expressing 4E-BP1 cell lines. We were also able to validate that 4E-BP1 overexpression increased sensitivity to mTOR inhibitors over a longer treatment period, namely 10 days, as shown in Figure 4.8D.

Together, the above results show that the level of 4E-BP1 protein expression is a determinant of sensitivity to mTOR inhibitors in breast cancer cell lines, and illustrates the utility of our computational model in identifying novel determinants of drug response in cell lines.

4.5. DISCUSSION

Cell line panels have the potential to provide us with a better understanding of the variability in drug response between patients. Previous efforts of linking molecular characteristics to drug sensitivity in cell line panels have identified several known and novel associations [10, 11, 13–16]. Here we showed that by combining extensive measurements with mathematical modeling, a more detailed understanding of variability in drug sensitivity can be achieved. The use of Bayesian statistics allowed for the simultaneous integration of diverse data types with prior knowledge from the literature.

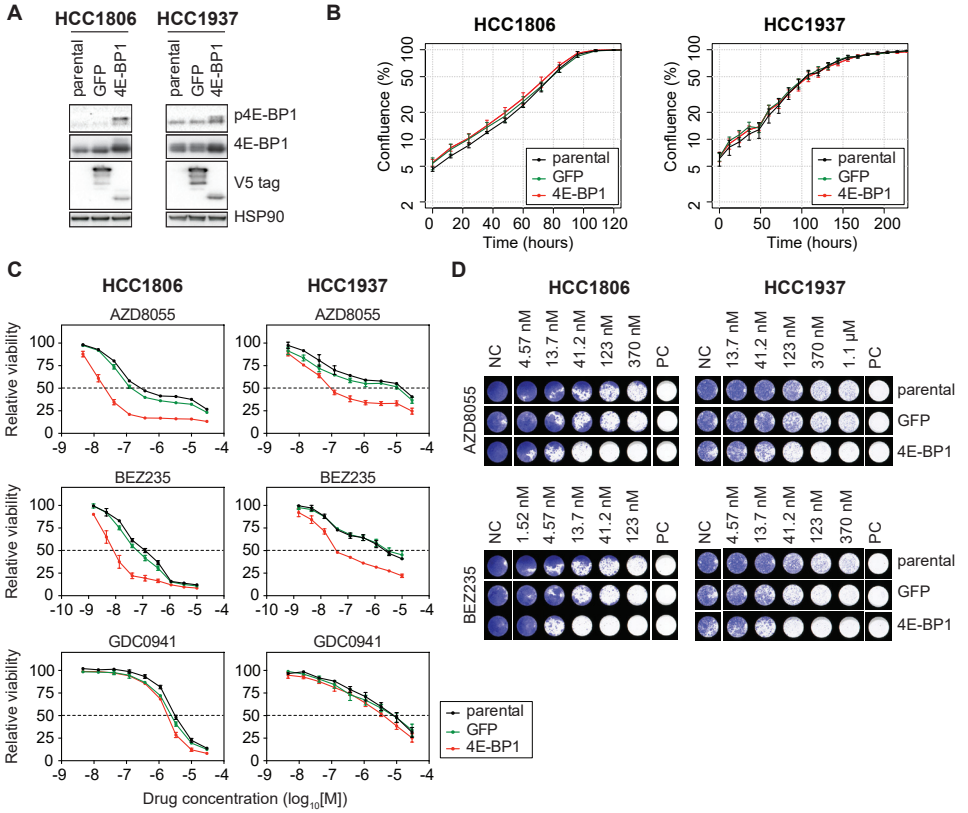


Figure 4.8: **Overexpression of 4E-BP1 in HCC1806 and HCC1937 cell lines is sufficient to increase their sensitivity to mTOR inhibitors.** (A) Western blotting of lysates from HCC1806 and HCC1937 cell lines stably overexpressing a 4E-BP1 construct from the CCSB-Broad Lentiviral Expression Collection as compared to the parental cell lines and a GFP-overexpressing controls. (B) Proliferation assay of the 4E-BP1 overexpressing lines as compared to the parental and GFP-expressing controls. (C) Dose response assays (72 hr) to AZD8055 (mTOR inhibitor), BEZ235 (dual mTOR/PI3K inhibitor) and GDC0941 (PI3K inhibitor) in the parental, GFP-expressing and 4E-BP1-expressing HCC1806 and HCC1937 cell lines. Data represents three independent replicates +/- SEM. (D) Long term (10 day) dose response assay to AZD8055 and BEZ235 in the parental, GFP-expressing and 4E-BP1-expressing HCC1806 and HCC1937 cell lines. Representative of three independent experiments is shown.

Models are by definition a simplified representation of the system. Two major simplifications that were used here are the assumption of quasi-steady state and the absence of feedback signaling. The benefit of using these simplifications is that significantly more components of cellular signaling can be included in the model. It is reassuring that a model with these simplifications can describe a large part of the variability in short-term drug response. Studying longer-term drug response and, for example, adaptive resistance will likely require the incorporation of dynamics and feedback mechanisms. Using dynamical models or the inclusion of feedback mechanisms does not pose any theoretical problems for the modelling approach we used. However, the computational cost is significantly higher, and additional intervention or time-course data would be needed to constrain the parameters.

Recently, Fey et al. described a computational model of JNK signaling, containing 5 signaling proteins that provided prognostic information in neuroblastoma patients [32], showing that such computational models may be useful also in a clinical setting. Their model was informative in a specific subset of patients, namely those with MYCN-amplified tumors, whereas in the general population individual biomarkers were still more informative. This indicates that it is necessary for models to incorporate various different oncogenic drivers, residing in multiple signaling pathways, in order to capture the variability across a wide range of patients.

A method that integrates multiple data types to obtain pathway activation status is PARADIGM [33]. Heiser et al. have used a modified version of this method, SuperPathway analysis, to link pathway activation status to drug response in a large breast cancer cell line panel [12]. They found, among others, that upregulation of DNA damage response pathways was associated with sensitivity to cisplatin, although this was not further tested experimentally. Indeed, PARADIGM and SuperPathway analyses do not shed light on how this association might work mechanistically, and the involvement of individual components of the DNA damage response pathway, such as TP53, ATM and BRCA1/2, is not investigated. In contrast, the approach presented here provides detailed signaling flows and estimates the relative contributions from all drivers and sensitivity mechanisms included in the model, making the in silico findings amenable to experimental validation.

In the course of refining our model, we found that elevated 4E-BP1 protein expression played an important role in the response of breast cancer cell lines to AZD8055 and BEZ235, two inhibitors targeting the mTOR kinase. We further validated this observation in 4E-BP1 overexpression studies, showing that this provides a pool of an endogenous translation inhibitor available for activation, and thus inhibition of cap-dependent translation, via mTOR inhibitor treatment.

On a mechanistic level, these findings are in line with prior work using transformed mouse embryonic fibroblasts, which showed that the ratio of eIF4E/4E-BP1 expression can predict response to mTOR-directed therapy [34]. That is, a higher ratio of eIF4E/4E-BP1 predicted poorer response to mTOR inhibitors. Consistent with this, EIF4E amplification was reported as a mechanism of AZD8055 resistance in a SW620 colorectal cell line model [35]. Conversely, a lack of 4E-BP1 expression in lymphoma cells, thus a reduced ability to restrain eIF4E activity, has been shown to lead to resistance to mTOR inhibition, an effect reversed by exogenous expression of the 4E-BP1 protein [36], similar

to our findings in 4E-BP1 overexpressing HCC1806 and HCC1937 cell lines. An exception to these findings is the report of elevated 4E-BP1 protein levels in mTOR inhibitor treatment-resistant luminal subpopulation of prostate cancer cells [37], although it remains to be investigated whether this observation is restricted to prostate cancer or the luminal subtype. Most recently, a study by Wang et al [38] added further weight to the hypothesis that elevated 4E-BP1 expression can be a marker of mTOR inhibitor sensitivity. They show that the combination of an mTOR inhibitor and an HDAC inhibitor — the latter acting to de-repress Snail-mediated 4E-BP1 transcriptional inhibition — can synergize to inhibit tumor growth in mice.

Notably, in our cell line panel, the overexpression of 4E-BP1 resulted from a copy number gain in the genomic region encoding EIF4EBP. Focal amplification of the 8p11-12 region, the region containing EIF4EBP1, is a known event in breast cancer occurring in approximately 15% of cases, and patients who harbor this event in their primary tumor have a much higher likelihood of relapse [39]. Previous studies have identified various genes in this region as potential oncogenes, with most evidence so far supporting FGFR1 [40–42] and ZNF703 [43–45]. These two genes lie close to and on either side of EIF4EBP1. Our cell line experiments show that overexpression of 4E-BP1 alone does not affect viability in vitro. While we cannot exclude that amplification of 4E-BP1 can contribute to a transformed phenotype in tumors (such as affecting invasion, migration or cell viability in vivo), various studies indicate that the cellular function of 4E-BP1 is consistent with a role as a tumor suppressor gene, rather than as an oncogene [34–36, 46–49]. It therefore seems plausible that the amplification of EIF4EBP1 is a passenger event. This raises the possibility that by selecting for amplification of nearby oncogenes, the tumors have also introduced a specific 'passenger vulnerability', that is, a passenger aberration that introduces a vulnerability to a particular drug. If the drug sensitivity association translates from cell lines to patients, testing for EIF4EBP1 amplifications could identify patients who may benefit most from treatment with mTOR inhibitors such as AZD8055 and BEZ235.

One difference between the cell line and patient data is that the subtypes in which the gain of EIF4EBP1 is present vary. While we identified this event predominantly in the triple negative and ER+ breast cancer cell lines, in patients, the 8p11-12 amplicon is present almost exclusively in ER+ tumors. Interestingly, phase III clinical trials with the allosteric mTOR inhibitor, everolimus, have been carried out in the ER+ setting showing a significant improvement in progression free survival (PFS) in patients who received everolimus versus placebo in addition to the aromatase inhibitor, exemestane [50]. A subsequent analysis of this data, exploring associations of PFS with common genetic aberrations have found no improvement in PFS in patients with FGFR1 gene amplification (generally co-amplifying the EIF4EBP1 gene) [51]. While this may suggest that there is no increased clinical benefit for mTOR inhibitors in patients with 8p11-12 amplified tumors, the inhibitory activity of everolimus, a rapalog, versus active-site inhibitors, such as AZD8055 and BEZ235, differs substantially. While everolimus is an allosteric inhibitor of predominantly mTOR complex 1 (mTORC1), active-site inhibitors are able to target both mTORC1 and 2 [30]. Most importantly, the active-site inhibitors have been shown to result in a much more potent inhibition of downstream mTOR signaling, specifically with respect to the inhibition of 4E-BP1 phosphorylation, whereas in

the case of rapalogs this phosphorylation is restored within hours of treatment [52, 53]. As such, it is likely that the absence of an association between amplification in the 8p11-12 region containing the FGRF1 and EIF4EBP1 genes and PFS with everolimus treatment observed by Hortobagyi et al. [51] stems from limited inhibition of 4E-BP1 phosphorylation. Together, these studies suggest that the benefit of rapamycin treatment in ER+ tumors results from a mechanism distinct to that of 4E-BP1 inhibition, but also emphasize the need for further study in order to determine the efficacy of active-site mTOR inhibitors in patients with 8p11-12 amplifications.

We conclude that the combination of mathematical modelling of signaling in response to drug treatment with a large panel of molecularly characterized cell lines demonstrates the usefulness of combining large data sets with prior knowledge to uncover key determinants of drug sensitivity. While the work presented here applied the ISA methodology to explain drug response in a panel of cell lines, our ultimate goal is to develop this approach into a tool for predicting response in patients. Further validation on clinical samples is of course required, but a clear benefit is that the molecular characterization, needed as input for a model, can be performed directly on biopsy material. This circumvents a need for ex vivo culture or lengthy response profiling, making the approach amenable to clinical application. We believe that such systematic, quantitative approaches to the understanding of drug responses hold promise as tools for achieving the goal of precision medicine in cancer.

ACKNOWLEDGMENTS

We are grateful to Thomas Kuilman for performing the CopywriteR analysis as well as the NKI Genomics Facility for performing high throughput sequencing experiments and subsequent data processing. We are grateful to Jordi Vidal Rodriguez for help with the drug response assays and maintenance of the cell line panel. We thank the Beijersbergen, Wessels and Bernards groups for helpful discussions.

REFERENCES

[1] R. Siegel, J. Ma, Z. Zou, and A. Jemal, *Cancer statistics, 2014*, CA Cancer J Clin **64**, 9 (2014).

[2] The Cancer Genome Atlas Network, *Comprehensive molecular portraits of human breast tumours*. Nature **490**, 61 (2012).

[3] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, *et al.*, *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups*. Nature **486**, 346 (2012).

[4] E. H. Romond, E. A. Perez, J. Bryant, V. J. Suman, C. E. Geyer, *et al.*, *Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer.* The New England journal of medicine **353**, 1673 (2005).

[5] C. E. Geyer, J. Forster, D. Lindquist, S. Chan, C. G. Romieu, *et al.*, *Lapatinib plus capecitabine for HER2-positive advanced breast cancer.* The New England journal of medicine **355**, 2733 (2006).

[6] F. H. Groenendijk and R. Bernards, *Drug resistance to targeted therapies: Deja vu all over again,* Molecular Oncology **8**, 1067 (2014).

[7] K. Berns, H. M. Horlings, B. T. Hennessy, M. Madiredjo, E. M. Hijmans, *et al.*, *A Functional*

Genetic Approach Identifies the PI3K Pathway as a Major Determinant of Trastuzumab Resistance in Breast Cancer, Cancer Cell **12**, 395 (2007).

[8] Y. Nagata, K. H. Lan, X. Zhou, M. Tan, F. J. Esteva, *et al., PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients,* Cancer Cell **6**, 117 (2004).

[9] T. R. Wilson, J. Fridlyand, Y. Yan, E. Penuel, L. Burton, *et al., Widespread potential for growth-factor-driven resistance to anticancer kinase inhibitors.* Nature **487**, 505 (2012).

[10] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, *et al., Systematic identification of genomic markers of drug sensitivity in cancer cells.* Nature **483**, 570 (2012).

[11] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. a. Margolin, *et al., The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature **483**, 603 (2012).

[12] L. M. Heiser, A. Sadanandam, W.-L. Kuo, S. C. Benz, T. C. Goldstein, *et al., Subtype and pathway specific responses to anticancer compounds in breast cancer.* Proceedings of the National Academy of Sciences of the United States of America **109**, 2724 (2012).

[13] A. Daemen, O. L. Griffith, L. M. Heiser, N. J. Wang, O. M. Enache, *et al., Modeling precision treatment of breast cancer,* Genome Biology **14** (2013).

[14] B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, *et al., Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset,* Cancer Discovery (2015).

[15] D. C. Kirouac, J. Y. Du, J. Lahdenranta, R. Overland, D. Yarar, *et al., Computational modeling of ERBB2-amplified breast cancer identifies combined ErbB2/3 blockade as superior to the combination of MEK and AKT inhibitors.* Science signaling **6**, ra68 (2013).

[16] B. Klinger, A. Sieber, R. Fritsche-Guenther, F. Witzel, L. Berry, *et al., Network quantification of EGFR signaling unveils potential for targeted combination therapy,* Molecular Systems Biology **9** (2013).

[17] F. Iorio, T. A. Knijnenburg, D. J. Vis, J. Saez-Rodriguez, U. Mcdermott, *et al., A Landscape of Pharmacogenomic Interactions in Cancer,* Cell **166**, 740 (2016).

[18] C. Sun, L. Wang, S. Huang, G. J. Heynen, A. Prahallad, *et al., Reversible and adaptive resistance to BRAF(V600E) inhibition in melanoma,* Nature **508**, 118 (2014).

[19] B. Thijssen, T. M. H. Dijkstra, T. Heskes, and L. F. A. Wessels, *BCM: toolkit for Bayesian analysis of Computational Models using samplers,* BMC Systems Biology **10**, 100 (2016).

[20] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, *et al., Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.* Nucleic acids research **41**, D955 (2013).

[21] C. L. Arteaga and J. A. Engelman, *ERBB receptors: From oncogene discovery to basic science to mechanism-based cancer therapeutics,* Cancer Cell **25**, 282 (2014).

[22] D. W. Rusnak, K. Lackey, K. Affleck, E. R. Wood, K. J. Alligood, *et al., The effects of the novel, reversible epidermal growth factor receptor/ErbB-2 tyrosine kinase inhibitor, GW2016, on the growth of human normal and tumor-derived cell lines in vitro and in vivo.* Molecular cancer therapeutics **1**, 85 (2001).

[23] V. Serra, B. Markman, M. Scaltriti, P. J. a. Eichhorn, V. Valero, *et al., NVP-BEZ235, a dual PI3K/mTOR inhibitor, prevents PI3K signaling and inhibits the growth of cancer cells with ac-*

tivating PI3K mutations. Cancer research **68**, 8022 (2008).

[24] N. Ilic, T. Utermark, H. R. Widlund, and T. M. Roberts, *PI3K-targeted therapy can be evaded by gene amplification along the MYC-eukaryotic translation initiation factor 4E (eIF4E) axis,* Proceedings of the National Academy of Sciences **108**, E699 (2011).

[25] Y. Kataoka, T. Mukohara, H. Tomioka, Y. Funakoshi, N. Kiyota, *et al.*, *Foretinib (GSK1363089), a multi-kinase inhibitor of MET and VEGFRs, inhibits growth of gastric cancer cell lines by blocking inter-receptor tyrosine kinase networks,* Investigational New Drugs **30**, 1352 (2012).

[26] J. Das and H. Yu, *HINT: High-quality protein interactomes and their applications in understanding human disease,* BMC Systems Biology **6**, 1 (2012).

[27] R. E. Kass and A. E. Raftery, *Bayes factors,* Journal of the American Statistical Association **90**, 773 (1995).

[28] J. E. Korkola, E. A. Collisson, L. Heiser, C. Oates, N. Bayani, *et al.*, *Decoupling of the PI3K Pathway via Mutation Necessitates Combinatorial Treatment in HER2+ Breast Cancer,* Plos One **10**, e0133219 (2015).

[29] B. Thijssen, K. Jastrzebski, R. L. Beijersbergen, and L. F. A. Wessels, *Delineating feedback activity in the MAPK and AKT pathways using feedback-enabled Inference of Signaling Activity,* bioRxiv (2018).

[30] M. Bhat, N. Robichaud, L. Hulea, N. Sonenberg, J. Pelletier, and I. Topisirovic, *Targeting the translation machinery in cancer.* Nature reviews. Drug discovery **14**, 261 (2015).

[31] J. Pelletier, J. Graff, D. Ruggero, and N. Sonenberg, *Targeting the eIF4F translation initiation complex: A critical nexus for cancer development,* Cancer Research **75**, 250 (2015).

[32] D. Fey, M. Halasz, D. Dreidax, S. P. Kennedy, J. F. Hastings, *et al.*, *Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients,* Science Signaling **8**, RA130 (2015).

[33] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart, *Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM,* Bioinformatics **26**, 237 (2010).

[34] T. Alain, M. Morita, B. D. Fonseca, A. Yanagiya, N. Siddiqui, *et al.*, *eIF4E/4E-BP ratio predicts the efficacy of mTOR targeted therapies,* Cancer Research **72**, 6468 (2012).

[35] C. L. Cope, R. Gilley, K. Balmanno, M. J. Sale, K. D. Howarth, *et al.*, *Adaptation to mTOR kinase inhibitors by amplification of eIF4E to maintain cap-dependent translation,* Journal of Cell Science **127**, 788 (2014).

[36] S. Mallya, B. A. Fitch, J. S. Lee, L. So, M. R. Janes, and D. A. Fruman, *Resistance to mTOR kinase inhibitors in lymphoma cells lacking 4EBP1,* PLoS ONE **9**, 1 (2014).

[37] A. C. Hsieh, H. G. Nguyen, L. Wen, M. P. Edlind, P. R. Carroll, W. Kim, and D. Ruggero, *Cell type–specific abundance of 4EBP1 primes prostate cancer sensitivity or resistance to PI3K pathway inhibitors,* Science Signaling **8**, ra116 (2015).

[38] J. Wang, Q. Ye, Y. Cao, Y. Guo, X. Huang, *et al.*, *Snail determines the therapeutic response to mTOR kinase inhibitors by transcriptional repression of 4E-BP1,* Nature Communications **8** (2017).

[39] V. Gelsi-Boyer, B. Orsetti, N. Cervera, P. Finetti, F. Sircoulomb, *et al.*, *Comprehensive profiling of 8p11-12 amplification in breast cancer.* Molecular Cancer Research **3**, 655 (2005).

[40] C. Theillet, J. Adelaide, G. Louason, F. Bonnet-Dorion, J. Jacquemier, *et al.*, *FGFR1 and PLAT genes and DNA amplification at 8p12 in breast and ovarian cancers,* Genes, chromosomes & cancer **7**, 219 (1993).

[41] F. Ugolini, J. Adelaide, E. Charafe-Jauffret, C. Nguyen, J. Jacquemier, B. Jordan, D. Birnbaum, and M. J. Pebusque, *Differential expression assay of chromosome arm 8p genes identifies Frizzled-related (FRP1/FRZB) and Fibroblast Growth Factor Receptor 1 (FGFR1) as candidate breast cancer genes,* Oncogene **18**, 1903 (1999).

[42] N. Turner, A. Pearson, R. Sharpe, M. Lambros, F. Geyer, *et al.*, *FGFR1 amplification drives endocrine therapy resistance and is a therapeutic target in breast cancer,* Cancer Research **70**, 2085 (2010).

[43] M. J. Garcia, J. C. M. Pole, S.-F. Chin, A. Teschendorff, A. Naderi, *et al.*, *A 1 Mb minimal amplicon at 8p11-12 in breast cancer identifies new candidate oncogenes.* Oncogene **24**, 5235 (2005).

[44] D. G. Holland, A. Burleigh, A. Git, M. a. Goldgraben, P. a. Perez-Mancera, *et al.*, *ZNF703 is a common Luminal B breast cancer oncogene that differentially regulates luminal and basal progenitors in human mammary epithelium.* EMBO molecular medicine **3**, 167 (2011).

[45] E. M. Slorach, J. Chou, and Z. Werb, *Zeppo1 is a novel metastasis promoter that represses E-cadherin expression and regulates p120-catenin isoform expression and localization,* Genes and Development **25**, 471 (2011).

[46] G. S. Ducker, C. E. Atreya, J. P. Simko, Y. K. Hom, M. R. Matli, *et al.*, *Incomplete inhibition of phosphorylation of 4E-BP1 as a mechanism of primary resistance to ATP-competitive mTOR inhibitors,* Oncogene **33**, 1590 (2014).

[47] Y. Martineau, R. Azar, D. Müller, C. Lasfargues, S. El Khawand, *et al.*, *Pancreatic tumours escape from translational control through 4E-BP1 loss,* Oncogene **33**, 1367 (2014).

[48] W. Cai, Q. Ye, and Q.-B. She, *Loss of 4E-BP1 function induces EMT and promotes cancer cell migration and invasion via cap-dependent translational activation of snail.* Oncotarget **5**, 6015 (2014).

[49] M. E. Martín, M. I. Pérez, C. Redondo, M. I. Alvarez, M. Salinas, and J. L. Fando, *4E binding protein 1 expression is inversely correlated to the progression of gastrointestinal cancers.* The international journal of biochemistry & cell biology **32**, 633 (2000).

[50] D. A. Yardley, S. Noguchi, K. I. Pritchard, H. A. Burris, J. Baselga, *et al.*, *Everolimus plus exemestane in postmenopausal patients with HR+ breast cancer: BOLERO-2 final progression-free survival analysis,* Advances in Therapy **30**, 870 (2013).

[51] G. N. Hortobagyi, D. Chen, M. Piccart, H. S. Rugo, H. A. Burris, *et al.*, *Correlative Analysis of Genetic Alterations and Everolimus Benefit in Hormone Receptor-Positive, Human Epidermal Growth Factor Receptor 2-Negative Advanced Breast Cancer: Results From BOLERO-2.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology **34** (2015).

[52] A. Y. Choo, S.-O. Yoon, S. G. Kim, P. P. Roux, and J. Blenis, *Rapamycin differentially inhibits S6Ks and 4E-BP1 to mediate cell-type-specific repression of mRNA translation,* Proceedings of the National Academy of Sciences **105**, 17414 (2008).

[53] A. Y. Choo and J. Blenis, *Not all substrates are treated equally: Implications for mTOR, rapamycin-resistance and cancer therapy,* Cell Cycle **8**, 567 (2009).

[54] H. G. Katzgraber, S. Trebst, D. A. Huse, and M. Troyer, *Feedback-optimized parallel tempering Monte Carlo,* Journal of Statistical Mechanics: Theory and Experiment **2006**, P03018 (2006),

0602085v3 .

[55] P. Del Moral, A. Doucet, and A. Jasra, *Sequential Monte Carlo samplers,* Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**, 411 (2006).

[56] D. Turek, P. de Valpine, C. J. Paciorek, and C. Anderson-Bergman, *Automated Parameter Blocking for Efficient Markov-Chain Monte Carlo Sampling,* Bayesian Analysis **12**, 465 (2017), 1503.05621 .

[57] A. Gelman and X.-L. Meng, *Simulating normalizing constants: from importance sampling to bridge sampling to path sampling,* Statistical Science **13**, 163 (1998).

[58] P. Del Moral, A. Doucet, and A. Jasra, *An adaptive sequential Monte Carlo method for approximate Bayesian computation,* Statistics and Computing **22**, 1009 (2011).

[59] I. Rodríguez-Escudero, M. D. Oliver, A. Andrés-Pons, M. Molina, V. J. Cid, and R. Pulido, *A comprehensive functional analysis of PTEN mutations: implications in tumor- and autism-related syndromes.* Human molecular genetics **20**, 4132 (2011).

[60] G. Singh, L. Odriozola, H. Guan, C. R. Kennedy, and A. M. Chan, *Characterization of a novel PTEN mutation in MDA-MB-453 breast carcinoma cell line.* BMC cancer **11**, 490 (2011).

[61] X. Wan, B. Harkavy, N. Shen, P. Grohar, and L. J. Helman, *Rapamycin induces feedback activation of Akt signaling through an IGF-1R-dependent mechanism,* Oncogene **26**, 1932 (2007).

[62] O. J. Shah, Z. Wang, and T. Hunter, *Inappropriate activation of the TSC/Rheb/mTOR/S6K cassette induces IRS1/2 depletion, insulin resistance, and cell survival deficiencies.* Current biology **14**, 1650 (2004).

[63] C. M. Chresta, B. R. Davies, I. Hickson, T. Harding, S. Cosulich, *et al.*, *AZD8055 is a potent, selective, and orally bioavailable ATP-competitive mammalian target of rapamycin kinase inhibitor with in vitro and in vivo antitumor activity,* Cancer Research **70**, 288 (2010).

[64] S.-M. Maira, F. Stauffer, J. Brueggen, P. Furet, C. Schnell, *et al.*, *Identification and characterization of NVP-BEZ235, a new orally available dual phosphatidylinositol 3-kinase/mammalian target of rapamycin inhibitor with potent in vivo antitumor activity.* Molecular cancer therapeutics **7**, 1851 (2008).

[65] A. J. Folkes, K. Ahmadi, W. K. Alderton, S. Alix, S. J. Baker, *et al.*, *The identification of 2-(1H-indazol-4-yl)-6-(4-methanesulfonyl-piperazin-1-ylmethyl)-4-morpholin-4-yl-thieno[3,2-d]pyrimidine (GDC-0941) as a potent, selective, orally bioavailable inhibitor of class I PI3 kinase for the treatment of cancer .* Journal of medicinal chemistry **51**, 5522 (2008).

[66] F. Qian, S. Engst, K. Yamaguchi, P. Yu, K.-A. Won, *et al.*, *Inhibition of tumor cell growth, invasion, and metastasis by EXEL-2880 (XL880, GSK1363089), a novel inhibitor of HGF and VEGF receptor tyrosine kinases.* Cancer research **69**, 8009 (2009).

[67] L. Yan, *Abstract #DDT01-1: MK-2206: A potent oral allosteric AKT inhibitor,* in *AACR Annual Meeting* (2009).

[68] S. D. Barrett, A. J. Bridges, D. T. Dudley, A. R. Saltiel, J. H. Fergus, *et al.*, *The discovery of the benzhydroxamate MEK inhibitors CI-1040 and PD 0325901,* Bioorganic and Medicinal Chemistry Letters **18**, 6501 (2008).

SUPPLEMENTARY MATERIAL

4.5.1. CELL LINE PANEL

A 30 breast cancer cell line panel was assembled from various sources, the details and growth conditions of which are listed in Supplementary Table 1.

4.5.2. CELL LINE DOUBLING TIMES

Doubling times were determined using the IncuCyte FLR/ZOOM instrument (Essen Bioscience). Each cell line was seeded over a range of densities from 10,000 to 313 cells per 384-well plate well, prepared as a 1:2 dilution series, in technical quadruplicate. Percentage confluence was quantified every 4 h, over a period of 96 h. Exponential growth curves were fitted to these data to derive the doubling time of each cell using GraphPad Prism. The doubling times obtained from at least two seeding densities, representing the most complete proliferation curves, were averaged to obtain the final estimate. The resulting doubling times are provided in Supplementary Table 12.

4.5.3. DRUG RESPONSE ASSAYS

Prior to carrying out drug response assays, cell line seeding densities were optimized. Cells seeded as for the cell doubling time experiments were assessed at the 96 h endpoint for percentage confluence, and incubated with CellTiter-Blue (CTB; Promega) for a measure of metabolic activity. This was to ensure that cell lines did not exceed 90% confluence at assay endpoint and that the CTB signal at this density was not saturated. Seeding densities used for each cell line are listed in Supplementary Table 1.

For drug response assays, cells were seeded at the optimized density and 24 h later treated with a 10-point 1:3 dilution series of a number of inhibitors using a Microlab STAR workstation fitted with 8 x 1000µl channels and 96-probe head (Hamilton): AZD8055, top dose 3 x 10-5 M; BEZ235, 1 x 10-5 M; GDC0941, 3 x 10-5 M; MK2206, 3 x 10-5 M; PD0325901, 3 x 10-5 M; Lapatinib, 3 x 10-5 M; Foretinib, 3 x 10-5 M (all from Selleckchem). Each condition, including an untreated negative control and a phenyl arsine oxide (1 x 10-6 M) treated positive control, were set up in technical quadruplicate. Following a 72 h incubation, cells were stained with CTB (1:30 dilution) for 4 h and the signal measured using an Envision spectrophotometer (Perkin Elmer). In the case of the validation experiments with HCC1806 and HCC1937 cell lines expressing 4E-BP1 or GFP constructs, cells were treated in a 9-point 1:3 dilution series of AZD8055, BEZ235 or GDC0941 using a HP D300 Digital Dispenser (Hewlett-Packard), while all other experimental conditions remained the same. Each assay was carried out in biological triplicate. Each replicate of a dose response experiment was further analyzed by normalization to the negative and positive control (the normalized data are provided in Supplementary Table 6. For calculating the 50% inhibitory concentration (IC50, dose at which viability is 50% compared to the untreated control), the normalized data was fitted by a four-parameter sigmoid function. The IC50 estimates are provided in Supplementary Table 7. For the model inference, the full dose response curve data were used.

4.5.4. LONG-TERM DRUG RESPONSE ASSAYS

HCC1806 parental, GFP- and 4E-BP1-expressing cells were seeded at 600 cells/well, while the HCC1937 panel was seeded at 1200 cells/well, in 96 well plates. Cells were treated, 24 h after seeding, with a 9-point 1:3 dilution series of AZD8055 (top dose 3.3 x 10-6 M) or BEZ235 (1.1 x 10-6 M) using a HP D300 Digital Dispenser (Hewlett-Packard). Each condition, including an untreated negative control and a phenyl arsine oxide (1 x 10-6 M) treated positive control, were set up in technical duplicate. Media and drugs were changed every 3-4 days over a period of 10-11 days of treatment. Cells were then washed with PBS, fixed with 3.7% formaldehyde/PBS and stained in 0.1% crystal violet solution. Images of dried, stained cells were digitized on a Perfection V750 PRO scanner (Epson).

4.5.5. RNA EXPRESSION

Steady state RNA expression was determined from cells seeded in 60 mm dishes (densities in Supplementary Table 1) and grown for 48 h. To harvest, cells were washed once with ice cold PBS, and lysed in 2 ml Trizol by scraping. Lysate was collected and vortexed to ensure complete solubilization. RNA was purified by a standard Trizol extraction protocol. RNA TrueSeq libraries, using Illumina indexing, were prepared by the NKI Genomics Facility using standard protocols. The resulting count data were normalized for sequencing depth and log-transformed. The RNA sequencing data is available at ArrayExpress, reference E-MTAB-4801 and the normalized read counts are provided in Supplementary Table 8.

4.5.6. RPPA MEASUREMENTS

Steady state protein expression samples were prepared in biological triplicate from cells seeded in 60 mm dishes (densities in Supplementary Table 1) and grown for 48 h. Cells were lysed in RIPA buffer (20 mM Tris-HCl, pH 8, 150 mM NaCl, 1% NP40, 0.5% sodium deoxycholate, 0.1% SDS) supplemented with cOmplete protease and phosSTOP phosphatase inhibitor cocktails (Roche). Lysates were cleared by centrifugation at 4°C and 20,800 x g, and protein concentration determined using the Pierce BCA protein assay (Thermo Fisher Scientific). The supernatant was normalized to 1 µg/µl with RIPA buffer and supplemented with SDS sample buffer to a final concentration of 62.5 mM Tris-HCl pH 6.8, 10% glycerol, 2% SDS, 2.5% (v/v) 2-mercaptoethanol. Samples were further assayed at the MD Anderson Cancer Center RPPA Core Facility. Cell lysates were five times 2-fold serially diluted in dilution buffer (lysis buffer containing 1% SDS). Serially diluted lysates were arrayed on nitrocellulose-coated slides (Grace Biolab) in an 11 x 11 format by Aushon 2470 Arrayer (Aushon BioSystems) alongside positive and negative controls composed of mixed cell lysates or dilution buffer, respectively. Each slide was probed with a primary antibody, followed by a biotin-conjugated secondary antibody. The signal obtained was amplified using a catalyzed signal amplification system (Dako) and visualized by DAB colorimetric reaction. Slides were scanned to 16-bit tiff images on a flatbed scanner. Spots were identified and analyzed by ArrayPro. Each dilution curve was fitted with a logistic model ("Supercurve Fitting" developed by the Department of Bioinformatics and Computational Biology at the MD Anderson Cancer Center; http://bioinformatics.mdanderson.org/OOMPA), which fits a single curve using all the samples of a dilution series on a slide with the signal intensity as the response variable

and the dilution steps as an independent variable. The fitted curve is plotted with the signal intensities - both observed and fitted - on the y-axis for diagnostic purposes. The protein concentrations of each set of slides are then normalized by median polish, which is corrected across samples by the linear expression values using the median expression levels of all antibody experiments to calculate a loading correction factor for each sample. The normalized RPPA expression values are provided in Supplementary Table 9.

4.5.7. DNA CAPTURE, MUTATION SEQUENCING, COPY NUMBER ANALYSIS

Genomic DNA samples were obtained for each cell line of the panel from pellets of 0.5 x 106 cells, extracted using a DNeasy Blood and Tissue kit (Qiagen) by standard methods. DNA TruSeq libraries, using Illumina indexing, were prepared by the NKI Genomics Facility using standard protocols. Capture enrichment was performed using the human kinome DNA capture baits (Agilent Technologies). Five to six libraries were pooled for each capture reaction at 150 ng of each library. Custom blockers were added to prevent hybridization to adapter sequences - B1: 5'AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG/3'ddC; B2: 5'CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT/3'ddC. Captured libraries were sequenced on an Illumina HiSeq2000 platform with a paired end 51 base protocol. Samples were aligned against the human genome (build GRCh37.55) using BWA (version 0.5.10). Potential PCR duplicates were filtered using Picard MarkDuplicates (http://broadinstitute.github.io/picard). GATK (version 2.3-23) was used for local realignment in the capture target regions and for calling SNPs, and small indels were called with Pindel (version 0.5.7). Mismatches to the reference occurring in 1 read only were filtered out. Genome-wide copy number profiles were estimated from the off-target DNA sequencing reads with CopywriteR, using 20 kb bin sizes. The DNA sequencing data is available at the European Nucleotide Archive, reference PRJEB14120. The mutations obtained from the mutation calling are provided in Supplementary Table 10, and the copy number estimates are provided in Supplementary Table 11.

4.5.8. IMMUNOBLOTTING

Cells were treated as outlined in the figure legends and lysates prepared in RIPA buffer as outlined for RPPA measurements. Equal amounts of protein were supplemented with Novex® LDS Sample Buffer and Sample Reducing Agent, heated at 70°C for 10 min and separated on 4-12% gradient gels (Thermo Fisher Scientific). Separated proteins were transferred onto Immobilon-P PVDF membranes (Merck Millipore) using a Trans-Blot® system (Bio-Rad). Blocking was performed in TBS supplemented with 0.1% Tween and 5% BSA (TBS-TB) for 1 h at room temperature, followed by overnight immunoblotting at 4°C with the following primary antibodies: pAKT (S473) (Cell Signaling Technology #9271); AKT (Cell Signaling Technology #9272); pS6 (S235/236) (Cell Signaling Technology #2211); S6 (Cell Signaling Technology #2217); p4E-BP1 (S65) (Cell Signaling Technology #9451); 4E-BP1 (Cell Signaling Technology #9452); pERK1/2 (T202/Y204) (Cell Signaling Technology #9101); ERK (Santa Cruz Biotechnology sc-93 and sc-154); HSP90 (Santa Cruz Biotechnology sc-7947); eIF4E (Santa Cruz Biotechnology sc-271480); eIF4G (Cell Signaling Technology #2498); V5 epitope tag (Thermo Fisher Scientific R960-25). Membranes were then washed with TBS supplemented with 0.1% Tween (TBS-T) and

probed with secondary goat anti-mouse or anti-rabbit HRP-conjugated antibodies (Bio-Rad) diluted in TBS-TB for 2 h at room temperature. Finally, membranes were washed in TBS-T, an ECL reaction was carried out using the ClarityTM Western ECL Substrate (Bio-Rad) and the signal detected using a ChemiDoc Touch instrument (Bio-Rad).

4.5.9. CAP-BINDING PULL DOWN ASSAYS

Cells were seeded in 100 mm dishes (BT549 and CAL-120 at 2.5 x 105; Hs 578T at 3 x 105; HCC1806 at 4 x 105; HCC1937 at 6.25 x 105) and cultured for 48 h, then treated with AZD8055 (1.11 x 10-7 M), BEZ235 (3.7 x 10-8 M) or vehicle (DMSO) for a further 24 h. Cells were washed once with ice-cold PBS and lysed in lysis buffer (25 mM Tris-HCl, pH 7.6, 1% Triton X-100, 1 mM DTT) supplemented with cOmplete protease and phosSTOP phosphatase inhibitor cocktails (Roche). Lysates were cleared and assayed for protein concentration, then total protein samples were prepared using 20 µg of protein lysate as outlined above. Cap pull-down samples were prepared by combining 50 µg of total lysate with 20 µl pre-washed m7GTP-agarose (Jena Bioscience), made up to a total volume of 500 µl with lysis buffer and tumbled at 4°C overnight. The following day, cap pull-downs were washed 3 x in ice-cold lysis buffer, then heated at 70°C for 10 min in 20 µl 1x Novex® LDS Sample Buffer and Sample Reducing Agent. The eluate from the cap pull-downs as well as the total protein control samples were then immediately separated on Novex® 4-12% gradient gels and immunoblotted as outlined above using primary antibodies to 4E-BP1, eIF4G and HSP90 (for total lysates samples only), then reprobed to detect eIF4E protein.

4.5.10. GENERATION OF 4E-BP1 OVEREXPRESSING CELL LINES

pLX304-4E-BP1 was obtained from the CCSB-Broad Lentiviral Expression Collection, while the pLX304-GFP control construct was generated as outlined previously. To produce lentiviral particles, HEK293T cells were co-transfected with the pLX304-4E-BP1 or -GFP bearing construct and a lentiviral packaging mix (pRSV-Rev, pMDLg/pRRE, pCMV-VSV-G; Addgene) using Polyethylenimine (PEI, Linear MW 25,000; Polysciences Inc.). Media was changed 24 h after transfection. After a further 24 h, viral supernatant was collected and 0.45 µm-filtered. HCC1806 and HCC1937 cells were transduced in the presence of hexadimethrine bromide (Sigma-Aldrich) and following 48 h selected using blasticidin.

4.5.11. PROLIFERATION OF 4E-BP1 OVEREXPRESSING CELL LINES

HCC1806 parental, GFP- and 4E-BP1-expressing cells were seeded at 800 cells/well, while the HCC1937 panel was seeded at 1000 cells/well, in 384 well plates with 4-6 replicates per condition. Proliferation was monitored using the IncuCyte ZOOM instrument (Essen Biosciences).

4.5.12. MODEL DESCRIPTION

The final model we obtained is depicted in Figure 2B. This depiction consists of nodes and edges that connect the nodes. The nodes represent molecules (ligands, surface receptors, signaling molecules in the MAPK an PI3K pathways), mutations and copy number aberrations that have been shown to play a role in breast cancer, the six drugs we

employed in our experiments and proliferation as final output. The edges represent the effects the nodes have on each other. For example, the inhibitory effect of a drug on its target is represented by an inhibitory edge. The purpose of the modelling process was to integrate the different data types in such a way that the variability in the response of the 30 cell lines to the seven drugs can be explained, and to provide estimates of the latent variables, i.e. the signaling activities and the strengths of the edges. In the Model Construction section we have already provided some detail regarding the Bayesian modelling process. In this section we provide the exact mathematical equations that were employed in the modelling. It is important to note that while this type of model can in principle be viewed as a Bayesian network, the Bayesian statistics operate on the parameters and not on the edges between the nodes. The edges between the nodes, i.e. the functions for calculating the value of the nodes from their parent nodes, are deterministic. This approach is distinct from a different common use of Bayesian networks, where the links between nodes in the network are probabilistic. We also do not estimate the structure of the network, but rather estimate the parameters and the marginal likelihood of a given network structure.

The full models with all equations are included in Supplementary Data 1 as model description files which can be run directly in the BCM inference software (see Model Inference section). Below follows a description of how these models are constructed.

The regulatory signaling in the cell lines is modelled with continuous variables $A_{i,j}$, which describe the steady state activity of the ith signaling molecule in the jth cell line. These variables can assume values between 0 and 1, and are deterministic functions of the upstream signals, as well as of the total measured expression level of the signaling molecule. Specifically:

$$A^*_{i,j} = E_{i,j}(b_i + \sum_{k \in \text{parents}_A(i)} s_{k,i} A_{k,j} + \sum_{k \in \text{parents}_M(i)} s_{\text{mut},k,i} M_{k,j} \tag{4.5}$$

$$A_{i,j} = \max(\min(A^*_{i,j}, 1), 0) \tag{4.6}$$

The parameter b_i represents the basal activity of the ith signaling molecule in the absence of any upstream input, and can assume values between 0 and 1. The parameter $s_{k,i}$ represents the strength of the activation of signaling molecule i by signaling molecule k and is defined in more detail below. $M_{k,j}$ is a binary variable representing a point mutation or copy number aberration, which is set to 1 if the mutation or copy number aberration is present and 0 otherwise. The parameter $s_{\text{mut},k,i}$ represents the strength of the activation signal arising from such a mutation or copy number aberration in the kth molecule affecting signaling molecule i. Finally, the variable $E_{i,j}$ represents the expression level of the ith signaling molecule in the jth cell line and is also defined in more detail below. In Figure 4.3, for the example of S6K, the variable $A_{i,j}$ is represented by S6K signal, $s_{k,i}$ is represented by mTORC1->S6K strength for the link between mTORC1 and S6K and bi is represented by S6K base signal.

The parameters for these functions (b_i , $s_{k,i}$ and $s_{mut,k,i}$) are shared between all cell lines. Specifically, the parameters assume a single value which remains constant for all cell lines. For example, each cell line can have a different level of MEK activity, perhaps due to the presence or absence of a KRAS- or BRAF-mutation. As a result of the different

levels of MEK activity in each cell line, the downstream signaling molecule ERK can assume different activity levels in each cell line. However, a given level of MEK activity will always result in the same input signal to ERK, in each cell line.

The strength parameter $s_{k,i}$ can assume values between 0 and 5 for activating signals. As a result of Equation 4.5, a value between 0 and 1 leads to a diminished signal from the upstream to the downstream signaling molecule. This allows a signaling molecule with multiple inputs to receive contributions from each upstream molecule without resulting in an excessively large total input signal. A value between 1 and 5 results in an amplification of the signal, such that a small upstream signal leads to a large downstream signal. In an analogous fashion, the $s_{k,i}$ parameter for inhibitory signals can assume values between -5 and 0, such that these signals lead to an inhibition of the target signaling molecule. The strength parameter for mutations, $s_{\text{mut},k,i}$, can assume values between 0 and 1. As described in Equation 4.5, the summed upstream signal together with the base signal is multiplied by the expression level of the signaling molecule. The total signal is then clamped between 0 and 1.

In the initial model, the expression of each signaling protein i in cell line j, $E_{i,j}$, was based only on the binarized mRNA expression data. That is, $E_{i,j}$ was set to 1 if the RNA expression measurement for that gene was above a threshold, and 0 otherwise. After the model iteration where protein expression was included (see the section "Searching for additional explanatory factors of drug sensitivity reveals novel associations" in the main text), the expression $E_{i,j}$ is given by the equation:

$$E_{i,j} = \begin{cases} (p_i P_{i,j}) + (1 - p_i) & \text{when RPPA data is available} \\ 0 & \text{when no RPPA data is available and } R_{i,j} \leq T_{\text{RNAseq}} \\ 1 & \text{when no RPPA data is available and } R_{i,j} > T_{\text{RNAseq}} \end{cases} \quad (4.7)$$

Here $P_{i,j}$ is the normalized protein expression level of protein i in cell line j measured by RPPA, p_i is the expression coefficient which is defined in more detail below, $R_{i,j}$ is the normalized, log transformed mRNA expression level of the gene coding for protein i in cell line j measured by RNAseq and T_{RNAseq} is a threshold for expressed versus not-expressed genes selected based on Gamma mixture modeling (see Supplementary Figure 10).

For proteins where we do not have protein expression data available, the mRNA level is used as a proxy. The mRNA level is binarized and the expression variable set to 1 or 0 to signify whether the gene is expressed or not. For proteins where we do have protein expression data available, the expression is modelled as a coefficient times the normalized expression value. This expression coefficient, p_i, assumes values between 0 and 1, and it controls whether the amount of protein is a limiting factor for the signal transduction. Specifically, a small expression coefficient indicates that even small amounts of protein can fully transmit the signal, while a large coefficient indicates that having only small amounts of protein strongly limits the signal that can be transmitted. In Figure 4.3, for the example of S6K, since RPPA data is available, the top condition in Equation 4.7 applies, and p_i is represented by S6K expression coeff and $P_{i,j}$ is represented by S6K expression (RPPA).

The effect of the drugs on the cells is modelled in two parts: an on-target effect and an off-target effect. The on-target effect is a non-linear inhibition of the known target or targets of the mth drug:

$$A_{i,j,\text{inhibited}} = A_{i,j}(K_{\text{ontarget},m} + \frac{1 - K_{\text{ontarget},m}}{10^{h_{\text{ontarget},m}(c - c_{\text{IC50,ontarget},m})} + 1}), \tag{4.8}$$

where $K_{\text{ontarget},m}$ is the maximum inhibition, $h_{\text{ontarget},m}$ is the steepness, c the concentration of the drug in $\log_{10}$ scale and $c_{\text{IC50,ontarget},m}$ the half-maximal inhibition concentration also in $\log_{10}$ scale, all for the mth drug. When the drug is administered, and the ith signaling molecule is a target of the drug, then Equation 4.8 is applied, otherwise $A_{i,j}$ remains unaltered. In some cases, a drug has multiple targets; for example lapatinib targets both EGFR and ERBB2. In this case, each target has separate parameters, since the drug may have different affinities and effects for each target. Note that the $c_{\text{IC50,ontarget},m}$ used here is distinct from the IC50 values estimated for data exploration purposes earlier, and is estimated along with all other parameters during the model inference. The off-target effect follows later.

Proliferation is modeled as exponential growth, where the growth rate is a linear combination of the signaling molecules that signal directly to proliferation:

$$r_j = r_b + \sum_{k \in \text{parents(proliferation)}} k_k A_{k,j}, \tag{4.9}$$

where r_j is the exponential growth rate of cell line j, r_b is the base growth rate in the absence of any signal, and k_k is a parameter describing how strongly signaling molecule $A_{k,j}$ gives rise to a proliferation signal. The parents of the proliferation node are those signaling molecules which affect proliferation or survival directly, using a mechanism not otherwise covered by the model. For example, considering Figure 4.2, both AKT and 4E-BP1 are proliferation effector molecules. AKT affects proliferation, for example by affecting cell survival through the modulation of apoptosis (among other effects). Those mechanisms are not covered in more detail in the model, and therefore AKT is a parent of the proliferation node directly, to encompass these mechanisms. 4E-BP1 affects protein translation which is required for cell growth and proliferation, and is therefore also a parent of the proliferation node. AKT also signals to 4E-BP1 through mTOR, but since this is covered by the signaling network, the indirect effect that AKT has on proliferation through mTOR-4E-BP1 is not covered by the direct signal from AKT to proliferation. In this way, the separate signal from AKT to proliferation allows a quantification of the mTOR-independent proliferation effects of AKT.

Under treatment with the mth drug, the proliferation rate r will be affected by the drug's effect on the signaling molecules. In addition to these on-target effects, the off-target effects of the drug are modelled by directly inhibiting the proliferation as well, giving the following equation for the drug-treated proliferation rate:

$$r_{j,\text{inhibited}} = (r_b + \sum_{k \in \text{parents(proliferation)}} k_k A_{k,j,\text{inhibited}})(K_{\text{offtarget},m} + \frac{1 - K_{\text{offtarget},m}}{10^{h_{\text{offtarget},m}(c - c_{\text{IC50,offtarget},m})} + 1}). \tag{4.10}$$

Finally, the drug response is obtained by normalizing the proliferation under drug treatment at a particular concentration to the proliferation in the untreated conditions:

$$D_m = \frac{x_{0,j}\, e^{r_{j,\text{inhibited}}\, t_{\text{treatment}}}}{x_{0,j}\, e^{r_j\, t_{\text{treatment}}}}, \tag{4.11}$$

where D_m is the response to the mth drug, $x_{0,j}$ is the starting number of cells (i.e. the seeding density, which is the same for treated and untreated conditions) for cell line j and $t_{\text{treatment}}$ is the treatment duration (72 hours).

4.5.13. LIKELIHOOD

All the data points used in the inference are measurements of independent biological replicates. Each biological replicate was a new experiment, with each experiment done on a separate day with a new batch of cells. We can therefore treat all data points as independent observations, and the full likelihood can be simplified to a multiplication of the likelihood functions for each data point. This gives the following likelihood function, for each data type:

$$P(y|\theta) = \prod_{i \in \text{observed-variables}} \prod_{j \in \text{cell-lines}} \prod_{k \in \text{replicates}} P(y_{i,j,k}|\theta). \tag{4.12}$$

Here θ is a vector containing all model parameters, thus including all variables defined in Equations 4.5 to 4.11, as well as the measurement variances, σ, as defined below in Equation 4.13, and y represents the measurement data. Note that for a particular measurement value $y_{i,j,k}$, the likelihood function depends on a subset of these model parameters, namely all parameters affecting the corresponding model variable and its upstream signals. The observed variable set includes all observed variables used in the likelihood, thus i here indexes all these variables.

The data is divided into two classes. The first class contains the variables observed through DNA and RNA sequencing. For these variables, the corresponding model variables were set directly to the measured value and were therefore not included in the likelihood function. The second class comprises variables observed through the drug response assays, the proliferation measurements and RPPA (specifically the phosphorylation epitopes). For these variables we used a Student's t-distribution as likelihood function. This t-distribution was chosen as a means of robust inference, to accommodate outlying measurement values which cannot be adequately described by the model. Since the t-distribution is used solely as a means of robust inference, the number of degrees of freedom is fixed at three, rather than including this parameter as a latent variable. This gives the following likelihood function for each data point:

$$P(y_{i,j,k}|\theta) = t(y_{i,j,k}|\mu = x_{i,j}(\theta), \sigma = \sigma_i, \nu = 3), \tag{4.13}$$

where θ is again the vector containing all model parameters, $y_{i,j,k}$ is the measurement data for observed variable i, cell line j, and replicate k, $x_{i,j}$ is the modeled variable (defined further below in Equation 4.14) and σ_i is the variance of observed variable y_i. The variance σ_i for variable y_i is shared by all cell lines and biological replicates.

The modelled variable xi,j depends on the data type:

$$x_{i,j}(\boldsymbol{\theta}) \begin{cases} g_i + (1 - g_i) A_{i,j}(\boldsymbol{\theta}) & \text{for RPPA data} \\ r_j(\boldsymbol{\theta}) & \text{for growth data} \\ D_j(\boldsymbol{\theta}) & \text{for drug response data,} \end{cases} \tag{4.14}$$

where i again indexes over all observed variables, j indexes the cell lines and g_i is the background signal generated by aspecific binding of the antibody in the RPPA. In Figure 4.3, for the example of S6K phosphorylation, $x_{i,j}$ is represented by S6K phosphorylation (T389, RPPA), g_i is represented by S6K_T389 RPPA background, and σ_i is represented by S6K_T389 RPPA variance.

4

4.5.14. PRIOR

As prior, most parameters were given a uniform distribution, with exception of the proliferation signal rates (k) and measurement variances (σ) which were given exponential distributions, and the drug affinities(c_{IC50}) which were given semi-informative normal distributions. The precise prior distributions that were used are given in Supplementary Table 2. For the half-maximal inhibition concentrations of the drugs for their targets, the prior was set to a normal distribution on log10-scale with unit standard deviation, centered on the measurement of these parameters in biochemical assays found in the literature (see Supplementary Table 3), offset by 1 to account for a probable lower concentration of the drugs in the intracellular environment compared to the homogeneous in vitro environment.

4.5.15. POSTERIOR PREDICTIVE

The posterior predictive distribution is the probability distribution of a new set of data, given the model and the observed data. This distribution was approximated from the posterior Monte Carlo samples. The posterior predictive distribution includes two aspects of uncertainty. Firstly, the uncertainty in the model parameters, described by the posterior gives rise to an uncertainty in the regulatory signals and in the drug response. Secondly, there is an uncertainty in the data measurements themselves – and therefore also in future predicted measurements – which is reflected by the t-distributions in the likelihood. The posterior predictive distribution accounts for both sources of uncertainty, the parametric uncertainty as well as the data uncertainty.

4.5.16. MODEL INFERENCE

The posterior probability distributions and marginal likelihoods were calculated by Monte Carlo sampling using the BCM software package [19]. We used two variants of Monte Carlo: the posterior for all models was sampled with parallel tempered Markov Chain Monte Carlo (PT-MCMC) [54], and for each model iteration we also sampled the posterior for at least one drug with sequential Monte Carlo (SMC) [55] to verify that a different sampling algorithm gave similar results. Both sampling methods were run until apparent convergence and this convergence was then verified by checking whether the two methods gave the same result.

PARALLEL-TEMPERED MARKOV CHAIN MONTE CARLO

We used the algorithm described in [54] with slight modification. We started the inference with a pilot run with the following configuration: 32 parallel chains, a subsampling of 1 in 200, a burn-in period of 250 and a sampling period of 1,000. We updated the variances of the proposal distribution every 250 samples. One sample (before subsampling) corresponds to updating all parameters once. After this initial sampling period, we optimized the temperature schedule twice, with each subsequent sampling period having 5 times as many samples as the previous period. This configuration was sufficient to get at least 100 round trips from prior to posterior, for most drugs, and virtually no autocorrelation after subsampling.

As proposal distribution, we used the strategy from [56], to block the correlated parameters together. We start with a scalar Gaussian proposal distribution for each parameter. Then at each proposal-update-step, we calculated the empirical correlation between all parameters over the last period. We clustered this correlation matrix using hierarchical clustering with complete linkage. We then cut the hierarchical tree at a correlation of 0.5. In the next sampling period, the groups of parameters which are clustered together were then sampled together using a full covariance Gaussian proposal distribution. Parameters which were not correlated with any of the other parameters continued to use the scalar Gaussian proposal distribution.

The scale of the proposal distributions was continuously adapted to maintain a constant acceptance rate of 23% for each parameter or block thereof. The current average acceptance rate was calculated using an exponentially moving average with period equal to 1/10th of the proposal update interval period.

The marginal likelihood was calculated using thermodynamic integration with a trapezoidal integration rule [57].

SEQUENTIAL MONTE CARLO

We used the algorithm described in [55], with the temperature schedule automation described in [58]. Although the schedule automation of [58] is developed for Approximate Bayesian Computation (ABC), we did not use the ABC-approximation; we continued to use the full likelihood calculation.

We used the following configuration: a population size of 5120, an effective sample size ratio between each iteration of 0.99 (this is parameter α in [58]), and resampling when the effective sample size drops below half the population size. Resampling was done using residual resampling, which gave a lower estimation variance than multinomial resampling.

As proposal distribution we used a Markov chain Monte Carlo kernel, taking as many MCMC steps as necessary for the correlation with the previous temperature to be less than 0.95, for each parameter. This typically started with only a few MCMC steps at initial temperatures and increased to several hundred or thousand steps when the sampling approaches the posterior. Having a correlation of at most 0.95 was sufficient to prevent any correlation of the samples between resample steps.

As proposal of the MCMC chain we used a diagonal multivariate Gaussian distribution, with the diagonal variances based on the empirical variance of the samples of the previous population. This kernel was then scaled to achieve an average acceptance rate

of 23%. The backward kernels for calculating the weights were taken as described in section 3.3.2.3 in [55].

The marginal likelihood was calculated using the sample weights at each resample step, as described by equation 15 in [55].

4.5.17. CONVERGENCE MONITORING

Convergence was monitored in two ways: firstly by monitoring the convergence of the individual sampling methods, and secondly by comparing the results of the two sampling methods. The individual sampling methods were monitored as described below.

Convergence for PT-MCMC was monitored in three ways: by calculating the autocorrelation for all parameters, by visually inspecting the traces of all variables and of the posterior probability, and by monitoring the number of round trips from prior to posterior.

The autocorrelation in each parameter was calculated for lag $\tau = 1$ up to $\tau = N$, and the lag at which the autocorrelation dropped below $2/\sqrt{N}$ was required to be less than 5, where N is the sampling period of 1,000. Supplementary Figure 9 shows the traces for the posterior probability and for the parameter with the strongest autocorrelation, for the model in the context of lapatinib treatment. Both traces do not show any obvious autocorrelation or other patterns. In a scatter plot for the two most-correlated parameters (right panels of Supplementary Figure 9) no signs of inhomogeneously distributed samples are visible.

To ensure that there was sufficient global exploration, we monitored the number of round trips from prior to posterior. We required that there were at least 100 such round trips. This round trip is defined as the chain swapping steps required for a particular sample to be at the chain sampling from temperature=0, reach temperature=1, and diffuse back to temperature=0, independently of how it is perturbed by the MCMC kernel. We assume that when a sample is at the chain sampling from the prior, it will be immediately uncorrelated, and thus represent a new global starting point.

Convergence for SMC was monitored in two ways: by calculating the correlation between samples at resampling steps, and by visual inspection of the variable traces. Due to the residual resampling, samples which are duplicated during the resampling step are located in succession. It can thus be easily spotted when these duplicated samples are not sufficiently perturbed by the MCMC sampling kernel; no evidence of this is present in the sample traces.

4.5.18. MODEL COMPARISONS

To compare different models, we calculated the marginal likelihood along with the posterior probability distribution during the Monte Carlo sampling. For PT-MCMC, we used thermodynamic integration across the parallel chains [57]. For SMC, we used the weights at resampling steps as described in [55]. In both cases, sufficient samples were generated, and for PT-MCMC sufficient parallel chains were used, such that the estimated approximation error of the marginal likelihood was at most 1 on natural log scale. When comparing two models, a marginal likelihood of at least 3 points higher on natural log scale was taken as sufficient evidence for one model over the other [27]. Note that the marginal likelihood inherently penalizes models with too many parameters. Neverthe-

less, when marginal likelihoods of two models were comparable, we used the simpler of the two to keep the computation time of subsequent models manageable.

4.5.19. WEBSITE VISUALIZATION

The accompanying website, http://ccb.nki.nl/software/BCCL_KI_response_model/, shows the model estimates for each cell line, drug and drug concentrations. The strength for each link, shown as the level of gray, is the posterior mean of the part of Equation 4.5 that corresponds to that link; that is, each element $s_{k,i}A_{k,j}$. The transparency indicates the uncertainty, quantified by the standard deviation of the posterior. The estimates are from the last model iteration, including the model adaptations of foretinib-FGFR2 and the protein expression levels.

4.5.20. DIFFERENTIAL EXPRESSION ANALYSIS

The cell lines were divided into sensitive and resistant cell lines using Gaussian mixture modelling. A one-component model and a two-component model were fitted to log-transformed IC50 values for each drug. If the Bayesian information criterion for the two-component model was bigger than for the one-component model, the cells were classified as resistant (highest component) and sensitive (lowest component); otherwise no differential expression tests were performed. Differential expression was tested with t-tests, and corrected for multiple testing using false-discovery rate correction. The cell line classification and the subsequent differential expression results are provided in Supplementary Table 14.

4.5.21. ADDITIONAL MODEL DETAILS

Most cell lines which have lost PTEN expression have a frameshift or nonsense mutation in the gene. There are also cell lines with a missense mutation: CAMA1 has a D92H mutations and MM453 has an E307K mutation. The D92H mutation has been shown to abolish all PIP3 phosphatase activity [59], so we modelled this mutation simply as a loss of the protein. The E307K mutation is a gain-of-function mutation leading to increased membrane localization [60]; this mutation was therefore not modelled as a loss of the protein but rather as a gain of function mutation.

The RPPA data includes a validated antibody for PDK1 phosphorylation at S241. PDK1 phosphorylation at this epitope does not show any correlation with Akt phosphorylation at either T308 or S473 in our data. We assume that rather than the phosphorylation at S241, it is the co-localization of PDK1 and AKT which relays the signal from PIP3 through PDK1 to AKT. We therefore did not include PDK1 phosphorylation in the model inference.

According to the MD Anderson RPPA core facility, the antibody for EGFR phosphorylation at Y1068 likely sees ERBB2 phosphorylation at Y1248 as well, and vice versa. We accounted for this cross-reactivity by summing the two signals together before fitting them to the RPPA data, with an additional parameter specifying the amount of cross-reactivity. The various dimerization possibilities of the RTKs are not included in the model, as this was not identifiable with the present data.

4.5.22. SUPPLEMENTARY NOTE 1 – INFERRED SIGNALING ESTIMATES AGREE WITH ON-TREATMENT PHOSPHORYLATION MEASUREMENTS

The signaling activity estimates upon inhibitor treatment are inferred from untreated molecular data in combination with relative viability data after drug treatment. To further test whether the inferred on-treatment signaling activities are accurate, we compared them with measurements of phosphorylation levels of cell lines while under treatment. For this, we used the data provided by Korkola et al [28]. They performed time course RPPA measurements of 15 cell lines after treatment with lapatinib, the AKT inhibitor GSK690693, and a combination of the two. Of these 15 cell lines, 9 overlap with our cell line panel. Since we modeled steady state levels rather than time courses, we averaged the measured phosphorylation levels over the time points from 1 hour to 72 hours (the trajectories generally converge to a steady state; using individual time points such as the first or last point gives similar results). We then compared these averaged phosphorylation levels, for the lapatinib-treated condition, to the inferred signaling activities, also in the lapatinib-treated condition. Three of the five overlapping epitopes show a significant correlation, while the other two epitopes at least show the correct trend as well (Supplementary Figure 18). The scale is not always correctly predicted, but this is to be expected given that the model does not know the antibody binding affinity in this external dataset. This comparison indicates that the model is capable of producing reasonable estimates of on-treatment phosphorylation levels, based on pre-treatment phosphorylation levels combined with relative viability after treatment.

4.5.23. SUPPLEMENTARY NOTE 2 – A SIMPLIFIED MODEL CAN DESCRIBE THE RESPONSE TO SEVEN DRUGS SIMULTANEOUSLY

Having established a model that can explain the majority of the variability in response for the drugs in isolation, we wished to obtain a model that can explain the drug response of all drugs simultaneously. The signaling estimates of such a model would be a better representation of the cell lines, as they take the response to all the drugs into account. A joint model for all drugs also allows the exploration of drug combination effects in the future.

Due to the large computational demands, it was not feasible to fit the large model presented in the main text to seven drug response profiles simultaneously. To nevertheless obtain a joint model, we selected the signaling events and genetic aberrations that were most important for explaining the drug response for each drug in isolation. For example, for lapatinib the factors of ERBB2 amplification, PIK3CA and PIK3R1 mutations were important, but EGF expression contributed only a small part, and the contribution of HGF expression could not be identified. These latter two factors, EGF and HGF expression, were also not major factors for explaining response to the other drugs. For the mTOR inhibitors, 4EBP1 protein expression was an important contributor. Thus, from among these examples, we kept the factors of ERBB2 amplification, PIK3CA and PIK3R1 mutations and 4EBP1 expression in the joint model, but removed EGF and HGF expression. The resulting reduced model is shown in Supplementary Figure 13A. This model was then fitted to the seven drugs simultaneously. To adequately represent the drug response to the seven drugs simultaneously, it was necessary to add an additional nonlinearity to the model, specifically, the link from AKT to proliferation. This is due to the

relatively low AKT phosphorylation levels in the ERBB2-amplified cell lines, while other evidence suggests sensitivity to lapatinib is still determined through the AKT pathway. In order to allow a reduction from low to very low AKT activation to still affect proliferation, and thereby reconcile these two observations, a non-linear function was added specifically to this link. Ideally all activation functions would be non-linear (with the ability to reproduce linear functions with certain parameter values), but this would introduce too many parameters, thus we only added a non-linear function where necessary. Additionally, since there are now seven drugs being fitted simultaneously, the amount of drug response data (6,300 data points) heavily outweighs the phosphorylation data (90 data points). To balance this, we increased the weight of the phosphorylation data by 10, thus making sure the model does not ignore the phosphorylation data.

The resulting model fit is shown in Supplementary Figure 13B. The sensitivity to mTOR-inhibitors can be described well by this model. For lapatinib and the PI3K inhibitor GDC-0941, two cell lines still stand out (MDA-MB-468 for lapatinib and HCC38 for GDC-0941). For the MEK-inhibitor PD0325901, the single most sensitive cell line (MDA-MB-231) is described well, and for foretinib the sensitivity of MFM-223 is correctly recapitulated. The Akt inhibitor MM2206 shows the most discrepancies, with several cell lines showing unexplained behavior. In this case differential expression analysis did not give significant associations to provide additional clues for model extensions, and further experiments would be necessary to investigate these discrepancies further. Finally, apart from the drug response data, the phosphorylation data is also adequately described by the joint model (Supplementary Data 1). We conclude that the procedure of iteratively creating a literature-based model, followed by data-driven model expansion and subsequent model reduction can be used to construct a model capable of describing the majority of the variability in short-term drug response.

4.5.24. SUPPLEMENTARY NOTE 3 – INCLUSION OF FEEDBACK SIGNALING FROM MTORC1 TO PI3K DOES NOT IMPROVE THE FIT FOR DRUG RESPONSE TO AZD8055 AND LAPATINIB

We developed an extended version of the ISA modeling framework that is capable of including feedback signaling events [29]. Calculating the signaling activities in a model with feedback is computationally more expensive, which makes it impractical to infer the signaling activities of the large model presented in this manuscript while also including feedback events. To nevertheless explore the effect of feedback signaling on drug response, we constructed a small model of AKT signaling which includes the main explanatory factors for response to lapatinib and AZD8055 as identified in the large model presented in this manuscript. The AKT pathway was chosen for this test, since the cell lines show most variability to inhibitors in this pathway.

Supplementary Figure 17 shows the model that was used, and the goodness of fit for the drug response of AZD8055 and lapatinib, with and without the well-known feedback from MTORC1 to PI3K through IRS1 [61, 62]. It is clear that the goodness of fit is comparable for both models. For lapatinib, the evidence in fact decreases when including the feedback loop; any improvement in fit is very small, and it is outweighed by the addition of an extra free parameter. For AZD8055, the evidence increases with the addition of the feedback loop, meaning that there is indeed support for inferring the activity of

the feedback loop. Based on the evidence from the posterior distributions, the feedback loop is likely to be active (Supplementary Figure 17 C and E); see also [29] for a further exploration of the identifiability of feedback activities. However, despite the feedback loop being active and an improvement in overall goodness of fit, the drug sensitivity of the same number of cell lines is explained by both models. For AZD8055, both with and without feedback loops there are seven cell lines with a sum of squared error > 0.1.

Given that the computational cost of large models with feedback loops is so large, it is presently not feasible to test the effect of including feedback loops in large models, such as those employed in this manuscript. We therefore cannot fully rule out that the addition of feedback loops, in a large model, could be able to better explain the variability in drug response. Indeed feedback signaling is an important feature of cellular regulatory networks, and likely to be important in many situations. However, it appears that feedback is not necessary for explaining the majority of the variability in short-term drug response, as demonstrated for the two kinase inhibitors we studied in this note.

Cell line	Short name	Growth medium	Source	Catalogue #	DR seeding
BT-20	BT20	MEM + 10% FBS + Penstrep	ATCC	HTB-19	1250
BT-474	BT474	DMEM/F12 + 10% FBS + Penstrep	ATCC	HTB-20	5000
BT-549	BT549	RPMI + 10% FBS + 1 ug/ml insulin + Penstrep	ATCC	HTB-122	500
CAL-120	CAL120	DMEM + 10% FBS + Penstrep	DSMZ	ACC 459	500
CAL-148	CAL148	DMEM + 20% FBS + 10 ng/ml EGF + Penstrep	DSMZ	ACC 460	1500
CAL-51	CAL51	DMEM + 20% FBS + Penstrep	DSMZ	ACC 302	625
CAMA-1	CAMA1	MEM + 10% FBS + Penstrep	ATCC	HTB-21	5000
HCC1187	HCC1187	RPMI + 10% FBS + Penstrep	ATCC	CRL-2322	5000
HCC1395	HCC1395	RPMI + 10% FBS + Penstrep	ATCC	CRL-2324	1500
HCC1419	HCC1419	RPMI + 10% FBS + Penstrep	ATCC	CRL-2326	5000
HCC1500	HCC1500	RPMI + 10% FBS + Penstrep	ATCC	CRL-2329	10000
HCC1569	HCC1569	RPMI + 10% FBS + Penstrep	ATCC	CRL-2330	2000
HCC1806	HCC1806	RPMI + 10% FBS + Penstrep	ATCC	CRL-2335	800
HCC1937	HCC1937	RPMI + 10% FBS + Penstrep	ATCC	CRL-2336	2000
HCC1954	HCC1954	RPMI + 10% FBS + Penstrep	ATCC	CRL-2338	1250
HCC38	HCC38	RPMI + 10% FBS + Penstrep	ATCC	CRL-2314	800
HCC70	HCC70	RPMI + 10% FBS + Penstrep	ATCC	CRL-2315	5000
Hs 578T	HS578T	RPMI + 10% FBS + 10 ug/ml insulin + Penstrep	ATCC	HTB-126	800
MCF7	MCF7	RPMI + 10% FBS + 10 ug/ml insulin + Penstrep	ATCC	HTB-22	2500
MDA-MB-157	MM157	RPMI + 10% FBS + Penstrep	ATCC	HTB-24	1250
MDA-MB-231	MM231	RPMI + 10% FBS + Penstrep	ATCC	HTB-26	2500
MDA-MB-361	MM361	RPMI + 20% FBS + Penstrep	ATCC	HTB-27	5000
MDA-MB-436	MM436	RPMI + 10% FBS + 10 ug/ml insulin + 16 ug/ml glutathione + Penstrep	ATCC	HTB-130	2000
MDA-MB-453	MM453	RPMI + 10% FBS + Penstrep	ATCC	HTB-131	2500
MDA-MB-468	MM469	RPMI + 10% FBS + Penstrep	ATCC	HTB-132	2500
MFM-223	MFM223	MEM + 15% FBS + 2 mM GlutaMax + 1xITS + Penstrep	DSMZ	ACC 422	2000
SK-BR-3	SKBR3	RPMI + 10% FBS + Penstrep	ATCC	HTB-30	1000
SK-BR-7	SKBR7	DMEM/F12 + 10% FBS + Penstrep	See ref		1250
T47D	T47D	RPMI + 10% FBS + 10 ug/ml insulin + Penstrep	ATCC	HTB-133	2500
ZR-75-30	ZR7530	RPMI + 10% FBS + Penstrep	ATCC	CRL-1504	5000

Table 4.1: **Supplementary Table 1**

Parameter	Description	Prior
b	Base signal	Uniform(a=0,b=1)
s	Signal strength	Uniform(a=0,b=5)
s_{mut}	Mutation -> signal	Uniform(a=0,b=1)
p	Protein expression coefficient	Uniform(a=0,b=1)
k	Signal -> proliferation	Exponential($\lambda = 25$)
$c_{\text{IC50,ontarget}}$	Drug affinity for target (x = in vitro IC50 + 1)	Normal($\mu = x, \sigma = 1$) – log10 scale
$c_{\text{IC50,offtarget}}$	IC50 for off-target effects	Uniform(a=-6,b=0) – log10 scale
h	Steepness of dose-response effects	Uniform(a=-1,b=1) – log10 scale
K	Max inhibition by drugs	Uniform(a=0,b=1)
σ_{RPPA}	Variance for RPPA data	Exponential($\lambda = 5$)
$\sigma_{\text{growthrate}}$	Variance for growth rate	Exponential($\lambda = 100$)
$\sigma_{\text{drugresponse}}$	Variance for drug response data	Exponential($\lambda = 10$)
g	RPPA background signal	Uniform(a=0,b=1)

Table 4.2: **Supplementary Table 2** We used the following prior distributions. The drug affinities are assigned semi-informative priors based on the biochemically measured affinities reported in the literature.

Drug	Target	Target inhibition IC50	Reference
AZD8055	mTORC1/2	0.8 nM	[63]
BEZ235	mTORC1/2	20 nM	[64]
BEZ235	PI3K	5 nM	[64]
GDC0941	PI3K	3 nM	[65]
Foretinib	MET	0.4 nM	[66]
Foretinib	FGFR2	N/A, but likely <50 nM	[25]
Lapatinib	EGFR	10.8 nM	[22]
Lapatinib	ERBB2	9.2 nM	[22]
MK2206	AKT1	8 nM	[67]
MK2206	AKT2	12 nM	[67]
MK2206	AKT3	65 nM	[67]
PD0325901	MEK	0.33 nM (in colon cells)	[68]

Table 4.3: **Supplementary Table 3** The following table lists the drug targets that were included in the model, along with the concentrations at which they inhibit their target's activity by 50%, as measured by in vitro assays using the target enzyme and a substrate. Some drugs may have other targets beyond the ones listed here; the table lists the targets which are included as signaling molecules in the model.

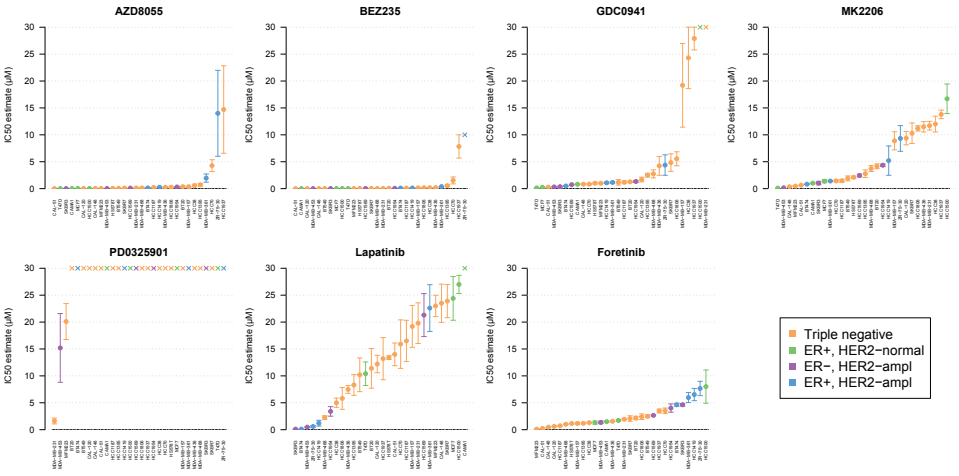


Figure 4.9: **Supplementary Figure 1**: Summary of the drug response data as IC50 estimates. Error bars indicate SEM and a cross indicates that a 50% viability reduction was not achieved in the screened concentration range.

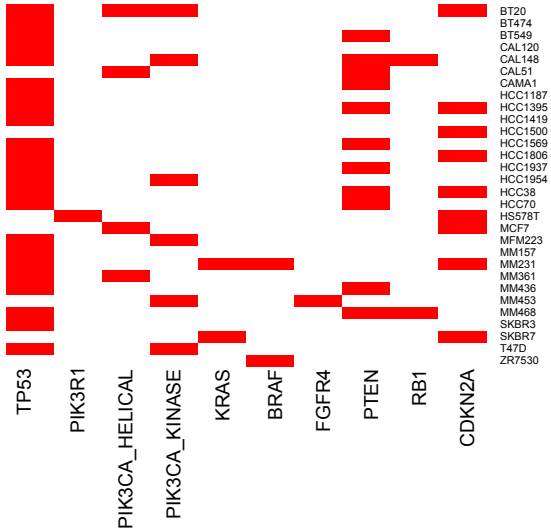


Figure 4.10: **Supplementary Figure 2**: Summary of a subset of the mutation data. Red boxes indicate that the particular gene is mutated in that cell line.

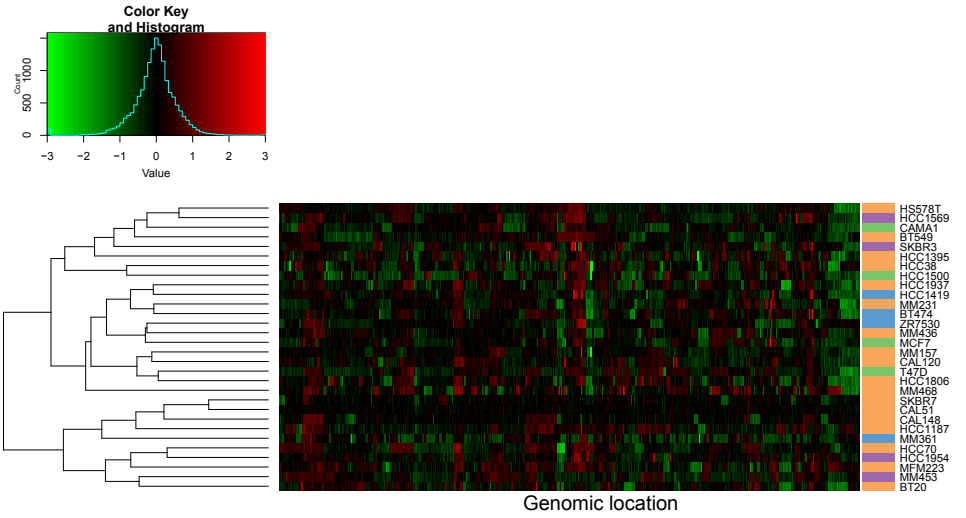


Figure 4.11: **Supplementary Figure 3**: Summary of the copy number estimates obtained from the off-target DNAseq reads using CopywriteR. Copy number estimates are log2 ratios of the region against the sample's average. Cell lines were clustered using correlation distance and Ward linkage.

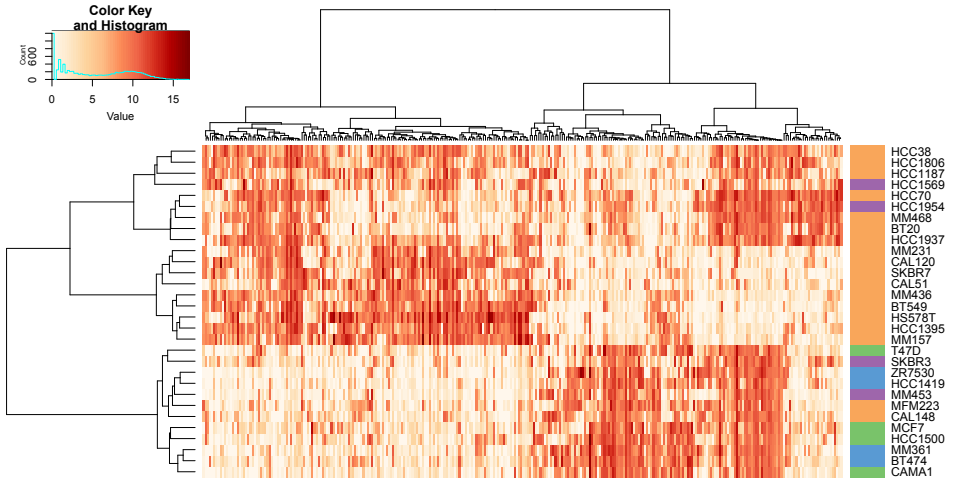


Figure 4.12: **Supplementary Figure 4**: Summary of the RNAseq data. The mRNA expression levels are log-transformed, normalized read counts. Genes were selected by taking the 300 most varying genes. Cell lines and genes were clustered using correlation distance and Ward linkage.

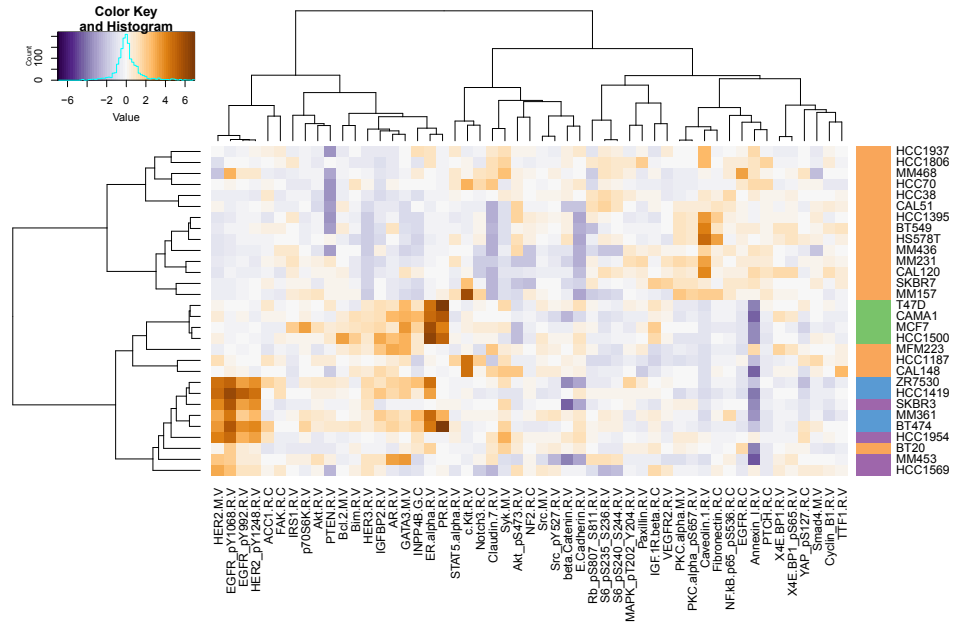


Figure 4.13: **Supplementary Figure 5**: Summary of the RPPA data. The protein expression levels are log-transformed, normalized RPPA signals. Epitopes were selected by taking the epitopes with variance > 0.3. Cell lines and epitopes were clustered using correlation distance and Ward linkage.

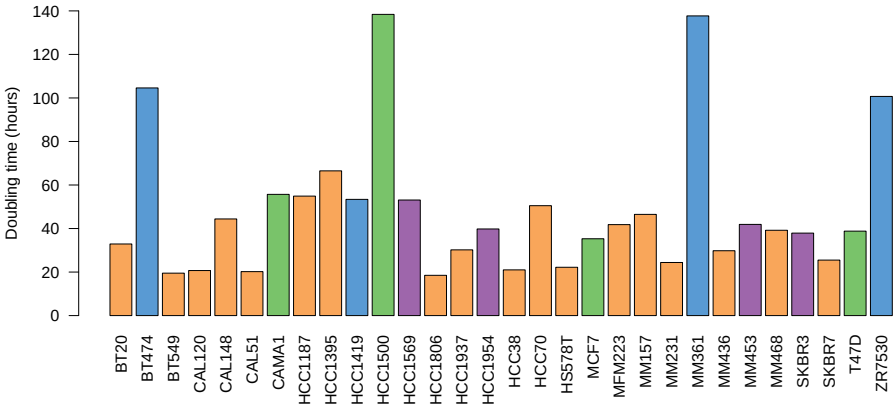


Figure 4.14: **Supplementary Figure 6**: Summary of untreated, steady-state proliferation data.

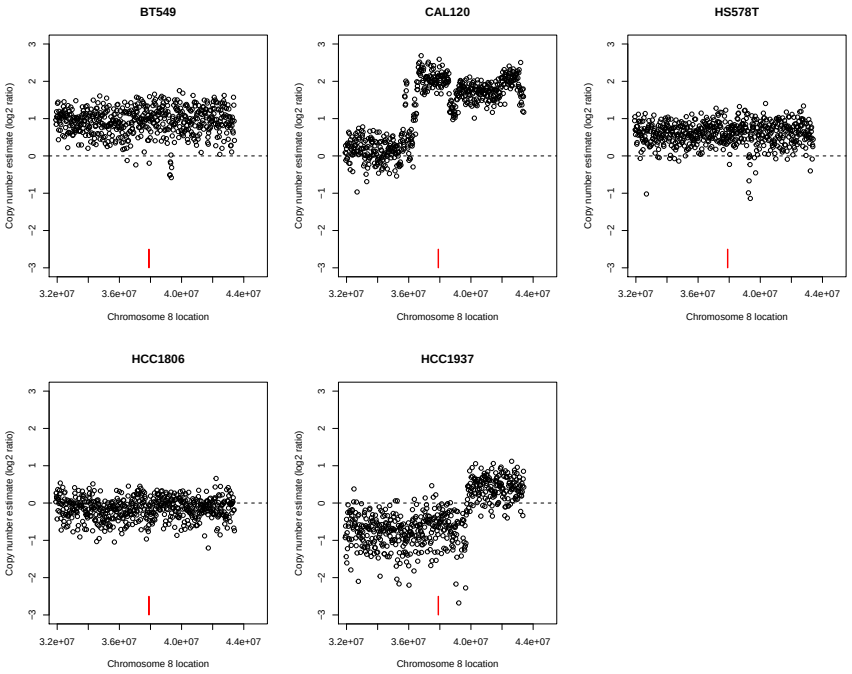


Figure 4.15: **Supplementary Figure 7**: Copy number estimates for the 8p11-12 locus for the five cell lines. Copy number estimates are log2 ratios of the locus against the sample's average. Each point represents a 20kb bin. The red box indicates the EIF4EBP1 gene. Cal-120 has a clearly focal amplification at 8p11-12, and BT549 and HS578T have a broader amplification.

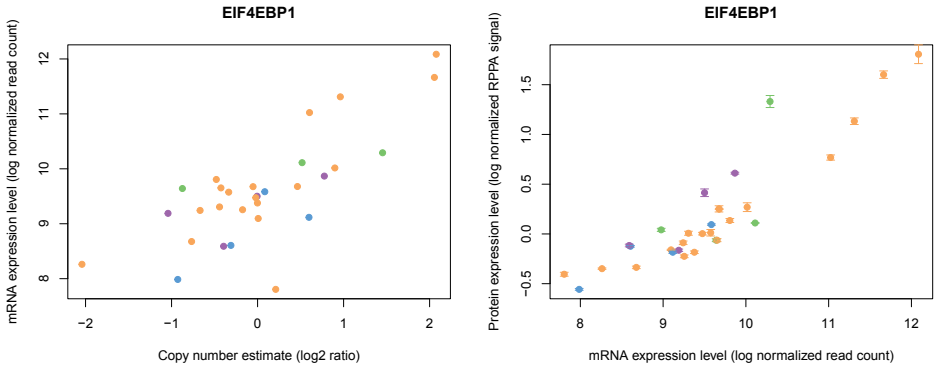


Figure 4.16: **Supplementary Figure 8**: Correlation of EIF4EBP1 copy number, mRNA expression level and protein expression level. Copy number estimates are log2 ratios of the region against the sample average. mRNA expression levels are log-transformed, normalized read counts. Protein expression levels are log-transformed, normalized RPPA signals. Error bars indicate SEM.

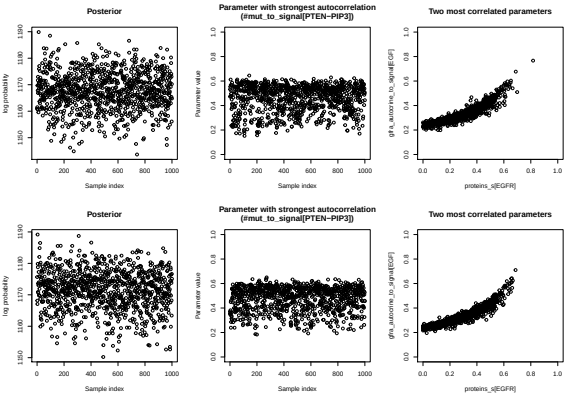


Figure 4.17: **Supplementary Figure 9**: Sample traces for convergence monitoring. The top row are samples obtained with PT-MCMC, the bottom row with SMC.

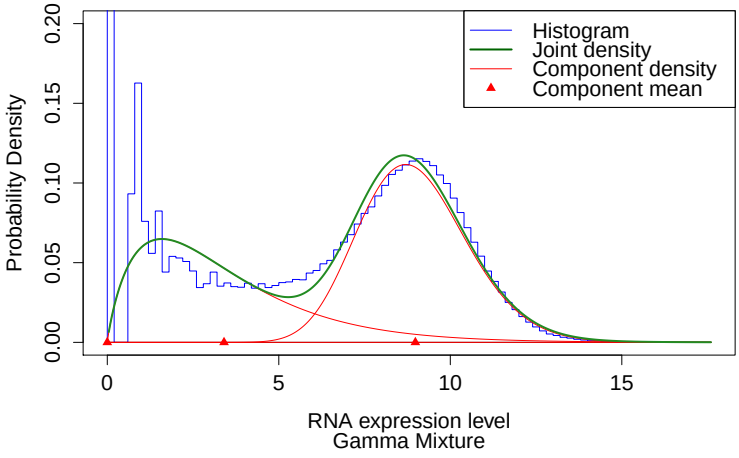


Figure 4.18: **Supplementary Figure 10**: Gamma mixture modeling for the binarization of the RNAseq data. RNA expression levels are log-transformed, normalized read counts.

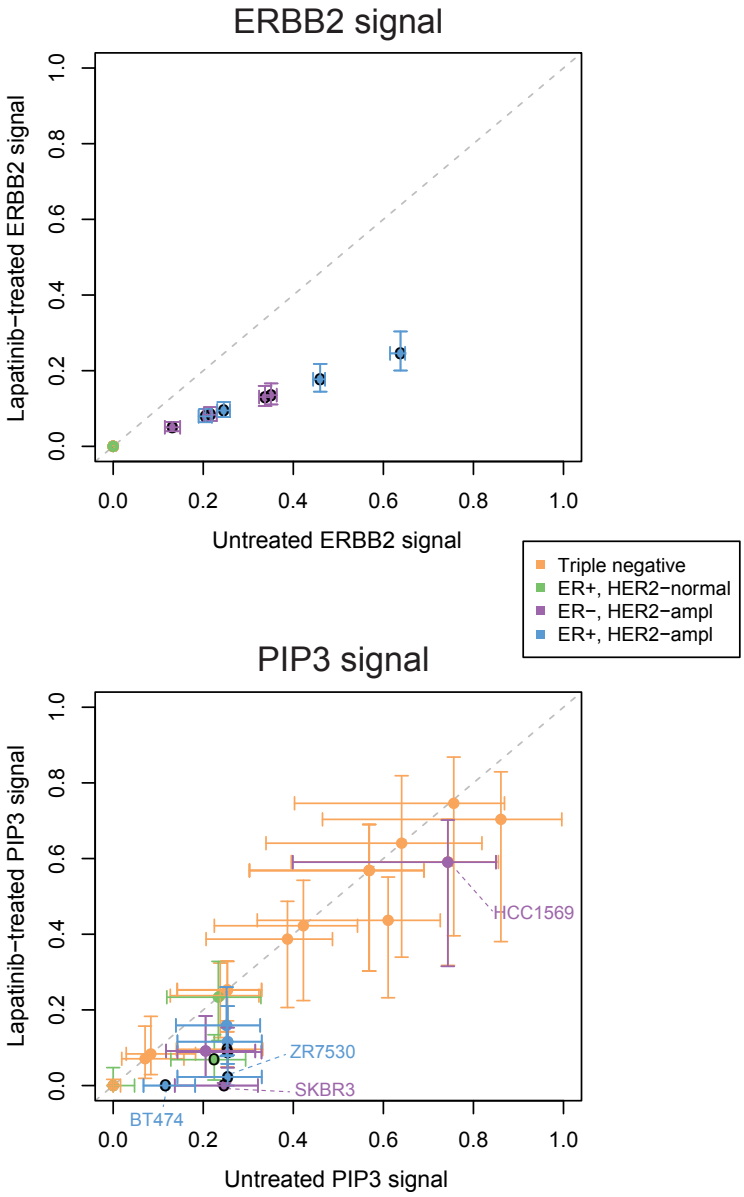


Figure 4.19: **Supplementary Figure 11**: Estimates of the activity of two signaling molecules, ERBB2 and PIP3, in untreated and lapatinib-treated conditions (expansion of Fig 3B). A black circle around a point indicates significant difference (posterior probability > 0.975 for the lapatinib-treated signal being less than the untreated signal). Error bars indicate the 90% confidence interval.

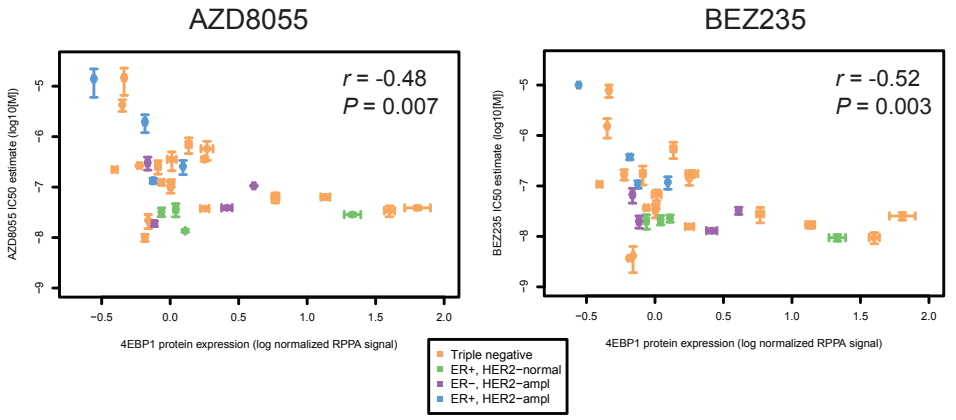


Figure 4.20: **Supplementary Figure 12**: Scatter plot of 4E-BP1 protein expression levels with sensitivity estimates to the two mTOR-inhibitors. Error bars indicate standard SEM.

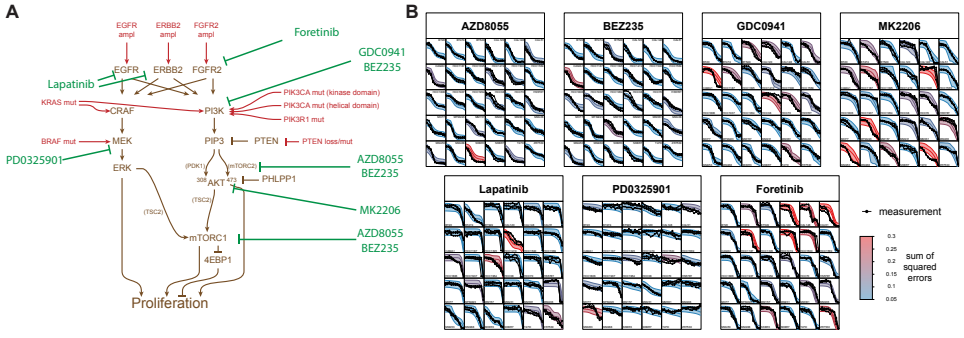


Figure 4.21: **Supplementary Figure 13**: Single model fitted simultaneously to the seven kinase inhibitors. (A) Schematic of the factors included in the reduced model. (B) Posterior predictive of the drug response to the seven kinase inhibitors.

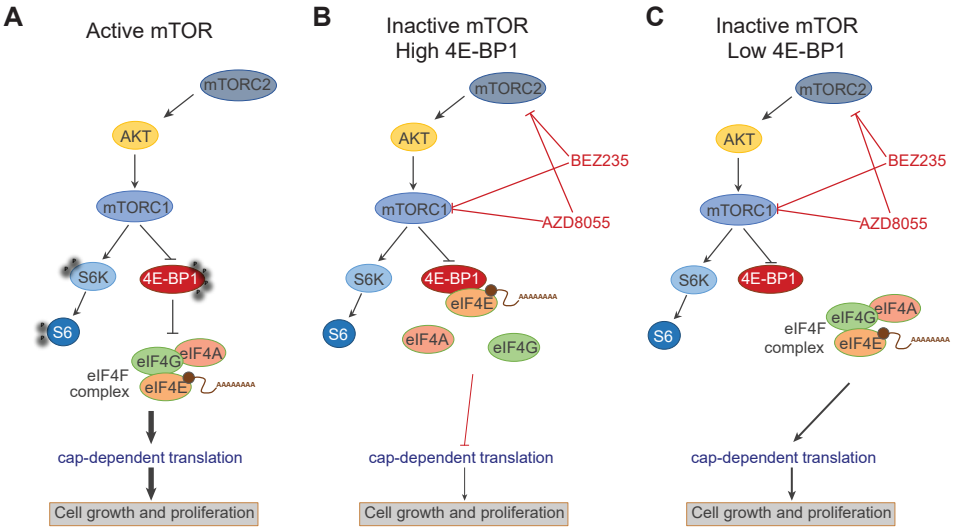


Figure 4.22: **Supplementary Figure 14**: Schematic of 4E-BP1 regulation and activity in the context of mTOR inhibitor treatment. (A) Under conditions where nutrients and growth factors are not limiting, active mTOR is able to phosphorylate and inactivate the 4E-BP1 protein. This allows the eIF4F translation initiation complex, composed of several proteins including eIF4E, eIF4G and eIF4A, to bind mRNA and drive cap-dependent translation, cell growth and proliferation. (B) Following treatment with mTOR inhibitors (or under conditions of nutrient or growth factor starvation), 4E-BP1 is no longer phosphorylated by mTOR, and becomes activated. This species of 4E-BP1 can bind the eIF4E translation initiation factor and sequester it from the eIF4F complex, attenuating cap-dependent translation, cell growth and proliferation. Our results indicate that a minimum threshold of 4E-BP1 expression is required in order to efficiently inhibit cap-dependent translation following mTOR inhibitor treatment. (C) In a situation where 4E-BP1 protein levels are below this threshold, eIF4E is not sufficiently sequestered away from the eIF4F complex, leading to ongoing cap-dependent translation and thus reduced sensitivity to mTOR inhibitors.

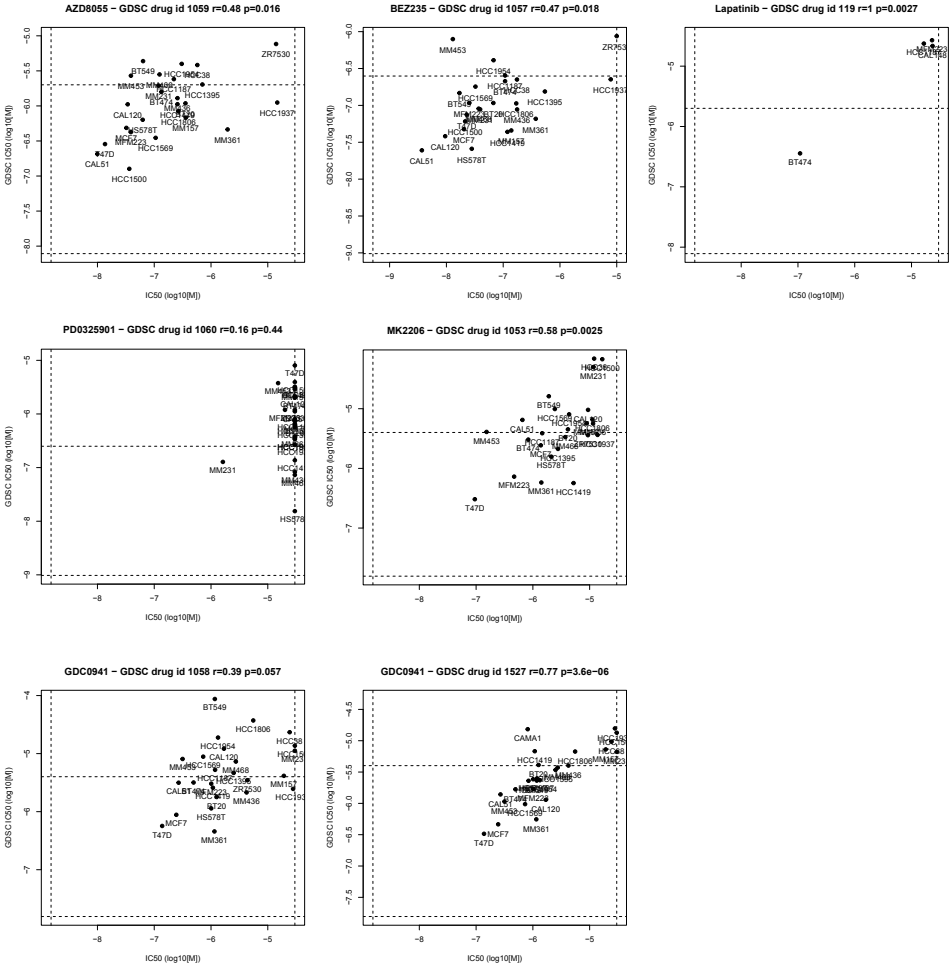


Figure 4.23: **Supplementary Figure 15**: Comparison of IC50 estimates with the Genomics of Drug Sensitivity in Cancer screen. Grey dashed lines indicate the screening ranges used in this work (vertical) and by GDSC (horizontal).

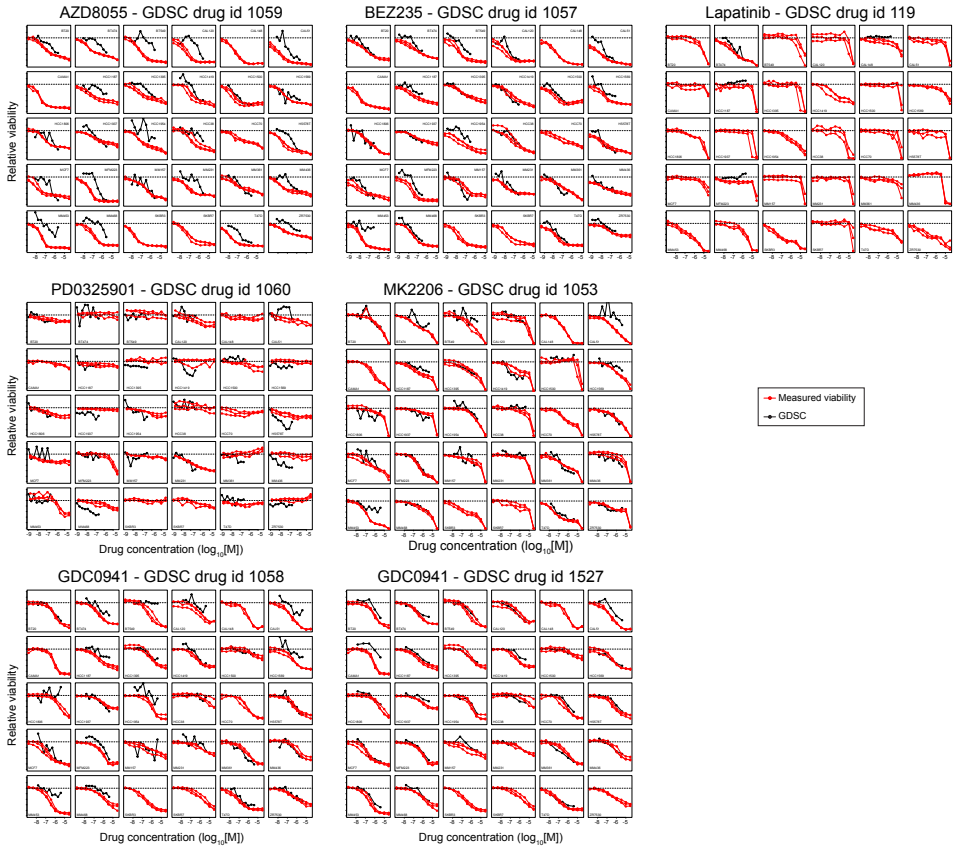


Figure 4.24: **Supplementary Figure 16**: Comparison of drug response data with the Genomics of Drug Sensitivity in Cancer screen.

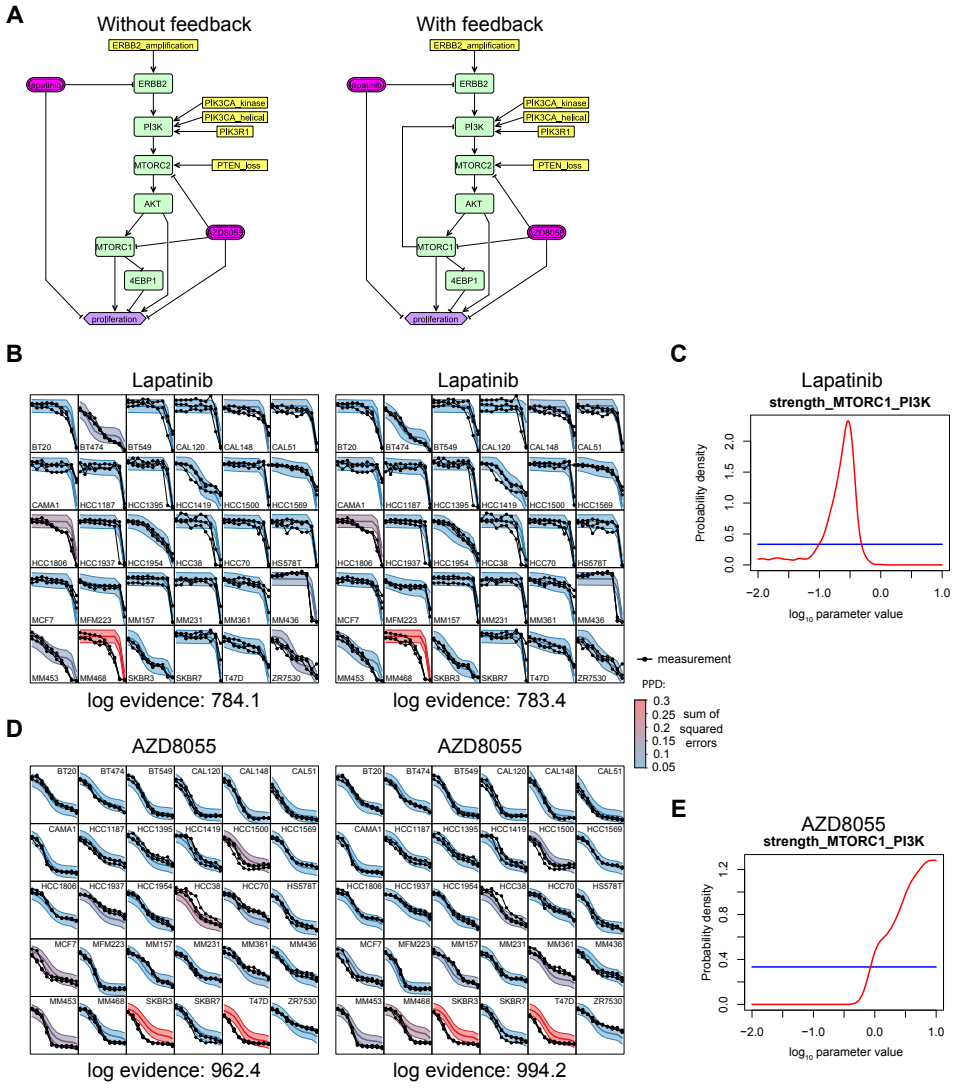


Figure 4.25: **Supplementary Figure 17**: Signaling model with and without feedback for the response of lapatinib and AZD8055. (A) Model structure in SBGN format, specified using feedback-ISA [29]. (B) Goodness of fit for lapatinib response. (C) Estimate of the activity of the feedback loop from MTORC1 to PI3K. (D) Goodness of fit for AZD8055 response. (E) As in C, now for AZD8055 response.

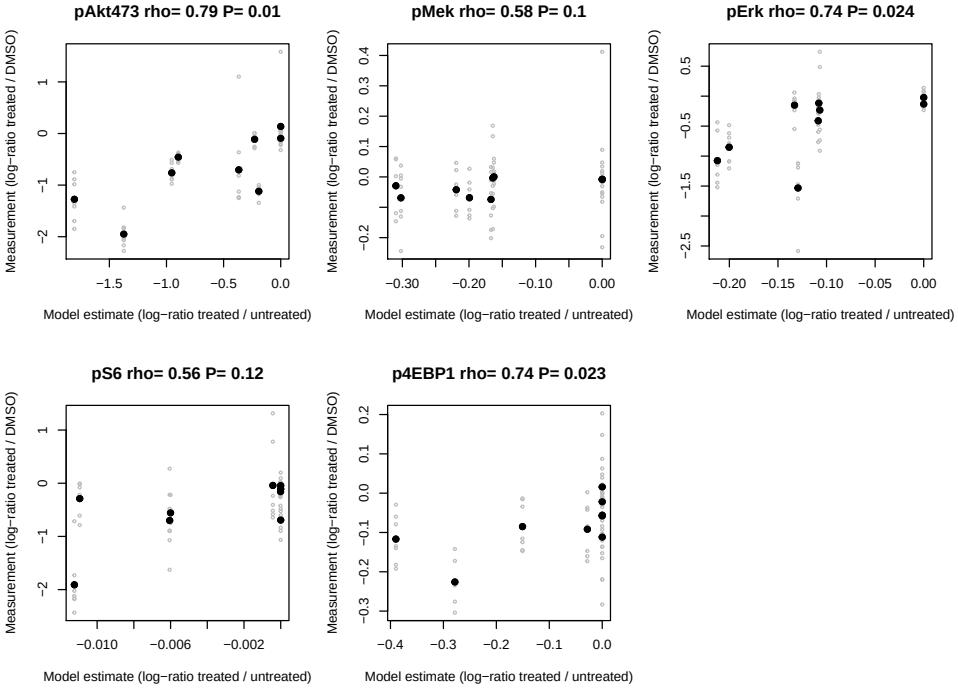


Figure 4.26: **Supplementary Figure 18**: Comparison of signaling estimates with on-treatment phosphorylation measurements by Korkola et al [28]. Each black dot indicates a cell line, showing the mean model prediction plotted against the mean on-treatment measurement value. The on-treatment measurements were summarized by taking the mean of the log-transformed measurement values over the time points from 1 hour to 72 hours (individual time points are shown in grey). For both the measured and inferred signals, the log ratio between treated condition and the untreated (or DMSO control) condition is depicted.

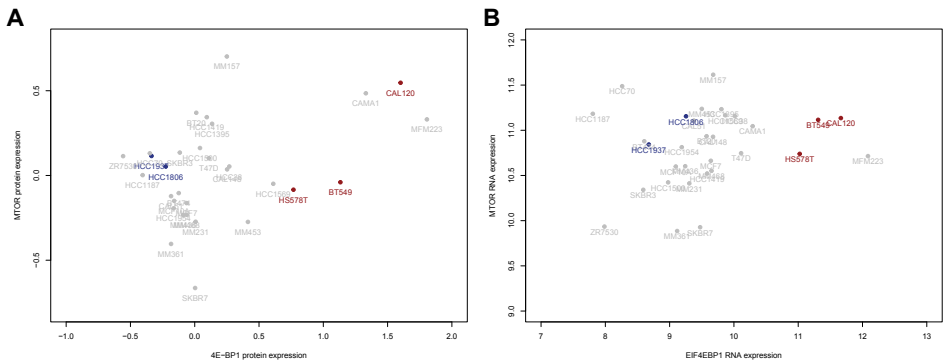


Figure 4.27: **Supplementary Figure 19**: Correlation between 4EBP1 and MTOR expression.

5

DELINEATING FEEDBACK ACTIVITY IN THE MAPK AND AKT PATHWAYS USING FEEDBACK-ENABLED INFERENCE OF SIGNALING ACTIVITY

Bram THIJSSEN
Katarzyna JASTRZEBSKI
Roderick L. BEIERSBERGEN
Lodewyk F.A. WESSELS

Parts of this chapter have been posted on bioRxiv (268359).

ABSTRACT

An important aspect of cellular signaling networks is the existence of feedback mechanisms. However, due to the complexity of signaling networks, as well as the presence of multiple interrelated feedback events, it can be difficult to identify which signaling routes are active in any particular context. We have previously shown that Inference of Signaling Activity (ISA) can be a useful method to study steady-state oncogenic signaling across different cell lines and inhibitor treatments. However, ISA did not explicitly include feedback signaling events. Incorporating feedback will increase the complexity and computational cost of the model, and more data is likely to be needed to infer feedback activities. Here, we developed feedback-ISA (f-ISA), an extension of the ISA modeling approach which incorporates feedback signaling events. It also includes integrated batch correction in order to fit the models to multiple, independent datasets simultaneously. We find that the identifiability of feedback activities can be counter-intuitive, which shows the importance of analyzing the full, joint uncertainty in model parameters. By iteratively adapting the model and including multiple datasets, including both steady state and intervention data, we constructed a model that can explain a large part of the phosphorylation levels of several signaling molecules in the MAPK and AKT pathways, across many breast cancer cell lines and across various conditions. The resulting model delineates which routes in the signaling network are likely to be active in each cell line and condition, given all of the data. Additionally, such models can indicate whether datasets agree with each other, and identify which parts of the data cannot be explained, thereby highlighting gaps in the current knowledge. We conclude that this modeling approach can be useful to quantitatively understand how complex cellular signaling networks behave across different cell lines and conditions.

5.1. INTRODUCTION

The mitogen-activated protein kinase (MAPK) and AKT signaling pathways are two central regulatory mechanisms which are often deregulated in cancer cells. Many inhibitors which block key kinases in these signaling pathways have been developed as potential cancer therapies. These kinase inhibitors can be very potent anticancer drugs, but cancer cells are often intrinsically resistant or develop resistance over time [1]. As a result, patients have a highly variable response to kinase inhibitors, and this variability is also seen in cell lines [2–4]. To understand this variability in response, and develop effective and selective (combination) therapies, a detailed quantitative understanding of cellular regulatory networks would be highly useful.

Signaling networks are complex, with many interrelated signaling events. One of the important features of cellular signaling networks is the existence of feedback mechanisms. These feedback loops can, for example, provide robustness to a signaling network or modulate the sensitivity to external inputs [5]. To date, many details of numerous signaling networks have been discovered, including various feedback events. In the MAPK pathway, an important feedback mechanism is the inactivation of RAF by ERK [6, 7]. In the AKT pathway, two feedback mechanisms involve the down-regulation of the insulin receptor substrate IRS1 [8], and the modulation of SIN1 activity [9, 10], a component of the MTORC2 complex. Nevertheless, several aspects of these feedback loops remain

unclear.

For instance, in the case of SIN1, there has been debate about the regulation and function of its phosphorylation sites, in particular T86 and T398. Liu et al [9] have shown that both S6K and AKT can phosphorylate SIN1, and that this phosphorylation suppresses MTORC2-mediated phosphorylation of AKT. Conversely, Humphrey et al [11] have shown that AKT phosphorylates SIN1 and that this stimulates MTORC2 instead. Yang et al [10] further investigated this in more cell lines and conditions, arguing that AKT rather than S6K is the major SIN1 kinase and that the feedback is positive, in line with Humphrey et al.

For IRS1, one aspect which is unclear is the phosphorylation of S312 (in human IRS1). Note that this site corresponds to mouse S307, while the human S307 constitutes yet another phosphorylation site on IRS1. Signaling by IRS1 is regulated by multiple phosphorylation events [12], where serine/threonine phosphorylation inhibits its activity while tyrosine phosphorylation activates it. For S312, it is known that JNK (among other kinases) phosphorylates this site [13], but it has also been shown that this phosphorylation is MTORC1-dependent [14]. S6K phosphorylates IRS1 on several other sites, including S307 [15] and S270 [16]. Given the MTORC1-dependence of IRS1 S312 phosphorylation, it is sometimes assumed that S312 phosphorylation is also S6K-dependent. Indeed the databases Uniprot [17] and PhosphoSitePlus [18] currently list S6K as kinase for IRS1 S312 phosphorylation (putatively in the case of PhosphoSitePlus), even though to the best of our knowledge there is no direct evidence for this. Rather, it has been reported as unlikely [15], and S312 is also not part of an S6K target motif. MTORC1 may also affect S312 phosphorylation through an effect on protein phosphatase PP2A [19] or indirectly through dependencies between phosphorylation sites. In addition to these details, it is unclear in which contexts feedback through IRS1 — whether mediated by MTORC1, JNK, S6K or yet other regulators — is important. For example, feedback to IRS1 is involved in insulin resistance [8], and mediates re-activation of the AKT pathway after rapamycin treatment [20], but seems not to be involved in re-activation of AKT after AKT-inhibitor treatment [21, 22].

Computational models can be used to better understand these complex regulatory networks. Different modeling frameworks have been used to quantitatively study oncogenic cellular signaling pathways, including dynamic models [23, 24] and steady state models [25–27]. Dynamic models can describe detailed kinetics of a system, but they are costly to simulate, especially when the full parameter uncertainty is analyzed [28–30]. The computational cost is further exacerbated when multiple datasets are included, resulting in many model conditions which have to be evaluated for each parameter value. Logic models allow significantly larger models to be evaluated [31, 32], but it is more difficult to model quantitative differences between cell lines and conditions in such a framework.

We have previously developed Inference of Signaling Activity (ISA), a steady state modeling approach to study signaling activities across cell lines along with different inhibitor treatments [27]. One major assumption in ISA is the absence of feedback signaling, which allowed fast model evaluations and hence relatively large signaling models. Models without feedback can give good fits to drug response cell viability data [27], arguing that feedback events are not crucial to describe differences in the relative viability

between cell lines after 72 hour drug treatment. However, it is clear that feedback signaling events are a major component of cellular signaling, and are important to consider in many situations. We therefore set out to expand ISA with feedback signaling events, to explore whether the method could also be used to infer signaling activities in models that include feedback events.

5.2. METHODS

Below, we outline the substantial changes we have made to ISA [27], resulting in feedback-ISA (f-ISA), an approach capable of modelling feedback mechanisms in signaling pathways across different cell lines and conditions.

5.2.1. MODEL STRUCTURE

In ISA, the activities of signaling molecules are modeled by a continuous latent variable xi that can take values between 0 and 1. This signaling activity is denoted by x_i where i indexes the signaling molecule. This activity is a function of the upstream signaling nodes, as well as a basal activity, the expression of the signaling molecule itself and any kinase inhibitors that may be present. In the original ISA, the activity was restricted between 0 and 1 using a clamping function. In order to accommodate feedback loops, we now change the activation function to a logistic function. Specifically, we calculate the signaling activity as

$$x_i = u_b e_i \frac{1}{1+\exp(-k(b_i + (\sum_{j \in \text{parents}_x(i)} u_a u_s a_{j,i} x_j) + (\sum_{j \in \text{parents}_m(i)} a_{j,i} m_j) - s)} = h_i(\mathbf{x}). \tag{5.1}$$

Here u_a, u_b and u_s are the kinase inhibitor effects (defined in more detail later), e_i is the expression of signaling molecule i, b_i is the basal activity of the signaling molecule i, $a_{j,i}$ is the strength of signaling from molecule j to molecule i, m_j is a binary variable denoting whether mutation j is present, while k and s are the constant steepness and inflection point of the logistic function. The logistic function is more expensive to compute than a clamping function, but in contrast to a clamping function, the logistic function is smooth (i.e. continuously differentiable), which simplifies the process of solving the systems of equations with feedbacks. To keep the number of parameters manageable, we fix both the steepness k and the inflection point s to a set value. The steepness is set to 9.19024 such that the activity is 0.01 and 0.99 when the total input is 0 and 1 respectively, and the inflection point is set to 0.5. An example of an activity calculation is shown in Figure 5.1.

Without feedback events, the signaling activities can be calculated from the upstream molecules downwards. However, by including feedback events, the equations for the signaling activities become coupled, and as a result the activities have to be calculated by solving a system of nonlinear equations. We use the Newton-Raphson method for this; that is, we solve the equation

$$\mathbf{x} = \mathbf{h}(\mathbf{x}) \tag{5.2}$$

by iterating through

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \mathbf{J}^{-1}(\mathbf{x}^n)\mathbf{f}(\mathbf{x}^n) \tag{5.3}$$

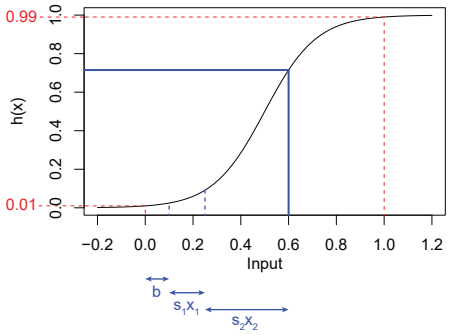


Figure 5.1: **Activation function used to calculate the activity of a signaling molecule from its input.** The red lines indicate the constraints which were taken to determine the fixed steepness and inflection point of the logistic function. The blue lines give an example; in this case the signaling molecule has two upstream inputs (x_1 and x_2), which are multiplied by the respective signal strengths (s_1 and s_2) and summed together with a basal activity (b) to give a final signaling activity of approximately 0.7.

where $\mathbf{x}^n$ represents the solution at the n-th iteration, and

$$\mathbf{f}(\mathbf{x}) = \mathbf{h}(\mathbf{x}) - \mathbf{x}. \tag{5.4}$$

The Jacobian matrix J is given by

$$J_{i,j} = \frac{\partial f_i}{\partial x_j} = \begin{cases} -1 & \text{if } i = j \\ 0 & \text{if } j \notin \text{parents}(i) \\ u_b u_a u_s e_i a_{j,i} \dfrac{k p_i}{(p_i+1)^2} & \text{if } j \in \text{parents}(i), \end{cases} \tag{5.5}$$

with

$$p_i = \exp(k(b_i + (\sum_{j \in \text{parents}_x(i)} u_a u_s a_{j,i} x_j) + (\sum_{j \in \text{parents}_m(i)} a_{j,i} m_j) - s)). \tag{5.6}$$

If the system of equations contains four or fewer signaling molecules, Equation 5.3 is most efficiently calculated by taking the inverse of the Jacobian directly. If the system contains more than four signaling molecules, it is more efficient to use LU-decomposition to solve Equation 5.3 instead. We stop the Newton-Raphson iteration when the last change in $\mathbf{x}$ is less than 10^{-5} in each direction.

Although it is possible to calculate the entire model with Equations 5.2 and 5.3, the LU-decomposition scales cubically in the number of signaling molecules. It would therefore be beneficial to decrease the system size as much as possible. The system of equations is not fully coupled however. Some signaling molecules are not involved in any feedback loop, but are only upstream or downstream of other molecules. The activities of these molecules can be calculated directly, without having to solve a system of equations. Additionally, some signaling molecules which are affected by a feedback loop, may not be affected by another loop. We can therefore decrease the size of the systems to be solved by decomposing the model into smaller systems composed of signaling molecules which are coupled by a feedback signaling loop. To identify which parts need

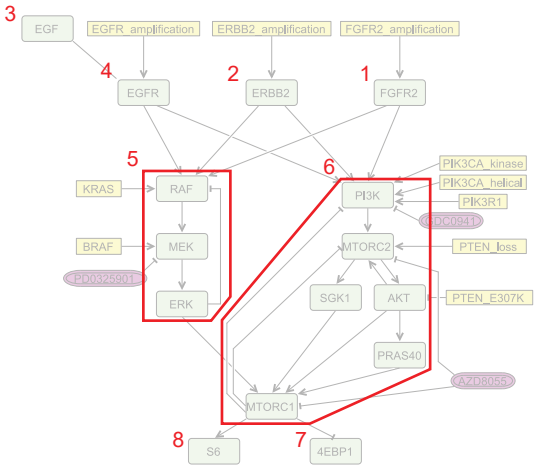


Figure 5.2: **Decomposition of the model into smaller modules.** The model is decomposed into modules that have to be solved as systems of equations (the red boxes; one system of 3 nodes and one system of 6 nodes). A topological ordering is also generated (the red numbers) giving the order in which to calculate the isolated nodes and systems.

5

to be solved as a system of equations, we use Tarjan's strong connectivity algorithm [33] as a preprocessing step. This algorithm identifies all the strongly connected components of a directed graph, as well as providing a topological ordering of the graph. To calculate the signaling activities, we then iterate over the topological ordering, and calculate the activities either directly (for connected components of size 1, i.e. individual signaling molecules that are not part of any feedback system) or by solving the corresponding system of equations (for connected components with size larger than 1). Figure 5.2 illustrates the decomposition of a model into individual molecules and systems.

Note that the signaling activity vector $\mathbf{x}$ is a function of the mutations, $\mathbf{m}$, protein expression, $\mathbf{e}$, and the presence of any inhibitors, u, (all of which can change between cell lines and conditions), as well as a number of parameters (which remain constant between cell lines and conditions). As such, the model is constrained to explain all observed signaling activities from the mutation and protein expression patterns. This is a fundamentally different approach from estimating signal strengths separately for each cell line and condition, as done for example by [24]. Rather than inferring cell-line specific signaling strengths, we attempt to infer signal strengths that can fit the data across all cell lines and conditions.

Kinase inhibitors in the model can be of two types: they can either inhibit the activity of their target, or they can inhibit the activation of the target. The type of the inhibitor is derived from the literature. For example, the AKT inhibitor GSK690693 is an ATP-competitive inhibitor which inhibits the activity of AKT, while the AKT inhibitor MK2206 is an allosteric inhibitor which inhibits the activation of AKT. For activity-inhibiting drugs we set $u_a = u$ in Equation 5.1 (for the specific signaling link that is inhibited), and for activation-inhibiting drugs we set $u_b = u$ (for the specific target that is inhibited). Note that u_a affects the activity of the children of the target, whereas u_b affects the activity of

the target itself. The type of the inhibitor is specified in the SBML annotation (described later). Regardless of whether u_a or u_b is used, the inhibitory effect is calculated as

$$u = \begin{cases} q & \text{if a single concentration is used} \\ q + \dfrac{1-q}{10^{k_u(c-c_{IC50})+1}} & \text{if more than one concentration is used} \\ 1 & \text{if the inhibitor is not present,} \end{cases} \tag{5.7}$$

where q is the maximal inhibitory effect (or the inhibitory effect at the particular concentration used if there is only one concentration), k_u is the steepness of the inhibitory curve, c is the concentration and c_{IC50} is the 50% inhibitory concentration. In contrast to the activation function in Equation 5.1, the steepness and inflection point of the kinase inhibition curves (i.e., k_u and c_{IC50}) are included as free parameters to be inferred. Finally an inhibitor may increase the susceptibility of its target to incoming signals, a process which has been named "inhibitor hijacking" [21]. The variable u_s is included in Equation 5.1 to reflect this effect. This parameter is only used if the drug is specified to alter the susceptibility of its target, and we use it specifically to allow ATP-competitive AKT inhibitors to alter the susceptibility of AKT phosphorylation by MTORC2.

5.2.2. LIKELIHOOD FUNCTION

Jastrzebski et al [27] used protein phosphorylation data as well as untreated proliferation rates and relative viability upon kinase inhibitor treatment to infer the signaling activities. However, since f-ISA is computationally more expensive than the original ISA framework, we only use protein phosphorylation data for the inference here. Including cell viability data, in particular dose response curves, would result in too many model conditions as well as additional parameters to be estimated. Mutation data, copy number data and total protein expression data are still used as before [27], by directly setting the corresponding variable (m_i or e_i) to the observed value. The structure of a small part of a model is shown in template notation in Figure 5.3.

The likelihood of an observed protein phosphorylation measurement is defined as

$$P(y_{i,j,k,l}|\boldsymbol{\theta}) = t(y_{i,j,k,l}|\mu = z_{i,j,k}(\boldsymbol{\theta}), \sigma = \sigma_{i,k}, \nu = 3), \tag{5.8}$$

where $\boldsymbol{\theta}$ is the vector containing all model parameters, $y_{i,j,k,l}$ is the measurement data for observed variable i, cell line j, dataset k and replicate l, $z_{i,j,k}$ is the modeled variable (defined further below in Equation 5.9) and $\sigma_{i,k}$ is the variance of observed variable y_i in dataset k. The variance $\sigma_{i,k}$ for observed variable i is shared by all cell lines and biological replicates, but is specific for each dataset k.

The modeled variable z is defined as

$$z_{i,j,k}(\boldsymbol{\theta}) = g_{i,k} + d_{i,k} x_{i,j}(\boldsymbol{\theta}) \tag{5.9}$$

where $g_{i,k}$ is the background signal generated by aspecific binding of the antibody and $d_{i,k}$ is a scaling factor to account for differences between datasets (recall that k indexes the dataset). If a particular phosphorylation is only measured by one dataset, then $d_{i,k}$ is set to 1, and if the phosphorylation is measured by more than one dataset, then $d_{i,k}$

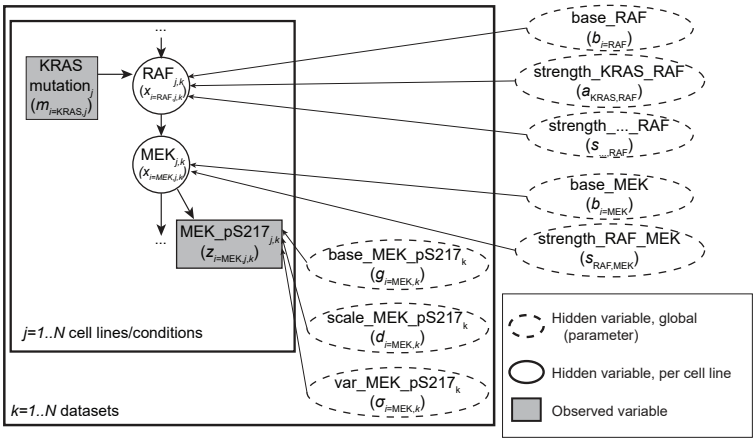


Figure 5.3: **Part of a signaling model shown in template notation.** This graph illustrates the latent variable structure and the integrated batch correction. The text names indicate the names in the code while the symbols in brackets correspond to the equations in the Methods section. Each dataset (indexed by k) has multiple cell lines or conditions (indexed by j). The signaling activities are unique for each condition. The likelihood of the data depends on two batch correction variables (base and scale) as well as the variance; these variables are specific for each dataset. The signaling strengths and base activities are global parameters, which are shared by all datasets and all conditions.

is included as a free parameter to be inferred, for every dataset. $g_{i,k}$ is always included as free parameter to be inferred for every epitope and dataset, regardless of whether the epitope is uniquely measured by only one dataset. Finally, we assume independence between the epitopes, datasets, cell lines and replicates, and the total likelihood is calculated as a product of the likelihoods of the individual data points.

Although the likelihood function includes a scaling variable to perform batch correction, it is useful to have the measurements on approximately the same scale between datasets before running the inference. Therefore, as a preprocessing step, the measured phosphorylation levels are reduced to [0,1] by dividing the measurements of each epitope by the maximum value observed for that epitope in any cell line or condition in that dataset, if that value is larger than 1. With the RPPA quantification procedure that was used (SuperCurve [34]), maximum observed values below 1 indicate that for that epitope, all conditions in that dataset have low phosphorylation levels, relative to the set of standard lysates. We therefore do not increase the signal to 1, to prevent amplifying experimental noise. With all measurement values in the range [0,1], the prior for the scaling parameters $d_{i,k}$ can then be set to a small uniform distribution, between 0 and 2, which reduces the parameter space that has to be searched.

5.2.3. PARAMETER PRIORS

The model contains several different types of parameters that are inferred from the data. Table 1 lists these parameters, along with the prior distribution that is used for each of them.

The strength parameter for reactions with a known influence sign is inferred on a

logarithmic scale. This is done to give equal prior weight to weak interactions (strength 0.01-0.1), interactions with intermediate strength (0.1-1) and strong amplifying interactions (1-10). If this range from 0-10 was not log-transformed and a uniform distribution would still be used, most of the prior weight would be on amplifying signals, and all signaling activities would be heavily saturated a priori, especially in a signaling cascade. When the sign of the influence is not known, that is, when the signal could be either activating or inhibiting, then the prior is set on a regular scale from -2 to 2.

For drug inhibition, the drug can be specified to either use a single drug concentration, giving only a single parameter for the inhibition at that concentration, or to use multiple drug concentrations, giving three parameters: the maximum inhibition, the 50% inhibitory concentration (IC50) and the steepness. The prior for the 50% inhibitory concentration is centered on the concentration at which it inhibits its target for 50%, as determined by in vitro inhibition experiments described in the literature.

5.2.4. MODEL STRUCTURE SPECIFICATION

The signaling graph of the model is constructed in CellDesigner [35], version 4.4. Using this tool, the model is specified in Systems Biology Graphical Notation as an Activity Flow diagram [36]. That is, rather than specifying precise molecule states and reactions, only the abstract influences between signaling molecules are specified, which fits with the modeling paradigm of ISA. The model is stored as an SBML file with CellDesigner extension annotation. This extension annotation allows specification of the types of the signaling molecules (i.e. whether a node is a signaling molecule, mutation or drug) as well as specification of the signaling influences and the types of drug inhibition.

5.2.5. INFERENCE

To infer the posterior distribution of the parameters, we use the BCM software package [37]. To incorporate f-ISA in BCM, we developed the tool sbmlpdinf, which can read the activity flow diagram described in the SBML/CellDesigner file, as well as two XML files specifying the likelihood and the prior. All data is stored in a single NetCDF4 file. The likelihood file specifies which measurement in the data file corresponds to which node in the model, as well as all the model conditions, such as the presence of a particular kinase inhibitor. The model simulation code was written in C++ and made thread-safe for use in the parallelized inference algorithms of BCM. The largest model presented here, with

Parameter	**Symbol**	**Prior**
base_[*molecule*]	b_i	uniform(a=0, b=1)
strength_[*molecule*]_[*molecule*]	$s_{j,i}$	uniform(a=-2,b=1) — $\log_{10}$ scale
maxinhib_[*drug*]_[*molecule*]	q	uniform(a=0, b=1)
[*drug*]_[*molecule*]_susceptibility	u_s	uniform(a=-2,b=1) — $\log_{10}$ scale
ic50_[*drug*]_[*molecule*]	c_{IC50}	normal($\mu = x$, σ=2) — $\log_{10}$ scale
logsteepness_[*drug*]_[*molecule*]	k_u	uniform(a=-1,b=1) — $\log_{10}$ scale
base_measurement_[*dataset*]_[*epitope*]	$g_{i,k}$	uniform(a=0, b=1)
scale_measurement_[*dataset*]_[*epitope*]	$d_{i,k}$	uniform(a=0, b=2)
variance_measurement_[*dataset*]_[*epitope*]	$\sigma_{i,k}$	exponential(λ=5)

Table 5.1: **Model parameters that are inferred from the data.**

18 signaling molecules and two feedback systems, could be evaluated at approximately 1.1 million model evaluations per second using 18 threads on an Intel Xeon E5-2697 v4 processor.

We sampled the posterior distribution using feedback-optimized, parallel tempered MCMC [38] with automated parameter blocking [39]. The marginal likelihood was calculated using thermodynamic integration [40]. The temperature schedule was optimized twice, which also served as burn-in period. After this optimization, 1,000 samples were generated from the posterior (after subsampling), with the amount of subsampling chosen such that the autocorrelation was negligible and at least 100 roundtrips from prior to posterior were performed. Specifically, when one dataset is included in the inference we use 24 parallel chains and run the inference with a subsampling of 1 in 2,000 and probability of choosing a temperature swap move of 0.9. Each MCMC move constitutes updating every parameter or parameter block once. When more than one dataset is included we increased the number of chains to 36 and use a subsampling of 1 in 4,000. Sample traces are provided in Supplementary File 1. Despite the high performance implementation, inference with a single dataset and a medium-sized model takes several hours, and the largest inference presented here, which included 108 model conditions and 139 free parameters, required approximately 48 hours to run with the aforementioned processor.

5.2.6. QUANTIFYING REDUCTION IN UNCERTAINTY

To quantify the reduction in uncertainty, we used the Occam factor introduced by MacKay [41], which quantifies the change in volume of the parameter space that is accessible from the posterior compared to the prior. In other words, it measures which fraction of the prior parameter space is consistent with the data. MacKay used a Gaussian approximation to calculate the Occam factor; but since we evaluated the marginal likelihood as part of the inference here, we can calculate the Occam factor directly by

$$\log \text{Occam factor} = \log P(y|\mathcal{M}) - \log P(y|\mathcal{M}, \boldsymbol{\theta}_{\text{MAP}}) \tag{5.10}$$

where $P(y|\mathcal{M})$ is the marginal likelihood of the data given the model and $P(y|\mathcal{M}, \boldsymbol{\theta}_{\text{MAP}})$ is the likelihood of the data at the maximum a posteriori value of the parameters.

5.2.7. CELL LINES

All cell lines used have been described, including their culture conditions, in Supplementary Table 1 of Jastrzebski et al [27].

5.2.8. MEASUREMENT OF ON-TREATMENT PHOSPHORYLATION

Cell lines were seeded in 60 mm dishes (BT549 at $4\text{x}10^5$ cells/dish; HCC1954 at $8\text{x}10^5$ cells/dish; MCF7 at $4\text{x}10^5$ cells/dish; MM231 at $8\text{x}10^5$ cells/dish; MM453 at $1\text{x}10^6$ cells/dish; MM468 at $1\text{x}10^6$ cells/dish; SKBR3 at $6\text{x}10^5$ cells/dish; T47D at $8\text{x}10^5$ cells/dish). Following 24 h of incubation, all but the exponentially growing cells were serum starved in unsupplemented base medium containing penicillin/streptomycin (Gibco) for a further 24 h. Cells were then treated with either DMSO vehicle control or one of the following three inhibitors — PD0325901 at 50 nM, GDC0941 at 10 μM or AZD8055 at 1 μM — for 30 min, after which stimulation with 10 ng/ml EGF, where indicated, was carried out for

a further 20 min. Cells were then placed on ice, washed with ice-cold PBS, and lysed in 150 µl/dish of RIPA buffer (20 mM Tris-HCl, pH 8, 150 mM NaCl, 1% NP40, 0.5% sodium deoxycholate, 0.1% SDS) supplemented with cOmplete protease and phosSTOP phosphatase inhibitor cocktails (Roche). Lysates were cleared by centrifugation at 4°C and 20,800 x g, and protein concentration determined using the Pierce BCA protein assay (Thermo Fisher Scientific). The supernatant was normalized to 1 µg/µl with RIPA buffer and supplemented with SDS sample buffer to a final concentration of 62.5 mM Tris-HCl pH 6.8, 10% glycerol, 2% SDS, 2.5% (v/v) 2-mercaptoethanol. Samples were further assayed at the MD Anderson Cancer Center RPPA Core Facility, as outlined previously [27].

5.2.9. IRS1 DISRUPTION EXPERIMENT

Cell lines were seeded in 6-well plates (BT549 at $4x10^5$ cells/well; HCC1954 at $2x10^5$ cells/well), and following a 48 h incubation, treated with either DMSO vehicle control or 5 µM NT157 for a further 24 h. They were then treated for a further 1 h with either DMSO vehicle control, 1 µM GSK-690693 or 5 µM PF-4708671. Cells were then lysed and protein concentration determined as outlined above. Twenty µg of protein were supplemented with Novex® LDS Sample Buffer and Sample Reducing Agent, heated at 70°C for 10 min and separated on 4-12% gradient gels (Thermo Fisher Scientific). Separated proteins were transferred onto Immobilon-P PVDF membranes (Merck Millipore) using a Trans-Blot® system (Bio-Rad). Blocking was performed in TBS supplemented with 0.1% Tween and 3% BSA (TBS-TB) for 1 h at room temperature, followed by overnight immunoblotting at 4°C with the following primary antibodies: IRS1 (MERCK/Millipore 06-248); pIRS1-S312 (Cell Signaling Technology 2381); AKT (Cell Signaling Technology 9272); pAKT-S473 (Cell Signaling Technology 4060); SIN1 (Cell Signaling Technology 12860); pSIN1-T86 (Cell Signaling Technology 14716); PRAS40 (Cell Signaling Technology 2610); pPRAS40-T246 (Cell Signaling Technology 2997); S6 (Cell Signaling Technology 2217); pS6 (S235/236) (Cell Signaling Technology 2211); Vinculin (Sigma V9131). Membranes were then washed with TBS supplemented with 0.1% Tween (TBS-T) and probed with secondary goat anti-mouse or anti-rabbit HRP-conjugated antibodies (Bio-Rad) diluted in TBS-TB for 2 h at room temperature. Finally, membranes were washed in TBS-T, an ECL reaction was carried out using the ClarityTM Western ECL Substrate (Bio-Rad) and the signal detected using a ChemiDoc Touch instrument (Bio-Rad). Quantification of the exposures was performed using ImageQuant software (Bio-Rad).

5.2.10. EXTERNAL DATA

In addition to the data we generated here, we included a part of the dataset provided by Korkola et al [42] in the inference. They performed RPPA measurements of 15 breast cancer cell lines, treated with either lapatinib (250 nM), GSK690693 (250 nM), a combination of the two, or a DMSO control, at 8 time points from 30 minutes to 72 hours. Since we focus here on fast-acting post-translational feedback here, we selected the 1-hour time point from their data, which most closely matched the 50-minute treatment of our on-treatment phosphorylation measurements. We used only the 9 cell lines which overlap with our cell line panel, since we have mutation and copy number data available for these cell lines. To have the measurements on the same scale, we reverse-log-transformed the data and divided by the maximum value for each epitope as described

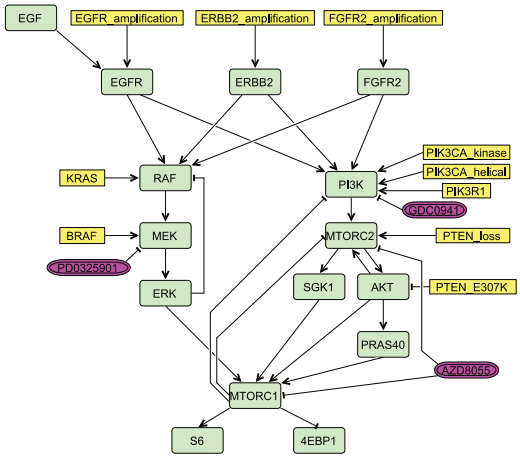


Figure 5.4: **Initial signaling model.** The graph shows the starting signaling model of MAPK and AKT signaling in breast cancer, based on the reduced, joint-drug model of Jastrzebski et al [27]. The model is depicted in Activity Flow format of the Systems Biology Graphical Notation. Yellow boxes indicate genetic events including mutations and copy number aberrations, green nodes are the signaling molecules, and purple nodes are the kinase inhibitors.

5

above.

5.3. RESULTS

5.3.1. CONSTRUCTING A MODEL OF STEADY STATE SIGNALING WITH FEEDBACK LOOPS

To better understand the signaling activities in the MAPK and AKT regulatory networks, we followed the ISA modeling approach [27] to construct a simplified model of these pathways. We based our starting model (see Figure 5.4) on the simplified, joint drug-model of [27]. The starting model includes three receptor tyrosine kinases (RTKs), a simplified representation of the MAPK and AKT pathways, and 10 important oncogenic mutations which are observed in breast cancer cells. In addition, we now added four known feedback signaling events.

Briefly, in ISA, the steady state activity of a signaling molecule is modeled as a latent variable with a continuous value between 0 and 1. The signaling activity is a function of a basal activity, the expression of the signaling molecule itself, the activities of upstream signaling molecules, the effect of mutations, and any kinase inhibitors that are present. Importantly, the parameters of the model are shared across all cell lines and conditions. That is, the parameter values are the same for all cell lines in all conditions. For example, while each cell line can have a different amount of AKT activity in a particular condition, a given amount of AKT signal always gives rise to the same amount of input signal to MTORC1.

To be able to include feedback events in ISA, we make two main changes to the modeling approach. First, to constrain the activity of the signaling molecules between 0 and

1, we use a logistic function rather than a clamping function (see Figure 5.1). Although this is computationally more expensive, it makes the activation function smooth, which greatly simplifies solving the systems of equations when feedbacks are present. Second, to calculate the activity of the molecules which are part of a feedback system, we use Newton-Raphson root-finding rather than calculating the activities in one pass from upstream to downstream, as was done before. Since the computation of one Newton-Raphson step scales cubically with the number of signaling molecules, a significant speed-up can be obtained if we restrict the root-finding to the part of the model that contains feedback. We therefore first identify the strongly connected components in the model, using Tarjan's strong connectivity algorithm [33], illustrated in Figure 5.2. This algorithm also provides a topological ordering, which is needed to calculate the signaling activities in the decomposed equations in the correct order. The system of equations corresponding to each strongly connected component is then solved separately, with the benefit that these systems are typically much smaller than the complete model. All equations and the methodology for solving them are described in detail in the Methods section.

5.3.2. FEEDBACK ACTIVITY FROM ERK TO RAF IS PARTIALLY IDENTIFIABLE FROM PRE-TREATMENT, NON-INTERVENTION DATA

Having a methodology to infer steady state signaling activities in models with feedback, we first wondered whether feedback loops are already identifiable using non-intervention data. To test this, we fitted the starting model (Figure 5.4) to the phosphorylation data of thirty, untreated breast cancer cell lines, grown under normal culturing conditions; i.e. the phosphorylation data of Jastrzebski et al [27]. We first included only the phosphorylation data of the downstream signaling kinases (Figure 5.5A). To our surprise, these non-intervention data already suggest that the feedback loop from ERK to RAF is likely to be active (Figure 5.5B). A possible explanation for the identifiability of feedback from non-intervention data may be that there are many inputs into the MAPK pathway; in the model we included three different RTKs, as well as mutations in KRAS or BRAF. These five inputs together could lead to over-activation of the MAPK pathway, so to prevent this, the inputs need to be restrained in some way. It is easier to accomplish this by a negative feedback loop, since this requires only one parameter to be given a high value, rather than by each input to the MAPK pathway being weak, which requires five parameters to have a low value. In such situations, the Bayesian inference follows the principle of Occam's razor by preferring a simple explanation over a more complex one.

To test whether it is indeed the inputs to the MAPK pathway that determine the high values for ERK-to-RAF feedback activity, we artificially forced the inputs to be weak by restricting the prior for the strength parameters of these inputs (Figure 5.5C). In this case, we indeed see that the feedback becomes weaker as well, indicating that the negative feedback is used to balance the inputs.

Nevertheless, the posterior probability distribution for the strength of the ERK-to-RAF feedback loop has non-zero probability for low values as well. This means that a weak or very weak feedback loop is also consistent with the data. A model entirely excluding this negative feedback loop is less likely to represent the data, but the difference

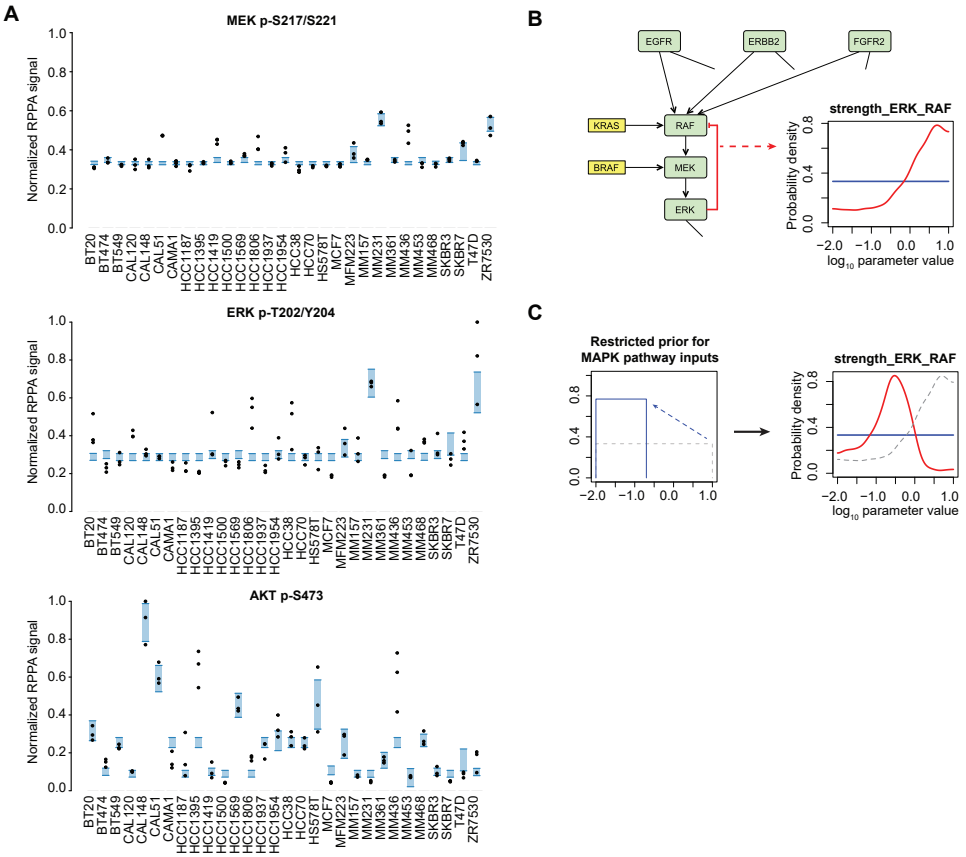


Figure 5.5: **Identification of feedback activity from ERK to RAF from pre-treatment, intracellular phosphorylation measurements.** (A) Data and posterior predictive for three of the six epitopes. Black dots indicate the measurement data and the shaded blue area is the 90% confidence interval of the posterior predictive distribution. (B) Posterior probability density for the strength of ERK->RAF feedback signal inferred from the data. (C) Restricting the prior for the inputs to the MAPK pathway to low values results in a weaker RAF->ERK feedback.

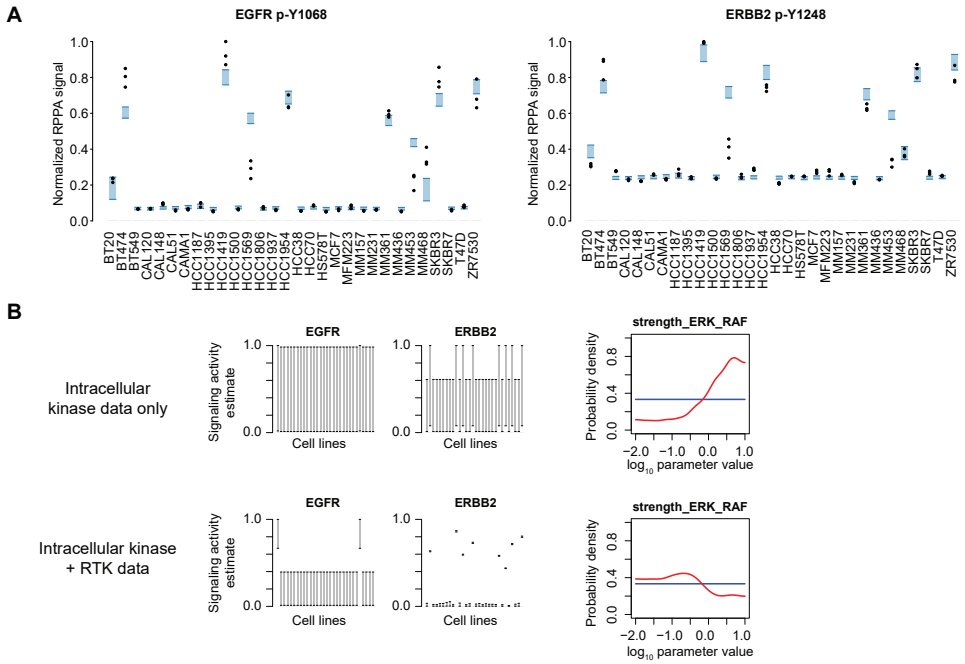


Figure 5.6: **Addition of RTK phosphorylation data increases the uncertainty in ERK to RAF signaling.** (A) Data and posterior predictive of the two epitopes that were added. Note that there is cross-reactivity between the antibodies against EGFR p-Y1068 and ERBB2 p-Y1248, hence both measurements are a linear combination of the EGFR and ERBB2 signaling estimates. (B) Posterior probability densities of the ERK->RAF feedback loop, and the signaling estimates of EGFR and ERBB2, with and without the RTK data. For the signaling estimates, the shaded gray areas indicate the 90% confidence interval.

is small (the log Bayes factor is 1.0 in favor of the model including the feedback loop). In summary, given the pre-treatment, non-intervention data of the intracellular kinases, this feedback loop is likely to be active, but based on the single dataset used so far we cannot rule out that it the feedback from ERK to RAF is inactive.

5.3.3. ADDING DATA CAN INCREASE THE UNCERTAINTY IN INDIVIDUAL PARAMETERS

We next tested whether the addition of more data helps in identifying the feedback loop. We first added data for the surface receptor tyrosine kinases, which provide information on the inputs to the pathways (Figure 5.6A). Interestingly, if we include phosphorylation and protein expression data of EGFR and ERBB2, the feedback activity from ERK to RAF in fact becomes less identifiable (Figure 5.6B). With these data, the model can infer that the activities of EGFR and ERBB2 are low in the majority of cell lines, and this constrains at least some inputs to the MAPK pathway to low strengths. Given this, a strong feedback is no longer required to balance the activation of the MAPK pathway. Since the feedback is not required, but still feasible, the activity of the feedback has become less certain.

To see whether the RTK data does provide some information overall, we tested whether the joint uncertainty in all parameters does decrease. This can be quantified using Occam's factors [41], which measure the reduction in the parameter space that is accessible from the posterior compared to the prior. The model with RTK data has a log Occam factor of -127.6, compared to -97.4 without the RTK data. The posterior space collapses more when the RTK data is included than when it is excluded, meaning that the uncertainty in all parameters together is reduced by the inclusion of the RTK data. Together, this shows that, while the total uncertainty is reduced by adding data, the uncertainty of individual parameters can increase.

5.3.4. INCORPORATING ON-TREATMENT MEASUREMENTS CONFIRMS ERK TO RAF FEEDBACK ACTIVITY

Intervention data should be more informative for identifying feedback loops. We therefore performed on-treatment measurements, using the MEK-inhibitor PD0325901, the PI3K inhibitor GDC0941 and the dual MTORC1/2 inhibitor AZD8055, in a smaller panel of cell lines (Figure 5.7A). Cells were either allowed to grow exponentially for 48 hours, or were starved for 24 hours, followed by 30 minute pre-treatment with inhibitors, before stimulation with EGF to increase signaling activity. The EGF stimulation is included in the model by setting EGF activity to 1 for the EGF-stimulated conditions and to 0 otherwise.

This on-treatment dataset was generated separately, at a different time, from the 30-cell line pre-treatment dataset. Although the RPPA measurements include a set of standard control lysates, the spot intensities are quantified separately and there are likely to be differences between the two data batches. Furthermore, we cannot always rely on a sufficient number of overlapping measurement conditions to align multiple datasets. We therefore incorporated batch correction directly into the inference, such that the differences between batches are automatically accounted for, and balanced against the most likely signaling strengths. This is accomplished by adding an offset and a scale parameter for each epitope in each dataset (see Figure 5.3 and the Methods section for details).

Focusing on the MEK inhibitor, we see that using the on-treatment data, feedback activity from ERK to RAF is clearly identifiable (Figure 5.7B, first column). This shows that on-treatment data is indeed more informative for inferring feedback activity, as expected. The result of the feedback is also clearly visible in the data (Figure 5.7A) given that MEK phosphorylation greatly increases after treatment with a MEK inhibitor. Although the increase in MEK phosphorylation upon MEK inhibitor treatment could also be a direct effect of the inhibitor rather than a feedback event (discussed in more detail for ATP-competitive AKT inhibitors later), PD0325901 is an allosteric inhibitor [43, 44], and the increased phosphorylation of MEK is rather more likely to be through ERK to RAF feedback signaling [45].

When considering several other parameters, we see that both the pre-treatment and on-treatment measurement provide useful information (Figure 5.7B). The pre-treatment measurements were collected for a single condition for more cell lines, while the on-treatment measurements included more conditions for fewer cell lines. Combining the two datasets provides a broad coverage over cell lines as well as intervention measure-

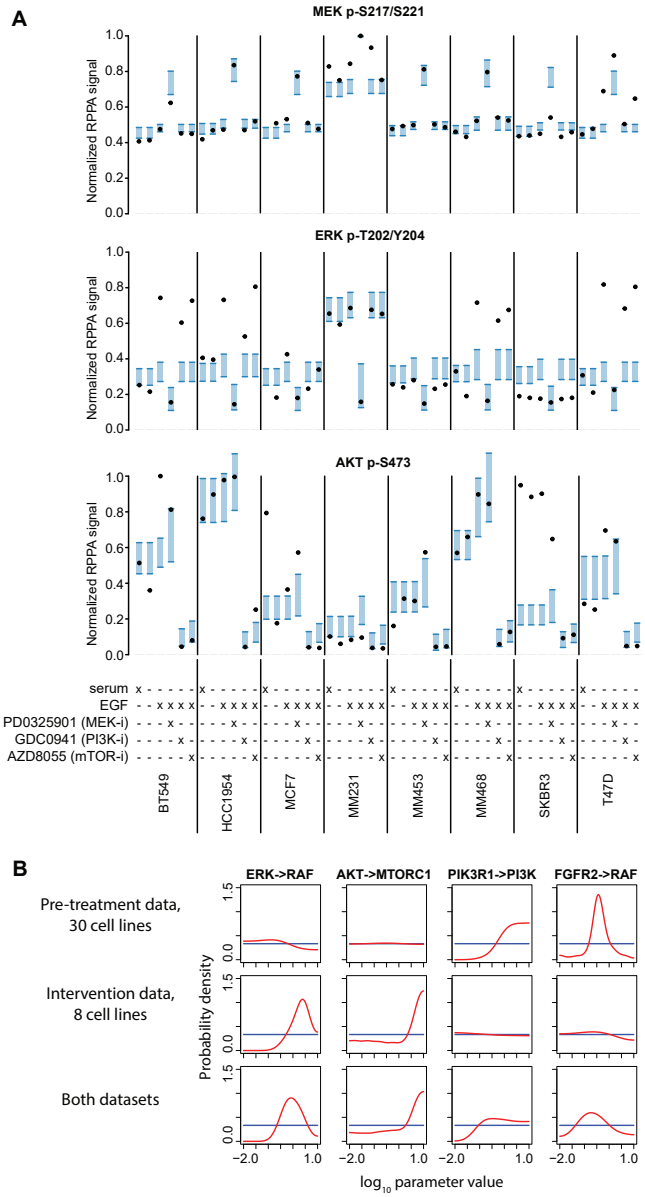


Figure 5.7: **On-treatment data constrains the feedback from ERK to RAF.** (A) Data and posterior predictive for three of the eight epitope added to the inference. The posterior predictive is obtained from inference with the pretreatment and intervention datasets combined. (B) Comparison of densities of the ERK->RAF feedback loop and three other parameters with either dataset alone or together.

ments in a smaller panel to help identify signal strengths. The on-treatment data not only helps to identify feedback activity from ERK to RAF, but also, for example, the strength of AKT to MTORC1 signaling. Conversely, the broader pre-treatment data helps identify, for example, the effect of PIK3R1 mutations and FGFR2 amplifications, neither of which were found in the smaller cell line panel used for generating the on-treatment data.

5.3.5. NEGATIVE FEEDBACK IS LIKELY TO BE ACTIVE IN THE AKT PATHWAY

One of the main advantages of f-ISA lies in identifying which signaling activities and strengths are most strongly supported by the data. This advantage becomes most apparent when there are multiple interrelated feedback events, combined with information from multiple datasets. In this case it is impossible to manually keep track of all the constraints, and computational modeling is necessary to obtain a quantitative understanding of the regulatory network. To illustrate this, we next focus on the AKT pathway, where several different feedback events exist, further complicated by conflicting reports in the literature. As a first approximation, and as can be seen in Figure 5.4, we incorporated the two hypotheses regarding feedback through SIN1, described in the introduction, through a negative link between MTORC1 and MTORC2 (representing inhibitory phosphorylation of SIN1 by S6K), and a positive link between AKT and MTORC2 (representing stimulating phosphorylation of SIN1 by AKT). The feedback through IRS1 is included as a negative link between MTORC1 and PI3K.

Using the two datasets described so far (see Figure 5.7A and 5.8A), the model identifies that there is likely to be a negative feedback in the AKT pathway (Figure 5B). The positive feedback through SIN1 is identified to be weak (Figure 5.8B, top-left panel), while the negative feedback through SIN1 is likely to be active (Figure 5.8B, top-right panel). This is partially driven by the same effect as in the MAPK pathway, that is, there are again many inputs into the AKT pathway, which are most easily balanced by a negative feedback loop. This effect can also be seen by a very strong correlation between the strength of signaling from PI3K to MTORC2 and the negative feedback from MTORC1 to MTORC2 (Figure 5.8B, bottom-right panel). However, apart from this effect, the model also infers that the negative feedback from MTORC1 to MTORC2 is used to reduce AKT-pathway activity in cell lines with high MAPK pathway activity, including MDA-MB-231 and ZR-75-30.

The negative feedback in the AKT pathway could be achieved in several different ways, either through IRS1 or through SIN1. To further resolve which of these proteins mediates the feedback, we extended the model to explicitly include IRS1 as signaling molecule (Figure 5.8C), and we incorporated information from the dataset of Korkola et al [42], who performed time-course RPPA measurements with a different panel of cell lines upon inhibitor treatment, including the AKT inhibitor GSK690693. We selected the 1 hour time point from their dataset as it is closest to the 50 minute inhibitor treatment (30 minute inhibitor pre-treatment plus 20 minute subsequent EGF stimulation) in our intervention dataset. To accommodate this dataset, we also added GSK690693 as inhibitor to the model.

A complication when using ATP-competitive AKT inhibitors like GSK690693 is that these inhibitors can directly alter the susceptibility of AKT phosphorylation by PDK1 and MTORC2, independently of AKT kinase activity [21]. This effect causes an increased AKT

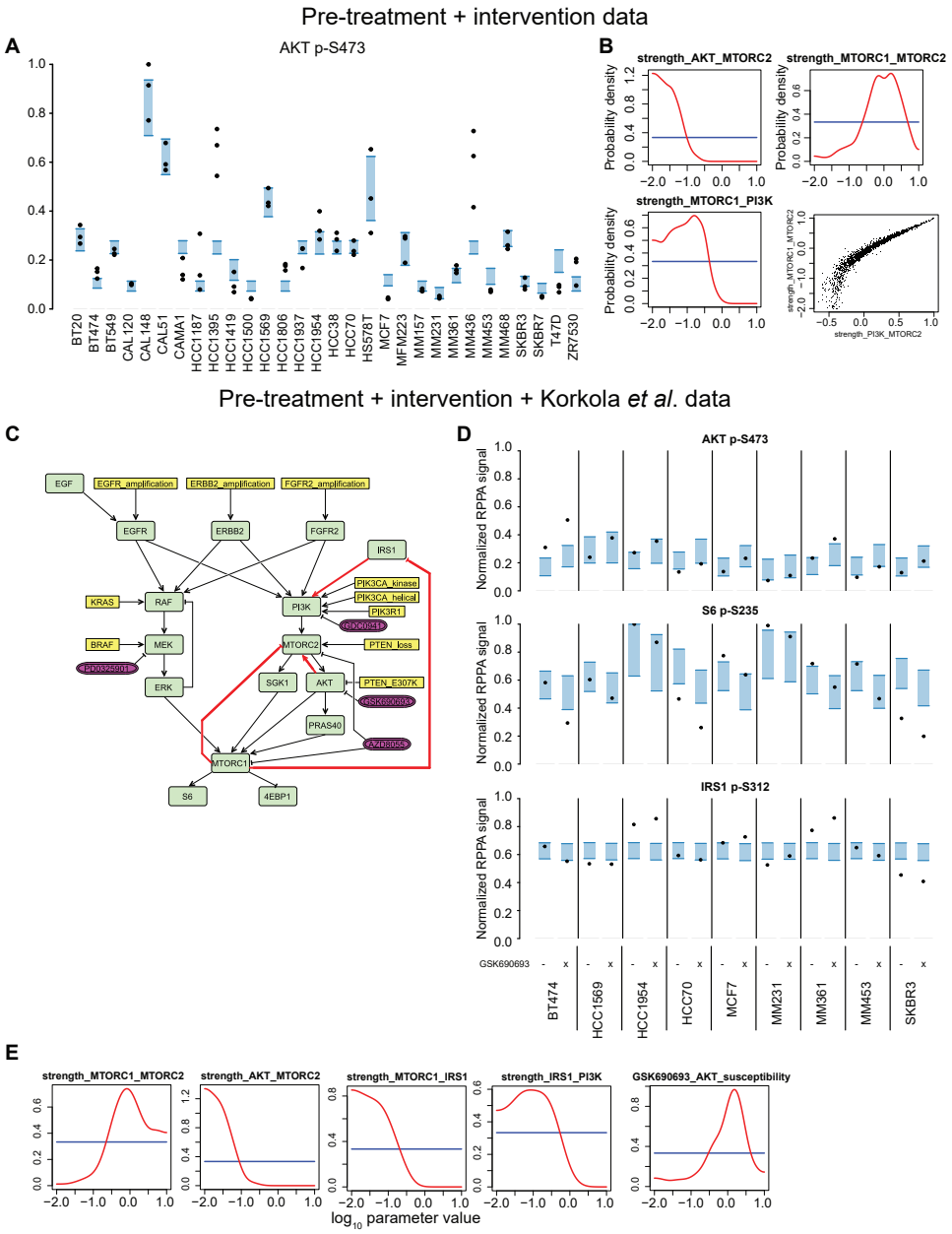


Figure 5.8: **Feedback in the AKT pathway is likely to be negative.** (A) Data and posterior predictive of AKT p-S473 in the pre-treatment dataset in the inference with the pre-treatment and intervention datasets combined; c.f. Figure 5.5A where the same data is shown but the intervention dataset was not included. (B) Feedback signaling strengths in the AKT pathway. The bottom right panel shows a scatter plot of the samples of the PI3K->MTORC2 and MTORC1->MTORC2 signal strengths. (C) Model with GSK-690693 as an additional inhibitor, and IRS1 explicitly included as a signaling molecule. The links for which the posterior is shown in E are highlighted in red. (D) Data and posterior predictive for three of the epitopes used for the inference; the data is the 1-hr time point of the dataset of Korkola et al [42]. The previous two datasets are included in the inference as well. (E) Posterior distribution of the feedback signaling events in the AKT pathway, given the three datasets together.

phosphorylation upon AKT inhibitor treatment, independent of any feedback events. We incorporated this effect in the model by allowing a kinase inhibitor to alter the strength of incoming signals to the targeted protein (see Methods section; note that this effect cannot be seen in the SBGN schematics but is included in the SBML model annotation).

The model can fit several aspects of the Korkola data, including the expected reduction in S6 phosphorylation and increased AKT phosphorylation upon AKT inhibitor treatment (Figure 5.8D), without compromising the fit of the other two datasets (Supplementary File 1). Furthermore, we see that the increased AKT phosphorylation upon AKT inhibitor treatment is indeed most likely explained by an increase in AKT phosphorylation susceptibility (Figure 5.8E, right-most panel). Despite this, a negative feedback in the AKT pathway is still likely to be present (Figure 5.8E, left-most panel). This, however, is unlikely to be through IRS1 S312 phosphorylation (Figure 5.8E, 4th and 5th panel). Consistent with this, the measured IRS1 S312 phosphorylation levels are not consistently changed upon AKT inhibitor treatment in this dataset (Figure 5.8D, bottom panel). Finally, the presence of a positive feedback in the AKT pathway, acting via SIN1, is still deemed unlikely in this model with IRS1 included as signaling molecule (Figure 5.8E, 2nd panel).

5.3.6. TESTING THE STRENGTH OF NEGATIVE FEEDBACK THROUGH IRS1 BY MODULATING ITS EXPRESSION

The above described inference indicated that while there is negative feedback in the AKT pathway, the feedback through IRS1 S312 phosphorylation is likely to be weak. However, there is uncertainty in the strength of IRS1 to PI3K signaling. To further resolve this, we experimentally disrupted the IRS1 feedback mechanism by pre-treating cells with the IGF1R inhibitor NT157, which has been shown to induce degradation of IRS1 [46, 47], and measured how this affects phosphorylation upon AKT- or S6K inhibitor treatment (Figure 5.9A). To accommodate this data more precisely, it is also useful to expand the model by including S6K and SIN1 as separate nodes (the resulting model is shown Figure 5.9B). Since we now explicitly include SIN1, we model the unknown effect of SIN1 on MTORC2 by allowing this signal to be either positive or negative.

With some exceptions, the model can describe the IRS1-disruption experiment very well (Figure 5.9C), without compromising the fit to the other datasets (Supplementary File 1). The data shows a clear increase in AKT phosphorylation after AKT-inhibitor treatment, but not after S6K-inhibitor treatment. We also see that the increased AKT phosphorylation is reduced by pre-treatment with NT157. This is consistent with the report that AKT-inhibitor-induced AKT-phosphorylation is dependent on MTORC2 activity [21], and MTORC2 being dependent on IRS1 signaling. The S6K inhibitor, although it does reduce S6 S235/236 phosphorylation, does not change the phosphorylation levels of AKT S473, SIN1 T86 or IRS1 S312, suggesting it is not critically involved in mediating the feedback loop under these conditions.

We can then use the model to delineate all the feedback activities (Figure 5.9D). In contrast to the inference based on the previous datasets, the model can now infer that the signal from IRS1 to PI3K is strong, perhaps even a point of signal amplification in the pathway, given that the inferred strength is larger than 1 (Figure 5.9D, panel 7). Hence a relatively weak negative feedback to IRS1 can still affect the AKT pathway. Further-

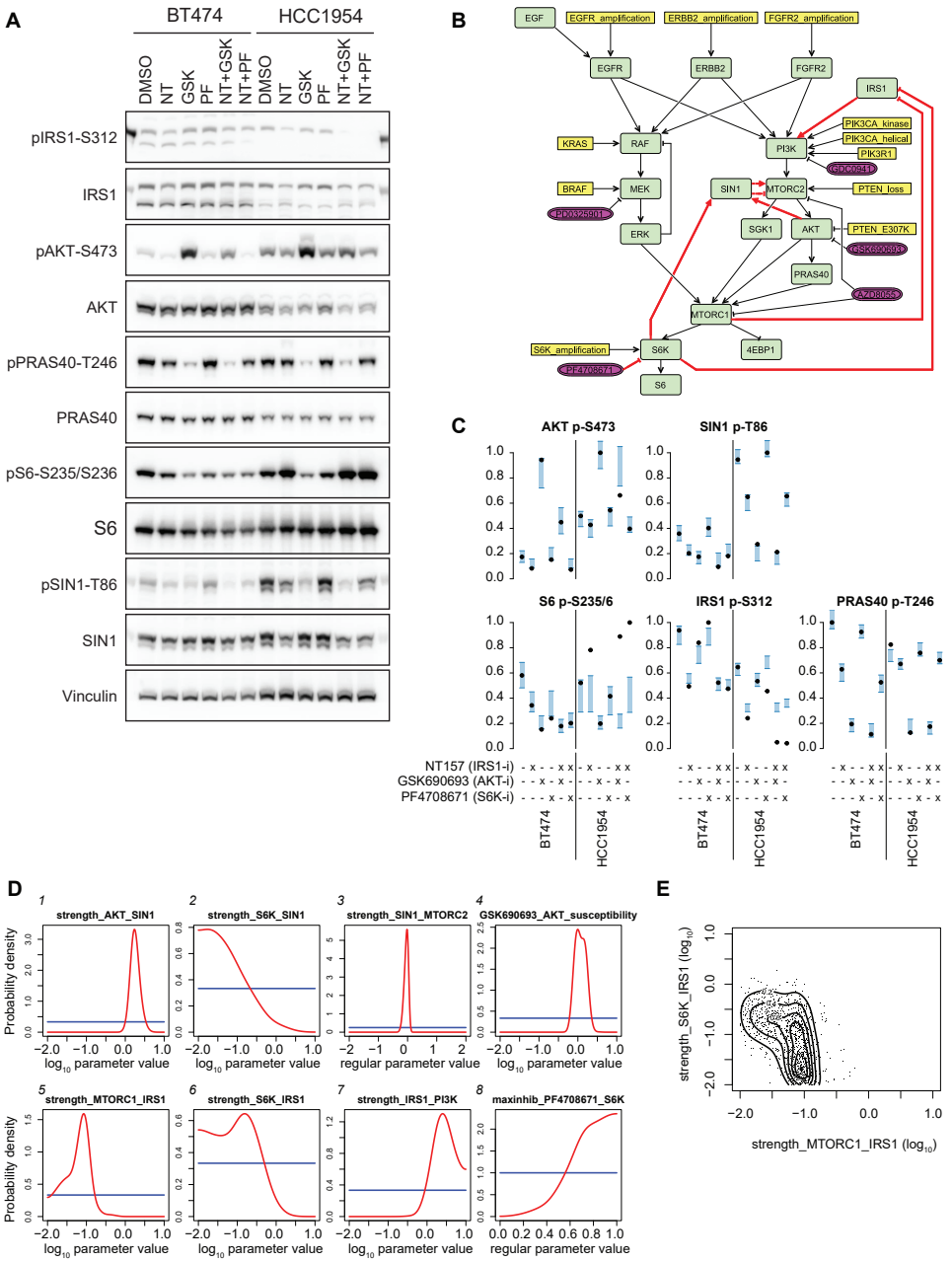


Figure 5.9: **Resolving the feedback through IRS1.** (A) BT474 and HCC1954 cell lines treated with the IRS1/2 inhibitor, NT-157 (NT), alone or in combination with either the AKT inhibitor, GSK690693 (GSK), or the S6K inhibitor, PF4708671 (PF), were analyzed by immunoblotting for expression and phosphorylation of key molecules in the AKT pathway. Control samples were treated with DMSO alone. (B) Model with S6K, SIN1 and PF4708671 added. The double, dashed arrows from SIN1 and MTORC2 indicate that the sign of this link is unknown. The links for which the posterior is shown in D are highlighted in red. (C) Quantification of the western blot signals shown in A, along with the posterior predictive of the fitted model. The previous three datasets are included in the inference as well. (D) Posterior distribution of the feedback signaling strengths in the AKT pathway, given the four datasets together. (E) Scatter plot of the posterior samples for the MTORC1->IRS1 and S6K->IRS1 signal strengths, with a 2D-KDE of the bivariate marginal posterior shown in contours.

more, there appears to be a weak, though nonzero, feedback to IRS1 (Figure 5.9E). The model cannot entirely resolve whether the feedback is mediated by MTORC1 only or also through S6K (Figure 5.9E), due to the uncertainty in how strongly the S6K inhibitor is inhibiting its target (Figure 5.9D, panel 8) - the more likely it is that PF4708671 is inhibiting S6K, the less likely S6K is to phosphorylate IRS1 on S312. Despite this uncertainty, an S6K-independent signal to IRS1-pS312 is approximately three times more likely than an S6K-dependent signal (Figure 5.9D and E).

The hyperphosphorylation of AKT upon GSK690693 treatment is still mainly explained by the inhibitor-induced change in phosphorylation susceptibility (Figure 5.9D, panel 4). It is also clear that SIN1 T86 phosphorylation is dependent on AKT (panel 1), providing further information for the debate on the regulation of this site. Furthermore, SIN1 T86 phosphorylation is unlikely to be mediated by S6K (panel 2), although there is again some amount of uncertainty caused by the poor and uncertain efficacy of the S6K inhibitor (panel 8). However, given the present data, feedback through SIN1 is predicted to have only a minor negative effect on MTORC2 activity, if any (panel 3).

Taking into account 1254 data points spread across 108 different conditions, the most likely feedback in the AKT pathway is a strong effect of AKT back to SIN1, which only weakly inhibits MTORC2, combined with a weak signal from MTORC1 to IRS1 S312 (but unlikely through S6K), which is however amplified by IRS1.

5.3.7. USING ISA TO TEST THE AGREEMENT BETWEEN DATASETS

Individual datasets are often insufficient to constrain most of the signaling strengths. Figure 5.10A shows how much each parameter is constrained by each dataset, as well as all datasets together. We can see that in many cases, a parameter is constrained by a single dataset. In this way, each dataset provides complementary information, and thus all datasets together are able to constrain most of the parameters to some extent.

Since datasets often provide different measurements, it is not clear how they can be directly compared against each other. However, we can use f-ISA to test whether datasets agree with each other given a model. To explore this, we calculated the disagreement in each parameter between any pair of the four datasets. We calculate the disagreement by taking one minus the overlap coefficient between the posterior 90% confidence intervals (CI) of the parameter given each dataset, where the overlap coefficients is the size of the intersection divided by minimum size of either CI. A disagreement of 1 means that the posterior 90% CIs do not overlap, whereas a disagreement of 0 means that the CI of one posterior is completely contained within the other, or are even exactly the same. Figure 5.10B shows this disagreement between datasets. Reassuringly, for most of the parameters the disagreement is small, indicating that in most cases the posterior distribution is accurate between datasets (an example is shown in Figure 5.10C). Note that this also includes many cases where one or both of the datasets simply do not constrain a parameter at all, hence the model may not be able to give a precise prediction for a parameter based on a dataset, but the posterior is still accurate in that it quantifies the uncertainty correctly.

There are, however, also some parameters for which the datasets disagree. Most disagreement occurs between the pre-treatment and intervention datasets. For example, the strength of signaling from MEK to ERK is estimated differently by the pre-treatment

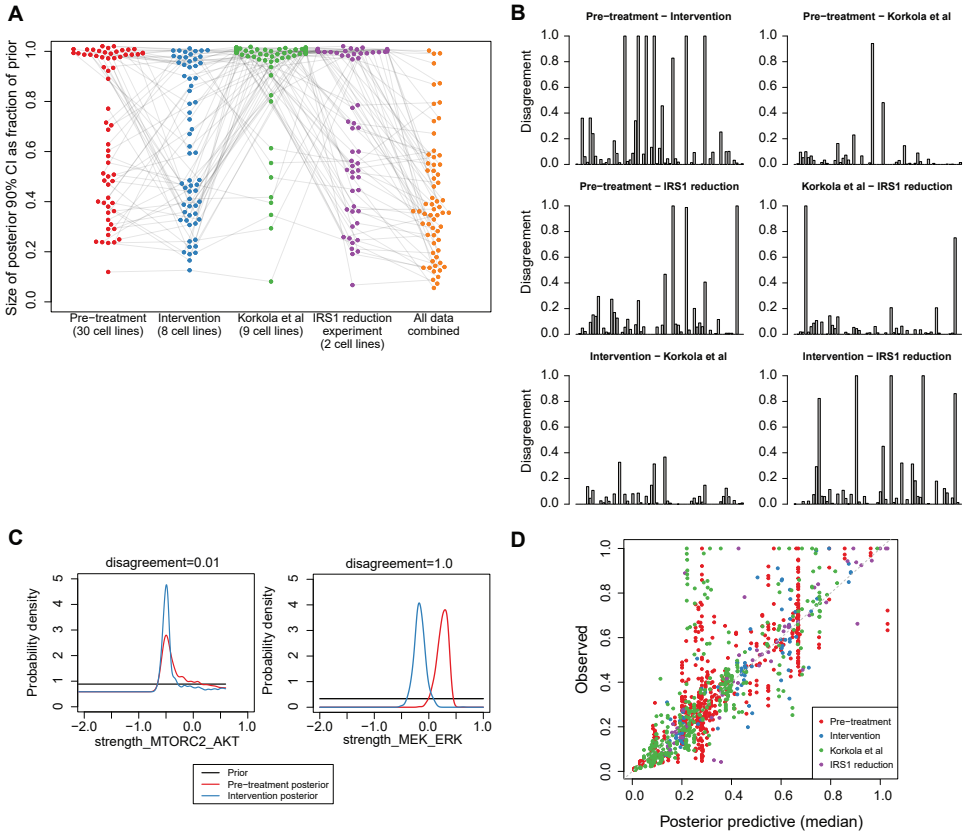


Figure 5.10: **Comparison of the model fit and parameter estimates across datasets.** (A) Reduction in uncertainty for each of the 63 signaling parameters, given each dataset separately and together. Each dot indicates one parameter. Grey lines connect the same parameters between datasets. Some parameters are constrained by multiple datasets, while other parameters are uniquely constrained by one dataset. (B) Disagreement in parameter estimates between datasets; each bar represents on of the 63 signaling parameters. Disagreement is calculated as one minus the overlap coefficient, that is, $1 - IS/\min(CI_1, CI_2)$, where CI_i is size of the posterior 90% confidence interval given each dataset, and IS is the size of the intersection of the two CIs. Detailed figures with labels are included as Supplementary Figure 1. (C) Examples of two parameters for which the pretreatment dataset and the intervention dataset either agree (top panel) or disagree (bottom panel). (D) Scatter plot of all fitted data, represented by the median of the posterior predictive distribution, against the observed values.

and intervention datasets. Assuming that the measurements in both datasets can be trusted, this would indicate that the model is overestimating its certainty. Nevertheless, despite the disagreement in some parameters, the model can provide a good fit against all data simultaneously (Figure 5.10D), and the resulting joint posterior is the most likely compromise between the posteriors inferred by the datasets separately.

5.3.8. IDENTIFICATION OF UNEXPLAINED DATA POINTS

Such models can also be used to identify unexpected parts of the data. Although the model can explain a large part of the data (Figure 5.10D), there are also data points which cannot be recapitulated by the model. For example, in the intervention dataset, EGF stimulation leads to a large increase in ERK phosphorylation in four of the eight cell lines, which the model cannot reproduce (see Figure 5.7A). The cell line specificity is partially explained by EGFR expression, but not completely, given that despite EGFR expression is included as input the model cannot explain the variability observed. In addition to the cell line specificity, there is only a very modest increase in MEK phosphorylation upon EGF stimulation, raising the question of how the signal from EGF to ERK is transmitted. One explanation for a difference in the changes of MEK and ERK phosphorylation could be that MEK amplifies the signal. However, the activation function allows for a signal amplification of up to 10-fold, and hence the constraints posed by all other data do not seem to allow this explanation. An explanation rooted in signal amplification also does not explain why this occurs only in some of the cell lines. Alternative explanations might be that there are additional feedback mechanisms or feed-forward signaling pathways which transmit the EGF signal downstream, that other phosphorylation sites are involved, or that the discrepancy is a result of the specific timing of starvation, inhibition and stimulation.

Another interesting part of the data that cannot be explained by the model, is the increased S6 phosphorylation in HCC1954 upon NT-157 pretreatment (Figure 5.9C). It has been shown that NT-157 treatment can induce MAPK pathway activation by inducing IGF1R-SHC complex formation [46], and the increased MAPK signaling could lead to elevated S6 phosphorylation [48, 49]. These are links that could be included in the model, although it would then still be unclear why the increased S6 phosphorylation is specific for HCC1954 and does not occur in the BT474 cell line. Alternatively, other pathways may be activated as a result of the disruption of IRS1 signaling, such as a stress response, potentially leading to S6K activation. Indeed, following treatment with NT-157, we did observe a reduction in cell viability of the HCC1954 cell line, while the BT474 cells appeared unaffected by the treatment.

These two examples show that the computational model can highlight cases where our knowledge, as summarized in the model, is incomplete and unable to fully describe the data. To resolve these discrepancies, additional rounds of model extension, combined with further experimental measurements to constrain the parameters, can be performed. Alternatively, the identified discrepancies can help define precise questions for follow-up functional genetic screening.

5.4. DISCUSSION

As our knowledge of biological signaling networks grows, it is increasingly difficult to fully understand the behavior of these networks. Computational modeling then provides a means to quantify the interactions as well as the contributions of different signaling molecules and genetic aberrations. We can also use these models to test whether our knowledge is sufficient to explain the data. As our models of biological signaling get more complex, we will also increasingly need to integrate multiple datasets to be able to identify the key parameters.

In this work, we used four datasets to constrain the parameters in a signaling model. Each of the data sets provided complementary information, and the largest number of the parameters could only be constrained with all four datasets employed simultaneously. We also described several non-intuitive behaviors of the identifiability of parameters. We found that feedback loops can sometimes be partially identified from non-intervention data, including the negative feedback from ERK to RAF, as well as a feedback from MTORC2 to MTORC1 used to reduce AKT signaling activity in cell lines with high MAPK pathway activity. Additionally, for the feedback from ERK to RAF, we found that adding more data to the inference can lead to an increased uncertainty in individual parameters. This is not caused by inconsistency of the data but by shifts in which signaling activities are more likely to explain all data simultaneously. We find that the aggregate uncertainty in all parameters does decrease when more data is employed in the inference.

Various other modeling approaches have previously been used to understand these signaling pathways. Thobe et al used logical modeling of MTORC2 signaling, including hypotheses of the regulation through SIN1 [50]. However, they were not able to resolve a preference between models, and given such a non-quantitative approach all models were found to be in agreement with the data. We find that using a quantitative approach as presented here, it is possible to deduce a preference for one feedback loop over another. Dalle Pezze et al constructed ODE models of mTOR signaling [51], although they did not include the SIN1 feedbacks (these had not been reported yet). In their ODE model, they used fixed parameter values, although they did perform a sensitivity analysis around the optimum. Given that in our study there are 108 model simulations required to evaluate the model with respect to the four datasets, and the many parameter values that need to be considered to characterize the multidimensional parameter space, it is presently not feasible to use such detailed ODE models in combination with a full uncertainty analysis. Nevertheless, with a quasi-steady state approach we were able to quantify the uncertainty in the feedback mechanisms.

Alongside the abovementioned advantages, the f-ISA modeling framework posseses several disadvantages as well. First, the model is restricted to known mechanisms. It is possible that other links provide an equally good fit. For example, there could be other negative feedbacks ending in or upstream of MEK and AKT, or the involvement of other phosphorylation sites on IRS1. Our approach does not search for additional links that may fit the data better. Rather, the goal is to test whether a particular model can describe the data well. Unfortunately, combining extensive parameter uncertainty analysis with a search for network topology is computationally intractable at present. Second, the calculation of signaling with feedback events as done here is significantly slower than

the original ISA modeling approach without feedback. It is therefore still impractical to infer signaling activities for large models with feedback while also including full dose response curves, especially with multiple drugs. For the dose response data presented in Jastrzebski et al [27], this would add another 300 model conditions per drug, as well as additional model parameters. It may be feasible to perform inference sequentially, thereby reducing the number of model evaluations required [52]. Alternatively, evaluating the gradient of the likelihood and using sampling algorithms that can leverage this gradient information may speed up the inference [53, 54].

A third limitation is that f-ISA is currently only able to handle feedback events which occur on one timescale. We focused here on fast-acting post-translational feedback. This approach could also be used to study transcriptional regulation; in this case total protein or mRNA levels could be used as inference data instead of phosphorylation data. It is also possible to include both transcriptional and post-translational feedback in the same model, but this would assume that the feedbacks occur on approximately the same timescale, which is probably unreasonable. To disentangle feedback mechanisms acting over different timeframes, it would be necessary to add further extensions that can calculate steady states with feedback loops occurring on multiple timescales, or it may be more beneficial to switch to dynamic models in this case.

Despite these limitations, we believe f-ISA is a useful approach to delineate context-dependent activities in signaling networks with multiple feedback paths. By iteratively incorporating additional detail in the signaling networks, and additional measurements to constrain the parameters, it is possible to obtain an increasingly thorough, quantitative understanding of a regulatory signaling network.

5.5. ACKNOWLEDGMENTS

We are grateful to Jordi Vidal Rodriguez for experimental support with the on-treatment, intervention RPPA experiment.

REFERENCES

[1] F. H. Groenendijk and R. Bernards, *Drug resistance to targeted therapies: Deja vu all over again,* Molecular Oncology **8**, 1067 (2014).

[2] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, *et al.*, *Systematic identification of genomic markers of drug sensitivity in cancer cells.* Nature **483**, 570 (2012).

[3] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. a. Margolin, *et al.*, *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature **483**, 603 (2012).

[4] B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, *et al.*, *Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset,* Cancer Discovery (2015).

[5] J. Stelling, U. Sauer, Z. Szallasi, F. J. Doyle, and J. Doyle, *Robustness of cellular functions,* Cell **118**, 675 (2004).

[6] M. K. Dougherty, J. Müller, D. a. Ritt, M. Zhou, X. Z. Zhou, *et al.*, *Regulation of Raf-1 by direct feedback phosphorylation.* Molecular cell **17**, 215 (2005).

[7] R. Fritsche-Guenther, F. Witzel, A. Sieber, R. Herr, N. Schmidt, *et al.*, *Strong negative feed-*

back from Erk to Raf confers robustness to MAPK signalling. Molecular systems biology **7**, 489 (2011).

[8] O. J. Shah, Z. Wang, and T. Hunter, *Inappropriate activation of the TSC/Rheb/mTOR/S6K cassette induces IRS1/2 depletion, insulin resistance, and cell survival deficiencies.* Current biology **14**, 1650 (2004).

[9] P. Liu, W. Gan, H. Inuzuka, A. S. Lazorchak, D. Gao, *et al.*, *Sin1 phosphorylation impairs mTORC2 complex integrity and inhibits downstream Akt signalling to suppress tumorigenesis.* Nature cell biology **15**, 1340 (2013).

[10] G. Yang, D. S. Murashige, S. J. Humphrey, and D. E. James, *A Positive Feedback Loop between Akt and mTORC2 via SIN1 Phosphorylation,* Cell Reports **12**, 937 (2015).

[11] S. J. Humphrey, G. Yang, P. Yang, D. J. Fazakerley, J. Stöckli, J. Y. Yang, and D. E. James, *Dynamic adipocyte phosphoproteome reveals that akt directly regulates mTORC2,* Cell Metabolism **17**, 1009 (2013).

[12] K. D. Copps and M. F. White, *Regulation of insulin sensitivity by serine/threonine phosphorylation of insulin receptor substrate proteins IRS1 and IRS2,* Diabetologia **55**, 2565 (2012).

[13] Y. H. Lee, J. Giraud, R. J. Davis, and M. F. White, *c-Jun N-terminal kinase (JNK) mediates feedback inhibition of the insulin signaling cascade,* Journal of Biological Chemistry **278**, 2896 (2003).

[14] C. J. Carlson, M. F. White, and C. M. Rondinone, *Mammalian target of rapamycin regulates IRS-1 serine 307 phosphorylation,* Biochemical and Biophysical Research Communications **316**, 533 (2004).

[15] O. J. Shah and T. Hunter, *Turnover of the Active Fraction of IRS1 Involves Raptor-mTOR- and S6K1-Dependent Serine Phosphorylation in Cell Culture Models of Tuberous Sclerosis,* Molecular and Cellular Biology **26**, 6425 (2006).

[16] J. Zhang, Z. Gao, J. Yin, M. J. Quon, and J. Ye, *S6K directly phosphorylates IRS-1 on Ser-270 to promote insulin resistance in response to TNF-α signaling through IKK2,* Journal of Biological Chemistry **283**, 35375 (2008).

[17] The UniProt Consortium, *UniProt: The universal protein knowledgebase,* Nucleic Acids Research **45**, D158 (2016), 1611.06654 .

[18] P. V. Hornbeck, B. Zhang, B. Murray, J. M. Kornhauser, V. Latham, and E. Skrzypek, *PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations,* Nucleic Acids Research **43**, D512 (2015).

[19] D. Hartley and G. M. Cooper, *Role of mTOR in the degradation of IRS-1: Regulation of PP2A activity,* Journal of Cellular Biochemistry **85**, 304 (2002).

[20] X. Wan, B. Harkavy, N. Shen, P. Grohar, and L. J. Helman, *Rapamycin induces feedback activation of Akt signaling through an IGF-1R-dependent mechanism,* Oncogene **26**, 1932 (2007).

[21] T. Okuzumi, D. Fiedler, C. Zhang, D. C. Gray, B. Aizenstein, R. Hoffman, and K. M. Shokat, *Inhibitor hijacking of Akt activation,* Nature Chemical Biology **5**, 484 (2009).

[22] E. K. H. Han, J. D. Leverson, T. McGonigal, O. J. Shah, K. W. Woods, T. Hunter, V. L. Giranda, and Y. Luo, *Akt inhibitor A-443654 induces rapid Akt Ser-473 phosphorylation independent of mTORC1 inhibition,* Oncogene **26**, 5655 (2007).

[23] B. N. Kholodenko, *Cell-signalling dynamics in time and space.* Nature reviews. Molecular cell biology **7**, 165 (2006).

[24] F. Eduati, V. Doldàn-Martelli, B. Klinger, T. Cokelaer, A. Sieber, *et al.*, *Drug resistance mechanisms in colorectal cancer dissected with cell type-specific dynamic logic models,* Cancer Research **77**, 3364 (2017).

[25] B. N. Kholodenko, A. Kiyatkin, F. J. Bruggeman, E. Sontag, H. V. Westerhoff, and J. B. Hoek, *Untangling the wires: A strategy to trace functional interactions in signaling and gene networks,* Proceedings of the National Academy of Sciences **99**, 12841 (2002).

[26] I. Stelniec-Klotz, S. Legewie, O. Tchernitsa, F. Witzel, B. Klinger, *et al.*, *Reverse engineering a hierarchical regulatory network downstream of oncogenic KRAS.* Molecular systems biology **8**, 601 (2012).

[27] K. Jastrzebski, B. Thijssen, R. J. C. Kluin, K. de Lint, I. J. Majewski, R. L. Beijersbergen, and L. F. A. Wessels, *Integrative modeling identifies key determinants of inhibitor sensitivity in breast cancer cell lines,* Cancer Research (2008).

[28] T.-R. Xu, V. Vyshemirsky, A. Gormand, A. von Kriegsheim, M. Girolami, *et al.*, *Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species.* Science signaling **3**, ra20 (2010).

[29] B. Thijssen, T. M. H. Dijkstra, T. Heskes, and L. F. A. Wessels, *Bayesian data integration for quantifying the contribution of diverse measurements to parameter estimates.* Bioinformatics **34**, 803 (2017).

[30] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf, *Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems,* Journal of The Royal Society Interface **6**, 187 (2009).

[31] J. Saez-Rodriguez, L. Alexopoulos, M. Zhang, M. K. Morris, D. A. Lauffenburger, and P. K. Sorger, *Comparing signaling networks between normal and transformed hepatocytes using discrete logical models,* Cancer research **71**, 5400 (2011).

[32] R. Schlatter, K. Schmich, I. Avalos Vizcarra, P. Scheurich, T. Sauter, *et al.*, *ON/OFF and beyond–a boolean model of apoptosis.* PLoS computational biology **5**, e1000595 (2009).

[33] R. Tarjan, *Depth-First Search and Linear Graph Algorithms,* SIAM Journal on Computing **1**, 146 (1972).

[34] K. R. Coombes, S. Neeley, C. Joy, J. Hu, K. Baggerly, and P. Roebuck, *SuperCurve: RPPA Analysis Package,* (2017).

[35] B. A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi, and H. Kitano, *CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks,* Proceedings of the IEEE **96**, 1254 (2008).

[36] N. Le Novère, M. Hucka, H. Mi, S. Moodie, F. Schreiber, *et al.*, *The Systems Biology Graphical Notation,* Nature biotechnology **27**, 735 (2009).

[37] B. Thijssen, T. M. H. Dijkstra, T. Heskes, and L. F. A. Wessels, *BCM: toolkit for Bayesian analysis of Computational Models using samplers,* BMC Systems Biology **10**, 100 (2016).

[38] H. G. Katzgraber, S. Trebst, D. A. Huse, and M. Troyer, *Feedback-optimized parallel tempering Monte Carlo,* Journal of Statistical Mechanics: Theory and Experiment **2006**, P03018 (2006), 0602085v3 .

[39] D. Turek, P. de Valpine, C. J. Paciorek, and C. Anderson-Bergman, *Automated Parameter Blocking for Efficient Markov-Chain Monte Carlo Sampling,* Bayesian Analysis **12**, 465 (2017),

1503.05621 .

[40] A. Gelman and X.-L. Meng, *Simulating normalizing constants: from importance sampling to bridge sampling to path sampling,* Statistical Science **13**, 163 (1998).

[41] D. J. MacKay, *Model Comparison and Occam's Razor,* in *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, 2003) pp. 343–354.

[42] J. E. Korkola, E. A. Collisson, L. Heiser, C. Oates, N. Bayani, *et al.*, *Decoupling of the PI3K Pathway via Mutation Necessitates Combinatorial Treatment in HER2+ Breast Cancer,* Plos One **10**, e0133219 (2015).

[43] J. F. Ohren, H. Chen, A. Pavlovsky, C. Whitehead, E. Zhang, *et al.*, *Structures of human MAP kinase kinase 1 (MEK1) and MEK2 describe novel noncompetitive kinase inhibition,* Nature Structural and Molecular Biology **11**, 1192 (2004).

[44] S. D. Barrett, A. J. Bridges, D. T. Dudley, A. R. Saltiel, J. H. Fergus, *et al.*, *The discovery of the benzhydroxamate MEK inhibitors CI-1040 and PD 0325901,* Bioorganic and Medicinal Chemistry Letters **18**, 6501 (2008).

[45] P. Lito, A. Saborowski, J. Yue, M. Solomon, E. Joseph, *et al.*, *Disruption of CRAF-Mediated MEK Activation Is Required for Effective MEK Inhibition in KRAS Mutant Tumors,* Cancer Cell **25**, 697 (2014).

[46] H. Reuveni, E. Flashner-Abramson, L. Steiner, K. Makedonski, R. Song, *et al.*, *Therapeutic destruction of insulin receptor substrates for cancer treatment,* Cancer Research **73**, 4383 (2013), 15334406 .

[47] N. Ibuki, M. Ghaffari, H. Reuveni, M. Pandey, L. Fazli, *et al.*, *The Tyrphostin NT157 Suppresses Insulin Receptor Substrates and Augments Therapeutic Response of Prostate Cancer,* Molecular Cancer Therapeutics **13**, 2827 (2014).

[48] M. Pende, S. H. Um, V. Mieulet, V. L. Goss, J. Mestan, *et al.*, *S6K1-/-/S6K2-/- Mice Exibit Perinatal Lethality and Rapamycin-Sensitive 5'-Terminal Oligopyrimidine mRNA Translation and Reveal a Mitogen-Activated Protein Kinase-Dependent S6 Kinase Pathway,* Molecular and Cellular Biology **24**, 3112 (2004).

[49] P. P. Roux, D. Shahbazian, H. Vu, M. K. Holz, M. S. Cohen, J. Taunton, N. Sonenberg, and J. Blenis, *RAS/ERK signaling promotes site-specific ribosomal protein S6 phosphorylation via RSK and stimulates cap-dependent translation.* The Journal of biological chemistry **282**, 14056 (2007), NIHMS150003 .

[50] K. Thobe, C. Sers, and H. Siebert, *Unraveling the regulation of mTORC2 using logical modeling.* Cell communication and signaling **15**, 6 (2017).

[51] P. Dalle Pezze, A. G. Sonntag, A. Thien, M. T. Prentzell, M. Godel, *et al.*, *A Dynamic Network Model of mTOR Signaling Reveals TSC-Independent mTORC2 Regulation,* Science Signaling **5**, ra25 (2012).

[52] B. Thijssen and L. F. A. Wessels, *Approximating multivariate posterior distribution functions from Monte Carlo samples for sequential Bayesian inference,* (2017), arXiv:1712.04200 .

[53] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, *Hybrid Monte Carlo,* Physics Letters B **55**, 2774 (1987).

[54] M. D. Hoffman and A. Gelman, *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo,* Journal of Machine Learning Research **15**, 1593 (2014).

SUPPLEMENTARY MATERIAL

Supplementary Material is available online at https://doi.org/10.1101/268359.

6

APPROXIMATING MULTIVARIATE POSTERIOR DISTRIBUTION FUNCTIONS FROM MONTE CARLO SAMPLES FOR SEQUENTIAL BAYESIAN INFERENCE

Bram THIJSSEN
Lodewyk F.A. WESSELS

Parts of this chapter have been posted on arXiv (1712.04200).

ABSTRACT

An important feature of Bayesian statistics is the possibility to do sequential inference: the posterior distribution obtained after seeing a first dataset can be used as prior for a second inference. However, when Monte Carlo sampling methods are used for the inference, we only have one set of samples from the posterior distribution, which is typically insufficient for accurate sequential inference. In order to do sequential inference in this case, it is necessary to estimate a functional description of the posterior probability distribution from the Monte Carlo samples. Here, we explore whether it is feasible to perform sequential inference based on Monte Carlo samples, in a multivariate context. To approximate the posterior distribution, we can use either the apparent density based on the sample positions (density estimation) or the relative posterior probability of the samples (regression). Specifically, we evaluate the accuracy of kernel density estimation, Gaussian mixtures, vine copulas and Gaussian process regression; and we test whether they can be used for sequential Bayesian inference. Additionally, both the density estimation and the regression methods can be used to obtain a post-hoc estimate of the marginal likelihood. In low dimensionality, Gaussian processes are most accurate, whereas in higher dimensionality Gaussian mixtures or vine copulas perform better. We show that sequential inference can be computationally more efficient than joint inference, and we also illustrate the limits of this approach with a failure case. Since the performance is likely to be case-specific, we provide an R package *mvdens* that provides a unified interface for the probability distribution approximation methods.

6.1. INTRODUCTION

In Bayesian statistics, variables are given a probability distribution that specifies our knowledge about the variables. This distribution can then be updated based on available data using Bayes' theorem. An important advantage of this approach is that inference can be done sequentially. That is, when we have obtained a posterior distribution after seeing a first dataset, we can use this posterior as prior for inference with a next dataset.

For complex models, Bayesian inference is often achieved with Monte Carlo sampling. This allows us to obtain samples from posterior distributions which would otherwise be intractable. However, when we want to use Monte Carlo sampling results for sequential inference, we only have the set of samples to use as prior. In principle we can use the samples directly for sequential inference, through importance reweighting, but the sequential posterior will then only be evaluated at those sample points and this will typically not be accurate. If we instead have a functional representation of the first posterior, we could use this functional representation as prior for the second Monte Carlo inference. To do sequential inference it is therefore necessary to estimate a functional representation of the posterior from the samples.

There are various methods which can estimate a functional approximation for the posterior distribution from samples. Broadly, this can be done in two ways. We can treat the posterior distribution estimation as a general density estimation task, where we estimate the density only from the location of the samples. Several popular density estimation methods include kernel density estimation (KDE), Gaussian mixtures (GM) and

copulas or vine copulas (VC). An alternative option is to treat the posterior distribution approximation as a regression problem, since alongside the sample positions, we usually also have the relative posterior probability at the sample locations. This has the advantage of using additional information regarding the posterior distribution, but presents its own challenges as well. In particular, the regression function must integrate to one over the prior domain for it to be a proper density function. It can be challenging to meet this constraint while fitting a function through many sample points. One regression method that has sufficient flexibility for this is Gaussian process (GP) regression.

In this manuscript, we will explore the use of density function approximations to enable sequential inference with Monte Carlo sampling. We will consider each of the aforementioned methods (KDE, GM and VC density estimation, and GP regression). We first test their performance in approximating a known density, then test their accuracy in approximating a posterior distribution, and subsequently test their performance in sequential inference. The posterior distribution approximations can also be used to obtain an estimate of the marginal likelihood. Finally, we test whether sequential inference of two datasets is computationally faster than inference with the two datasets jointly.

Posterior distribution approximations are also used in several other areas of Bayesian computation. First, in Monte Carlo sampling itself, a proposal distribution is used, and sampling is most efficient when the proposal distribution closely resembles the true target probability density. There have been many efforts in creating efficient proposal distributions, including using some of the density approximation methods that we consider here, for example with vine copulas [1] and Gaussian processes [2]. Second, posterior distribution approximations have been used in schemes for parallelizing MCMC inference [3]. In this case the inference is split into parts, and the resulting subposteriors are combined using a posterior distribution approximation to recover the full posterior. Third, in the area of Bayesian filtering [4], a posterior distribution is updated when new data arrives over time, which also relies on posterior distribution approximations. In the present study, we explicitly test the accuracy in approximating posterior distributions, and, apart from the use of such approximations in sequential inference, the results presented here may be relevant for these other areas as well.

6.2. METHODS

To use the posterior obtained from Monte Carlo sampling in sequential inference, we need to approximate the distribution

$$P(\mathbf{x}|y) = \frac{P(y|\mathbf{x})P(\mathbf{x})}{P(y)} \approx \hat{P}(\mathbf{x}),$$

where $\mathbf{x}$ is the D-dimensional variable of interest and y represents the inference data. In the notation of the approximation $\hat{P}(\mathbf{x})$ we have dropped the conditioning on y for brevity.

The approximation $\hat{P}(\mathbf{x})$ needs to be constructed from samples $\mathbf{x}_i$ that have been drawn from the posterior $P(\mathbf{x}|y)$. The approximations can be achieved using density estimation or through regression, see Figure 6.1. In all subsequent equations, N is the number of Monte Carlo samples and $\mathbf{x}_i$ is the D-dimensional value of the ith sample. While i indexes the samples, j indexes the dimensions, so note that x_j (non-bold, and

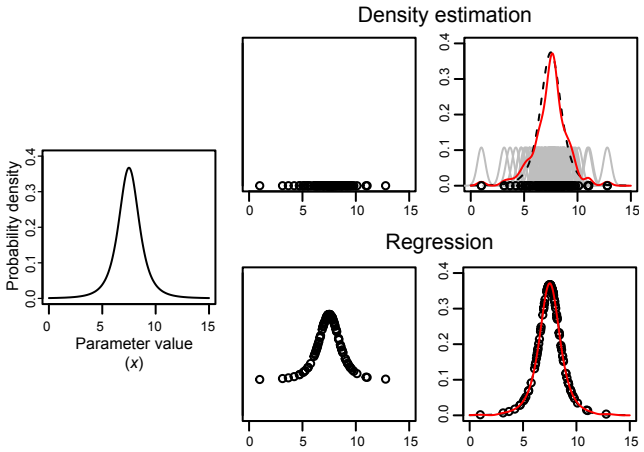


Figure 6.1: **Reconstructing a probability density function by density estimation or regression.** Density estimation uses the sample location, while regression uses both the sample location and the unnormalized, relative probability to reconstruct a normalized probability density function. The example function is a t-distribution with ν=4 centered at 7.5.

indexed by j) refers to the jth element of the D-dimensional value of $\mathbf{x}$. For the regression methods, we assume that the relative, unnormalized probability is available, and it is represented by p_i for sample i, (that is, $p_i = P(y|\mathbf{x}_i)P(\mathbf{x}_i)$).

6.2.1. DENSITY ESTIMATION

The density estimation methods use the sample positions, $\mathbf{x}_i$, to reconstruct an approximation to the probability density function. Below we briefly introduce three density estimation methods: kernel density estimation, Gaussian mixtures and vine copulas, with several variations; see Figure 6.2 for a bivariate illustration.

KERNEL DENSITY ESTIMATE

The kernel density estimate approximation is given by

$$\hat{P}_{KDE}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} K(\mathbf{x}-\mathbf{x_i}),$$

where $K(\mathbf{x}-\mathbf{x_i})$ is a kernel function. We take the kernel function to be a multivariate normal distribution $\mathcal{N}(0,\Sigma)$. When $D \leq 4$ we estimate a full covariance matrix using multivariate plug-in bandwidth selection [5]. When $D > 4$ we estimate a diagonal covariance matrix with the diagonal entries estimated using univariate plug-in bandwidth selection [6], scaled by a factor $N^{-1/(D+4)}$ based on the normal reference rule [7].

GAUSSIAN MIXTURE

The Gaussian mixture approximation is given by

$$\hat{P}_{GM}(\mathbf{x}) = \sum_{g=1}^{G} c_g \mathcal{N}(\mathbf{x}|\mu = \boldsymbol{\mu}_g, \Sigma = \Sigma_g)$$

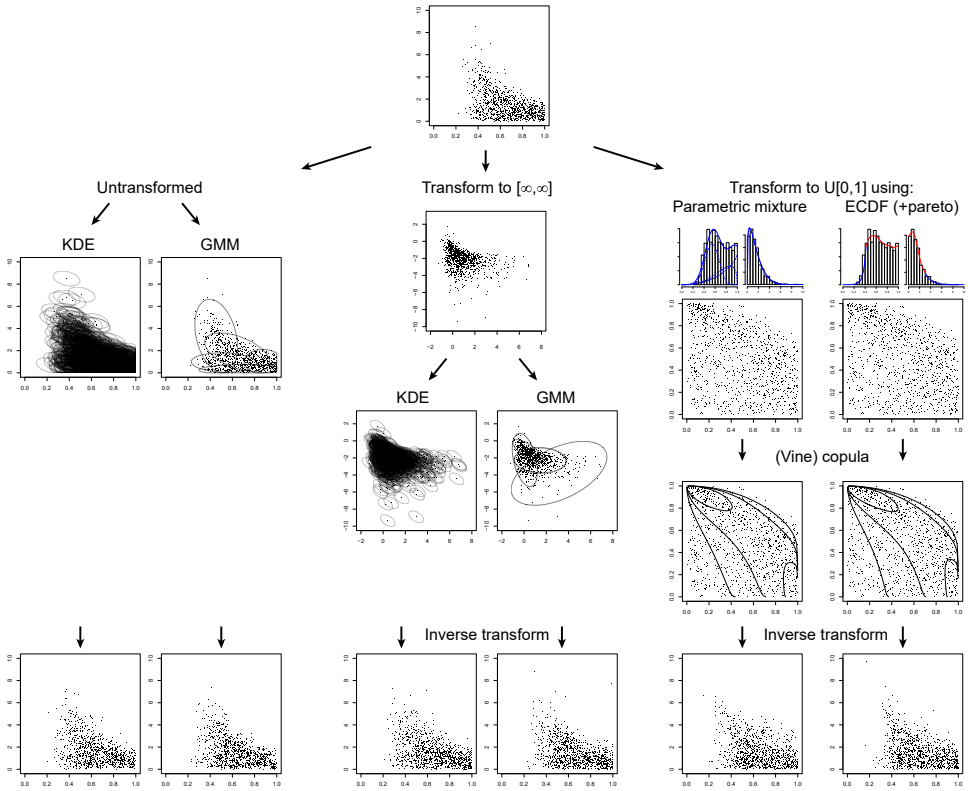


Figure 6.2: **Density estimation methods applied to a bivariate example.** At the top, we start with samples obtained through Monte Carlo sampling from the posterior of two variables. The two variables are β_{kill} and γ from the (bounded) Lotka-Volterra example discussed later. On the left, a kernel density estimate or a Gaussian mixture is fitted to the samples. In the middle, the variables are first transformed to an unbounded domain (in this case through a scaled logit transform) before a KDE or GM is fitted. On the right, the variables are transformed to have uniform marginal distributions between 0 and 1, using either a parametric mixture or an empirical cumulative distribution with Pareto tails. Subsequently, a copula function is fitted to the transformed variables. Finally, on the bottom row, new samples are drawn from each of the approximations. Where necessary, the new samples are transformed with the inverse of the original transformation. In each case the distribution of the new samples is similar to the original sample distribution, but slight differences between the approximations can be observed as well.

where c_g, $\boldsymbol{\mu}_g$ and Σ_g are the proportion, mean and covariance of the gth component, G is the number of mixture components, and $\sum c_g = 1$. We use a full covariance matrix, and the parameters c, $\boldsymbol{\mu}$ and Σ are estimated using expectation-maximization. The number of components is selected by minimizing the Bayesian information criterion (BIC).

TRUNCATED GAUSSIAN MIXTURE

When the prior probability distribution $P(\mathbf{x})$ is bounded, we can use truncated Gaussians with known bounds in the mixture:

$$\hat{P}_{TGM}(\mathbf{x}) = \sum_{g=1}^{G} c_g \mathcal{N}_T(\mathbf{x}|\mu = \boldsymbol{\mu}_g, \Sigma = \Sigma_g, a = \mathbf{a}, b = \mathbf{b}),$$

where **a** and **b** are the known lower and upper bounds respectively. The parameters are estimated using expectation-maximization and the number of components selected by minimizing the BIC.

VINE COPULA

With copulas, the multivariate distribution is decomposed into marginal distributions and a description of the dependency structure. The copula density approximation is then given by

$$\hat{P}_{cop}(\mathbf{x}) = c(F_1(x_1), ..., F_D(x_D)) \prod_{j=1}^{D} f_j(x_j),$$

where c is a copula function, f_j is the marginal probability density function for dimension j, and F_j the corresponding marginal cumulative density function. Various different families of copula function exist; using the *R* package *VineCopula* [8], we evaluate various commonly used families and their rotations and select the optimal function by minimizing the Akaike information criterion (AIC).

For $D > 2$; a multi-dimensional copula function could be used, but we instead model the approximation using regular vine copulas [9], given by the equation

$$\hat{P}_{vc}(\mathbf{x}) = \prod_{l=1}^{D-1} \prod_{k=1}^{D-l} c_{k,(k+l)|(k+1),...,(k+l-1)} \prod_{j=1}^{D} f_j(x_j),$$

where the first two products are the pair-copulas and the third product contains the marginal densities as before. The bivariate pair-copula functions are selected as before by minimizing the AIC, and the vine structure is selected using a maximum spanning tree with Kendall's tau edge weights [10].

For the marginal distribution and density functions, common choices include empirical distribution functions and parametric distributions. We will consider these two options, as well as using Pareto tails and parametric mixtures, as described below.

Empirical distribution marginal An empirical marginal distribution function is given by

$$F_j(x_j) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{x_{i,j} \leq x_j},$$

where $\mathbf{1}_{x_{i,j} \leq x_j}$ is the indicator function. A corresponding density function is constructed using a 1-dimensional kernel density estimate

$$f_j(x_j) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{N}(x_{i,j}, \sigma_j),$$

where σ_j is estimated using plug-in bandwidth selection.

For the quantile function (the inverse of the cumulative distribution function) we use a linear interpolation of the empirical distribution function. When the prior has bounded support, samples are mirrored across the boundary to improve the estimate near the boundaries.

Pareto tails Since an empirical distribution can be inaccurate in the tails, we also consider augmenting the empirical density with Pareto tails. The distribution is then split in three parts, a body described by the empirical distribution function and kernel density estimate as before, and two tails described by a generalized Pareto distribution (GPD). An important choice is where to put the threshold beyond which data are used to fit the tail distribution [11]. We use the simple rule of thumb of using 10% of the samples to estimate a tail [12]. Since we have a tail on each side, we use the middle 80% of the samples for the body, and the upper and lower 10% of the samples to estimate the Pareto tail on each side:

$$F_j(x_j) = \begin{cases} q - qF_{j,\xi_{j,1}}(\frac{t_{j,1} - x_j}{\sigma_{j,1}}) & \text{if } x_j \leq t_{j,1} \\ (1-q) + qF_{j,\xi_{j,2}}(\frac{x_j - t_{j,2}}{\sigma_{j,2}}) & \text{if } x_j \geq t_{j,2} \\ F_{j,ECDF}(x_j) & \text{otherwise,} \end{cases}$$

where

$$F_\xi(z) = 1 - (1 + \xi - z)^{-1/\xi}$$

is the GPD function, q is the quantile used for the threshold ($q = 0.1$ for the 10% rule), and $t_{j,1}$ and $t_{j,2}$ are the lower and upper qth quantile of x_j respectively. $F_{j,ECDF}(x_j)$ is the empirical distribution function as before. To ensure continuity in the density function between the Pareto tail and the ECDF body, we set $\sigma_j = q / f_{j,KDE}(t_j)$. The shape parameter ξ_j is estimated by maximum likelihood, separately for each tail. The density function of the tails is given by the GPD density, scaled by q:

$$f_\xi(z) = \frac{q}{\sigma_j}(\xi z + 1)^{-(\xi+1)/\xi}.$$

In the case of bounded support, we do not use a Pareto tail unless the empirical density at the boundary is less than a threshold ϵ (which we set to $1/N$). While a GPD can handle a bounded support (by taking $\xi < 0$), we find this often leads to a poorer approximation than an empirical estimate with mirroring across the boundary.

Parametric mixtures The marginal densities can also be approximated with mixtures of parametric distributions. For unbounded variables we use a mixture of normals:

$$f_j(x_j) = \sum_{g=1}^{G} c_g \mathcal{N}(x_j | \mu = \mu_g, \sigma^2 = \sigma_g^2).$$

When there are known bounds, we use gamma distributions (when there is only a lower or upper bound) or beta distributions (when there is both a lower and upper bound) instead of normal distributions; these distributions are scaled, shifted and/or reflected to match the bounds. The parameters are estimated using expectation-maximization, and we select the number of components by minimizing the BIC.

6.2.2. REGRESSION

When the relative probability density at the sample positions is available, the density function can be estimated by regression. Typically, only the relative, unnormalized probability density will be available. In these cases it will then be necessary to normalize the regression function to ensure that it integrates to one over the prior domain.

When an estimate of the marginal likelihood $P(y)$ is available in addition to the samples, then the probability values can be normalized before entering the regression. If the approximation is accurate, this would ensures that the regression function is properly normalized as well, but we don't further explore this option of normalization with a known marginal likelihood here.

GAUSSIAN PROCESS

As regression method we employ Gaussian process regression, since it provides flexibility for approximating arbitrary density functions, and it handles multivariate regressors naturally. In order to handle unnormalized input densities, we multiply the Gaussian process predictive distribution with a scaling parameter. By calculating the integral of the predictive distribution (see appendix A), we can constrain the distribution to integrate to one by setting the scaling parameter to the reciprocal of the integral.

The behavior of Gaussian processes is characterized by their mean and covariance functions. We set the mean function to be zero everywhere, as we expect the probability to go to zero in regions where we do not have any samples. The predictive mean of the Gaussian process function based on the input samples X is then given by:

$$\hat{P}_{GP}(\mathbf{x}) = \frac{1}{Z} K(\mathbf{x}, X) K(X, X)^{-1} \mathbf{p},$$

where Z is the normalizing constant (see appendix A), $K(X_1, X_2)$ is the matrix obtained by applying the covariance function $k(\mathbf{x_1}, \mathbf{x_2})$ to all pairs of X_1 and X_2 (see e.g. [13] for more details on Gaussian processes), and $\mathbf{p}$ is the vector of unnormalized posterior probability densities at the sample locations as defined earlier.

As covariance function we consider two commonly used kernels, the squared exponential

$$k_{SE}(\mathbf{x}, \mathbf{x}^*) = \exp(\frac{r^2}{2l^2})$$

and the Matérn kernel with $\nu = 3/2$

$$k_{Mat32}(\mathbf{x}, \mathbf{x}^*) = (1 + \frac{\sqrt{3}r}{l}) \exp(-\frac{(\sqrt{3}r)}{l}),$$

where r is the Euclidean norm $|\mathbf{x} - \mathbf{x}^*|$ and l a length scale parameter. The parameter l is optimized by minimizing the root mean square error of $\hat{P}_{GP}(\mathbf{x})$ in 5-fold cross-validation.

A downside of using Gaussian processes for probability densities is that they do not naturally allow for a constraint that the function is non-negative everywhere. As a result, negative probability densities can occur. This could be circumvented by transforming the densities (for example by log transform (as done in [14]) or logistic transform (as done in [2])), but then the predictive function can no longer be normalized to integrate to one in the untransformed space. We found that, in our test cases, constraining Z to be positive during the optimization of l prevented large negative densities, and any remaining negative densities were typically very small and were pragmatically set to zero.

6.2.3. IMPORTANCE REWEIGHTING

As reference for the approximation methods, instead of constructing an approximate distribution function, we can also use the Monte Carlo samples from the initial inference directly and reweight them given the likelihood of the second dataset. That is, the samples are given weights

$$w_i = P(y_2|\mathbf{x}_i) / \sum_{j=1}^{N} P(y_2|\mathbf{x}_j),$$

where y_2 indicates the data in the second inference and $\mathbf{x}_i$ are the sample positions from the first inference as before. This can be viewed as importance sampling from the joint posterior distribution with the posterior of the first dataset as proposal distribution, with the fixed set of samples.

6.2.4. TRANSFORMATIONS FOR BOUNDED VARIABLES

Some of the approximation methods can explicitly handle a bounded support. In the other cases, we can use rejection sampling to discard samples outside the prior support. Alternatively, the variables can be transformed to an unbounded domain before applying the posterior approximation methods. We consider a log transform (when there is only a lower or upper bound) or a logit transform (when there is both a lower and upper bound), and scale, shift or reflect the variables as necessary. The probability density function is corrected for the transformation by multiplying with the derivative of the transform.

6.2.5. MARGINAL LIKELIHOOD ESTIMATION

When the approximation of the posterior distribution function can be normalized such that it integrates to one (as is the case for all methods used here), we can use the approximation to obtain an estimate of the marginal likelihood. Since $\hat{P}(\mathbf{x}) \approx P(\mathbf{x}|y)$, and

$$P(\mathbf{x}|y) = \frac{P(y|\mathbf{x})P(\mathbf{x})}{P(y)},$$

we can use a linear regression of the approximation probability density against the unnormalized posterior probability at each sample position and obtain an estimate $\hat{P}(y)$ of the marginal likelihood from the slope of the regression. Depending on the setting, it may be beneficial to log transform the probabilities:

$$\log \hat{P}(\mathbf{x}) = \log(P(y|\mathbf{x})P(\mathbf{x})) - \log \hat{P}(y)$$

and get an estimate of the log marginal likelihood from the intercept of the regression.

6.2.6. MONTE CARLO SAMPLING

As Monte Carlo sampling algorithms, we made use of three variants: parallel tempered Markov chain Monte Carlo (PT-MCMC) [15] with automated parameter blocking [16], sequential Monte Carlo (SMC) with MCMC proposal distributions [17], and nested sampling [18]. Marginal likelihood estimates were obtained by thermodynamic integration (when using PT-MCMC), by the resampling weights (when using SMC) and by sampling the mass ratios (when using nested sampling). The sampling and marginal likelihood estimation were done using the Bayesian inference software package BCM [19].

6.3. RESULTS

6.3.1. APPROXIMATING A KNOWN TARGET DENSITY

To test whether the density approximation methods can adequately describe a multivariate density function, we first attempted to reconstruct a known target distribution. We take a mixture of two multivariate Gaussians,

$$P(\mathbf{x}) = \frac{2}{3}\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1) + \frac{1}{3}\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2),$$

with random covariance matrices and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$ for the first test case. Figure 6.3A (left panel) shows 300 random samples drawn from this distribution for $D = 2$. We then compared how well the approximation methods can reconstruct this density using an increasing number of samples, and at increasing dimensionality (Figure 6.3A, right panels).

In the lower dimensional setting, Gaussian processes give the best approximation. Since the Gaussian processes can use the relative probability density at the sample positions, they have more information to create a good approximation, which allows a very good reconstruction already with few samples. In the higher dimensional setting however, the Gaussian processes do not perform as well. This is likely due to having only a single length scale parameter l (since we only consider isotropic covariance functions). We find that fitting such a regression through high dimensional multivariate sample points leads to an overdispersed distribution, which is limiting the performance.

At D=10, the Gaussian mixture approximation achieves higher accuracy than all other approaches, including Gaussian process regression. Among the density estimation methods, it is to be expected that the Gaussian mixture approximation is most accurate, since it has the same functional form as the target density.

To test the performance of the approximation methods in a multimodal setting, we separated the two Gaussians in space by setting $\mu_2 = \mu_1 + 10$ in every dimension (see Figure 6.3B). All methods do at least slightly worse than in the unimodal case, as evidenced

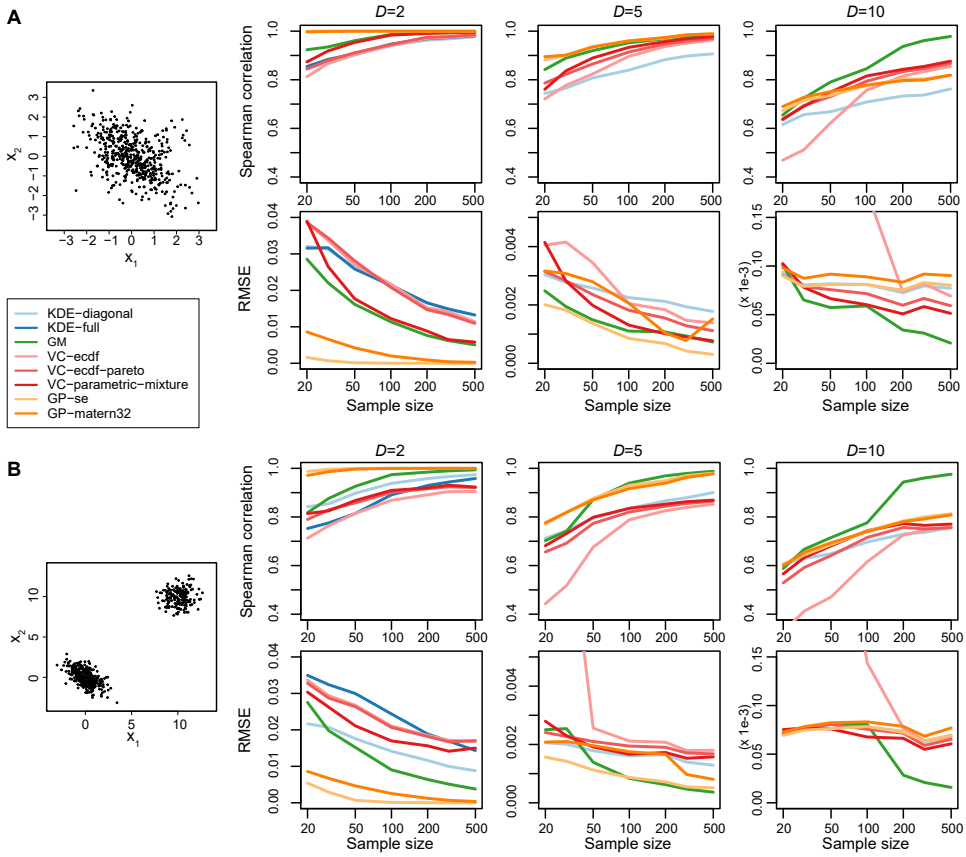


Figure 6.3: **Comparison of the approximation methods for reconstructing a Gaussian mixture with increasing dimensionality.** The density approximation was trained on the indicated number of training samples size, and the accuracy was evaluated by testing on 500 new samples regardless of training size. This procedure was repeated 100 times and the lines indicate the median Spearman correlation and root mean square error (RMSE) over these iterations.

by the lower Spearman correlations (note that the RMSE cannot be directly compared between the two cases because the mean density is different). In particular the vine copulas do significantly worse. This is likely due to the fact that the available copula functions are designed to describe the shape of a single mode, and are not necessarily suited for describing multimodal distributions. The behavior of the Gaussian mixture and Gaussian process approximations is similar to their behavior in the unimodal setting.

We also observe jumps in the performance of the Gaussian mixture approximation (e.g. between 100 and 200 samples for D=10). These jumps occur when the number of samples is sufficiently large for the Gaussian mixture approximation to identify that it is a mixture of two Gaussians rather than a single Gaussian distribution. For the GP kernels, we see that the squared exponential kernel has better performance than the heavier-tailed Matérn kernel in this case, which is to be expected given the exponential target distribution. For vine copulas, we see that using Pareto tails gives better performance than using only an ECDF/KDE marginal, especially for lower numbers of samples. However, even with Pareto tails, the empirical marginals are outperformed by parametric marginals in this case.

6.3.2. APPROXIMATING A POSTERIOR DISTRIBUTION

To test how the methods perform in approximating a posterior density function, we turned to a dynamic model of a predator-prey system. Specifically, we used a modified Lotka-Volterra system to model the interactions between the Canadian lynx and the showshoe hare [20]. This system was chosen because of the availability of several datasets, a modest number of parameters (5 dynamic parameters and 2 initial conditions for each dataset), and non-linearity in the system which likely leads to non-linearity in the posterior probability distribution of the parameters, making for a meaningful test case.

The model is given by the differential equations

$$\begin{aligned} \frac{dx}{dt} &= \alpha x - (\beta_{\text{kill}} + \beta_{\text{stress}}) xy \\ \frac{dy}{dt} &= \delta xy - \gamma y \end{aligned} ,$$

where x represents the hare population and y the lynx population. The populations are measured by their density, i.e. the number of individuals per area in arbitrary units. In the standard Lotka-Volterra model, there is a single parameter β for the effect of predation. We have split this effect into two parts, β_{kill} and β_{stress}, because it has been shown that at peak lynx density, the hares do not only die from increased predation, but also produce less offspring, which appears to be due to stress induced by the high threat of predation [20, 21]. The modeled natality (number of offspring per adult female in one breeding season) is given by $2 \cdot \exp(\alpha - \beta_{\text{stress}} y)$.

To accommodate multiple datasets, we include two parameters, $x_{0,j}$ and $y_{0,j}$ for each dataset j, giving the initial conditions that are used to simulate the model for that dataset. The model parameters can be inferred with each dataset separately by simulating the model from the respective initial conditions. Each dataset will then constrain the dynamic parameters and the initial conditions for that dataset, and will leave the initial

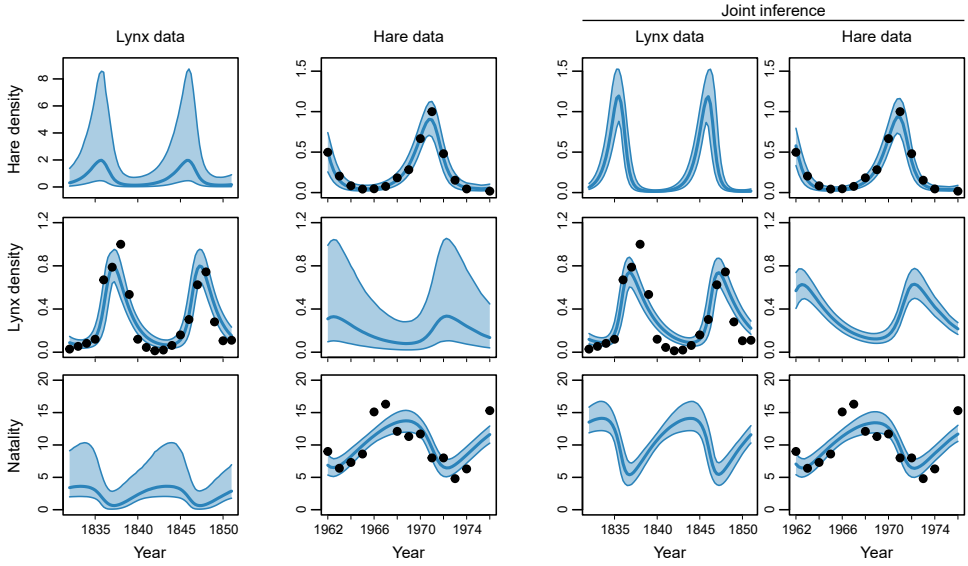


Figure 6.4: **Lynx-hare datasets and posterior predictive distributions.** The lynx data provides an estimate of lynx density (number of animals per surface area) and the hare data provides an estimate of hare density as well as natality (offspring per adult female per year). Black dots indicate the data, the thick blue line is the median and the shaded blue area the 90% confidence interval of the posterior predictive. The original data was normalized by dividing by the maximum observed value.

conditions of the other datasets unaffected. One could also have a single initial condition value x_0 and y_0 and simulate the model to cover the timespan of all datasets, but apart from the time difference, the measurements we used are also from different geographical regions, which do not necessarily have the same phase in the predator-prey cycle. We do assume that the dynamic parameters are the same for each dataset.

All of the parameters should be positive. To simplify the inference and approximations, we initially infer the parameters on log scale, so that there are no discontinuities in the posterior density (we lift this restriction of unbounded priors later).

We used two datasets to infer the parameters. The first dataset is the Hudson Bay Company data of lynx pelt records [22], which we will refer to as the *lynx data*; in particular we used the McKenzie River station data from 1832 through 1851. The second dataset is a study of a hare population and its reproductive output [23], from 1962 through 1976, which we will refer to as the *hare data*. Note that the lynx data only contains measurements of the lynxes while the hare data only contains measurements of the hares.

We then fitted the model to the hare and lynx dataset separately and to the two datasets together; Figure 6.4 shows the data and the posterior predictive distributions given the model. From the overlap of the posterior predictive and the data, it is clear that the model can adequately describe these datasets, both separately and jointly.

Figure 6.5A-C shows several aspects of the posterior obtained after seeing the lynx data. The posterior distribution is unimodal and has moderate and somewhat nonlinear correlations. The unimodality is the result of using relatively narrow priors. When

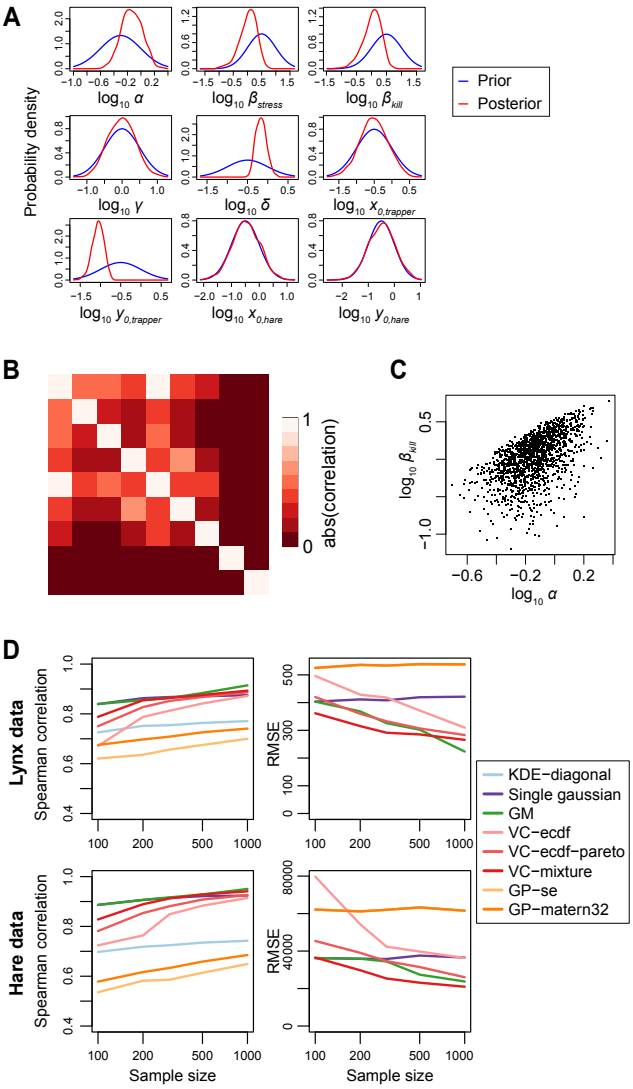


Figure 6.5: **Approximations of the posterior induced by the lynx and hare datasets.** (A) Marginal posterior densities after seeing the lynx data, the graphs are constructed using kernel density estimation with plug-in bandwidth selection. (B) Correlations between the parameters in the lynx posterior. (C) Scatter plot of the samples for one parameter combination. (D) Approximation accuracy as a function of sample size. Spearman correlation and root mean square error were calculated by taking the median over 100 iterations of repeated random subsampling validation, the training size (after subsampling) is indicated and a test size of 500 samples was used. 6.3.

wide priors are used, at least one other mode can be found, corresponding to oscillations through the data points at very high frequency; we therefore restricted the priors so that only the correct oscillation with a period of roughly 10 years is obtained.

Method	Lynx data	Hare data
Thermodynamic integration	0.46 ± 0.92	-34.7 ± 1.4
Sequential Monte Carlo	0.57 ± 0.42	-34.7 ± 0.36
Nested sampling	0.77 ± 0.65	-34.4 ± 0.64
Kernel density estimate	4.60	-29.1
Gaussian mixture	0.80	-34.5
Vine copula - mixture	1.20	-34.3
Gaussian process - SE	5.19	-28.8

Table 6.1: **Log marginal likelihood estimates.**

We then tested by cross-validation how well the approximations can describe the posterior distribution of the two datasets (see Figure 6.5D). As with the Gaussian mixture test case at similar dimensionality, the Gaussian mixture approximation gives good cross-validation performance. In this case, a vine copula with mixture marginals also produces a good approximation of the posterior (Spearman correlation $\rho \approx 0.9$ at a sample size of 1,000).

6.3.3. SEQUENTIAL INFERENCE

Having obtained reasonably accurate approximations of the posterior densities, we can test how they perform in sequential inference. To do this, we approximated the posterior from the lynx dataset with all methods using 1,000 samples, and use these approximations as prior for the hare dataset. If the approximations are accurate, the resulting posterior of the second inference should give the same result as a joint inference with the two datasets together.

Figure 6.6A shows the marginal probability density of one of the parameters, β_{kill}, from the datasets separately, the true joint, and with two approximation methods (importance reweighting and a gaussian mixture). As expected, importance reweighting provides a very poor approximation; a single sample receives almost all of the weight and the true joint posterior cannot be accurately estimated from essentially one sample. The Gaussian mixture approximation on the other hand provides a sequential posterior that is visually almost indistinguishable from the true joint. To quantify the performance, we calculated the Kolmogorov-Smirnov statistic for the marginal distribution of each of the parameters, based on the empirical cumulative distributions (see Figure 6.6B and C). Both Gaussian mixtures and vine copulas give sequential posteriors that are closest to the true joint. Gaussian processes and the KDE approximation perform worse, as expected given their poorer cross-validation performance.

6.3.4. MARGINAL LIKELIHOOD ESTIMATION

We can use the posterior distribution approximations to obtain an estimate of the marginal likelihood directly from the Monte Carlo samples (see Methods section). Table 6.1 shows the estimates obtained from three dedicated marginal likelihood estimation algorithms, compared to the estimates obtained directly from the samples using the posterior approximations. The posterior approximations that performed well in cross validation and sequential inference also provide accurate marginal likelihood estimates.

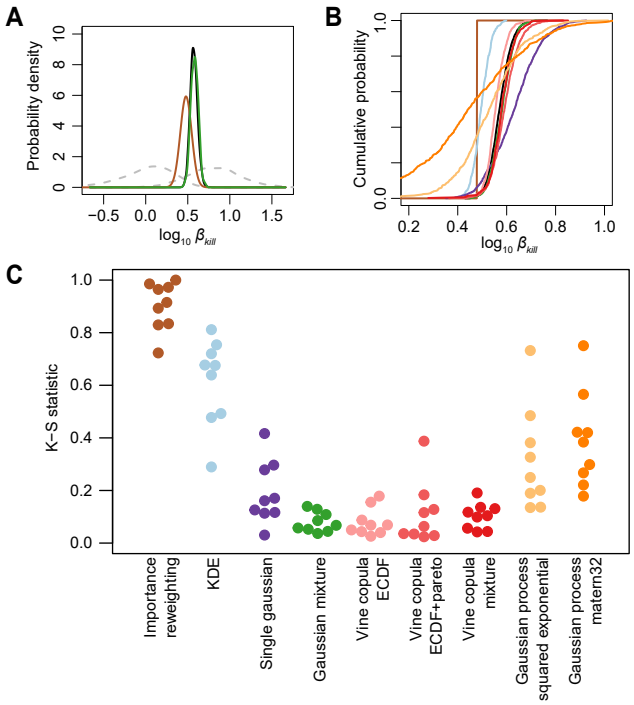


Figure 6.6: **Sequential inference performance with the lynx-hare datasets.** (A) Marginal density of one of the parameters (for clarity, only the GM approximation and importance reweighting result is shown). The dashed lines indicate the posterior of the two datasets separately, and the black line is the true joint. Other colors are the same as in C. (B) Empirical cumulative distribution of the same parameter, showing all approximation methods. (C) Kolmogorov-Smirnov statistics for the comparison of the marginal distributions of the true joint to the marginals of the posterior obtained after sequential inference with each of the approximation methods. Each dot indicates one of the parameters.

6.3.5. BOUNDED PRIORS

In practical applications, it is often the case that the prior probability distribution has a bounded domain, due to known constraints in any of the variables of interest. Some of the approximation methods can handle bounded distributions directly. Alternatively, the variables can be transformed to an unbounded domain (see Methods section). To test these options, we take the same predator-prey model, now inferring the parameters on natural scale and with uniform priors, thus resulting in hard bounds on both the prior and the posterior distribution. As before, the prior is chosen such that only the correct oscillation with a period of 10 years is obtained.

Figure 6.7A-C shows several aspects of the posterior distribution of the lynx data, as before in the log-transformed setting. It is clear that the bounds on the prior distribution leads to a large discontinuity in the posterior probability distribution at this bound for most parameters. The sequential inference test (Figure 6.7D) shows that for KDEs and GMs, it is beneficial to specifically handle these boundaries; either by variable transformation or using truncated Gaussians in the case of Gaussian mixtures. For vine copulas, the marginal transformations can handle bounded domains, but the performance is nevertheless worse than in the unbounded situation.

6.3.6. EFFICIENCY OF SEQUENTIAL VERSUS JOINT INFERENCE

One of the motivations for using posterior approximations and sequential inference is that it may allow a computationally faster evaluation of the joint posterior. For evaluating the posterior of a first dataset, the likelihood of the second dataset does not need to be evaluated and vice versa. More importantly, some of the parameters may only be relevant for one of the datasets and could thus be dropped from the inference, thereby reducing the dimensionality of the inference.

To test this, we return to the unbounded Lotka-Volterra system. If we are primarily interested in the five kinetic parameters, we can treat the initial conditions as nuisance parameters. We can then perform the inference in two steps with seven parameters each and one model evaluation per likelihood calculation; and compare it to joint inference with nine parameters and two model evaluations per likelihood calculation. For all inferences we use MCMC sampling with automated parameter blocking, with identical algorithm configuration, although the actual number of model evaluations differs as a result of the different dimensionality and different parameter blocks being chosen. We use Gaussian mixtures as posterior approximation method.

As shown in Table 6.2, the sequential inference is indeed faster than the joint inference, although this comes at the cost of precision. The reduced dimensionality in the separate sequential inference steps results in more efficient MCMC moves (and hence a decrease in autocorrelation and an increase in the effective sample size). The order of the sequential inference has some influence; including the hare data first is approximately 15% more efficient, and has a lower error, compared to including the lynx data first. We suspect this is the result of the lynx posterior being closer to the prior; and it would be more efficient to first take the larger step from prior to hare posterior and then refine the distribution based on the lynx data, than the other way around.

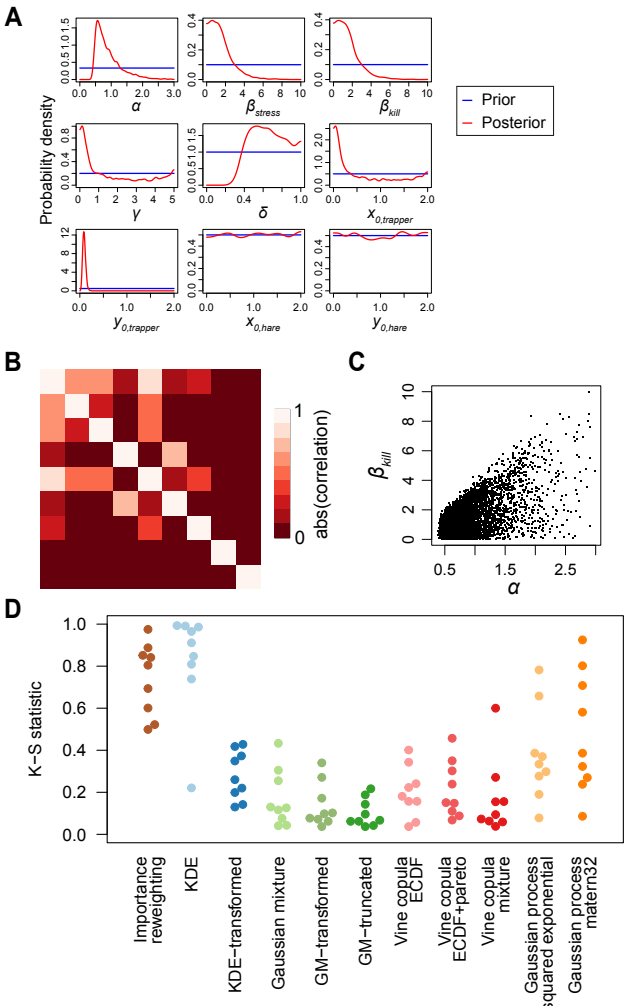


Figure 6.7: **Sequential inference with bounded priors.** (A) Marginal posterior densities after seeing the lynx data; compare with Figure 6.5A. (B) Correlations between the parameters in the lynx posterior. (C) Scatter plot of the samples for one parameter combination. (D) Sequential inference accuracy; same as in Figure 6.6, with the addition of transformed and truncated variations.

Inference	Model evaluations	Minimum ESS	ESS / 1,000 model evals	Maximum D
Lynx - hare	775,002	387	0.50	0.180
Hare - lynx	675,002	391	0.58	0.144
Joint	900,002	165	0.18	0.050

Table 6.2: **Sampling efficiency in sequential inference.** Sampling efficiency is judged by the minimum effective sample size (ESS), and the maximal Kolmogorov-Smirnov statistic (D), in any of the five kinetic parameters. For the calculation of D, the three inferences are compared to the joint inference with 20x more samples. The ESS in the sequential inference is the ESS from the second inference.

Method	Training	Evaluation
Kernel density estimate	N^2D	ND
Gaussian mixture	G^2ND^3	GD^2
Vine copula - ecdf	$N^2D + ND^2$	$ND + D^2$
Vine copula - mixture	$G^2ND + ND^2$	$GD + D^2$
Gaussian process	$N^3 + D$	ND

Table 6.3: **Time complexity of training and evaluation of the approximation methods.** Evaluation is the cost of evaluating one new sample. N = number of Monte Carlo samples used for the estimation, D = dimensionality, G = number of mixture components.

6.3.7. TIME COMPLEXITY

The approximation methods differ in the computational cost of training and evaluation. Table 6.3 lists the time complexity of each method.

Typically, the number of Monte Carlo samples N will be (much) larger than the dimensionality D. Since Gaussian mixtures and vine copulas with mixture marginals do not depend on the number of samples during evaluation, they can achieve the fastest performance when a large number of evaluations are needed in the sequential inference. Kernel density estimates, Gaussian processes and vine copulas with empirical marginals do depend on the number of samples and can thus be significantly slower when a large number of samples is used.

Gaussian processes have cubic scaling with respect to the number of samples for the training, which severely limits the number of samples that can be used. While there are approximation methods available for GPs with large input sizes [13], the use of GPs for posterior approximation appears to be best suited for low N and D.

6.3.8. FAILURE CASE

To illustrate the present limits of this approach to sequential inference, we also discuss a case where the approximations fail to provide an accurate posterior.

A more challenging test case is given by a model of biological signaling in cancer cells. The goal here is to explain how different breast cancer cell lines respond to kinase inhibitors by modeling how the signal arising from oncogenic driver mutations is propagated through a signaling network. These models are described in more detail in [24] and [25]. Here we will use a small test model using the model framework of [25]. The model is shown graphically in Figure 6.8A and the equations are given below. Briefly, the model contains four observed variables, namely the ERBB2 amplification status, PIK3CA mutation status and phosphorylation of AKT and PRAS40 (represented by **m**, **n**, **p** and **q** respectively). The amplification and mutation status is known with certainty, so the variables are directly set to 1 if the amplification or mutation is present and 0 otherwise. The remaining three variables, PI3K activation, AKT activation and PRAS activation (represented by **x**, **y** and **z** respectively) are latent variables, and the inhibitor concentration w is given.

The model is described by the equations

$$\begin{aligned}
\mathbf{x} &= f(b_1 + a_1\mathbf{m} + a_2\mathbf{n}) \cdot g(w) \\
\mathbf{y} &= f(b_2 + a_3\mathbf{x}) \\
\mathbf{z} &= f(b_3 + a_4\mathbf{y}) \\
P(\mathbf{p}|\mathbf{y}) &= t(\mathbf{p}|\mu = \mathbf{y}, \sigma = 0.2, \nu = 3) \\
P(\mathbf{q}|\mathbf{z}) &= t(\mathbf{q}|\mu = \mathbf{z}, \sigma = 0.2, \nu = 3),
\end{aligned}$$

where

$$\begin{aligned}
f(\mathbf{x}) &= 1.0/(1.0 + \exp(-9.19024(\mathbf{x} - 0.5))) \\
g(w) &= k + (1-k)/(10^{s(w-h)} + 1)
\end{aligned}$$

and t is Student's t-distribution with fixed $\nu = 3$ and $\sigma = 0.2$. The remaining 10 variables are scalar parameters to be inferred.

To test whether the sequential inference gives a good approximation also in this setting, we study sequential inference by incorporating parts of a dataset sequentially. The dataset contains measurements of protein phosphorylation without drug treatment (referred to as the pre-treatment data), as well as after 30 minutes of drug treatment (referred to as the on-treatment data), in eight cell lines (see Figure 6.8B). The drug concentration w is 0 in the pre-treatment setting and 1 μM in the on-treatment setting.

We first test sequential inference in the same way as for the lynx-hare model, by splitting the data by observable. That is, we first infer the posterior with observations of p, and subsequently update the posterior with observations of q. As can be seen in Figure 6.8C, sequential inference performs well in this case. The observations of q are correlated with p, and so the first posterior is only slightly refined by the further inclusion of q (in most dimensions).

A potentially more useful sequential inference would be to split the data by pre-treatment and on-treatment data; that is, use the observations of both p and q first for $w = 0$ and then for $w = 1$. Such a split would provide a potential speedup as discussed in the section of sequential inference efficiency, as it would allow us to drop the calculation of the drug effect and the corresponding parameters for the first inference. The accuracy of the sequential inference when split in this way is shown in Figure 6.8D. Unfortunately, none of the approximation methods gives posterior distributions that agree with the joint inference. For several parameters the resulting empirical distributions always have a large discrepancy. Figure 6.8E shows this in more detail for one of the parameters. When investigating this poor performance, we found that this is due to the pre- and on-treatment parts of the data inducing widely different posteriors. As shown in Figure 6.8F, the pre- and on-treatment data are essentially contradictory for the parameters b_2 and a_3: the on-treatment data indicates low values for both parameters, whereas the pre-treatment data indicates higher values. The model can still reconcile these data, as the joint inference shows that a high strength a_3 is favored by both datasets. To recover this joint posterior using approximations would require that the approximations are highly accurate in the tails of the posterior of the pre-treatment data. But standard Monte Carlo methods, and by extensions the approximation methods based on them, are typically not well suited for estimating the tails of a distribution, since most samples

will be concentrated in the body of the distribution. Sequential inference with posterior approximations therefore seems to be unsuitable when the separate datasets give rise to strongly divergent posterior distributions.

6.4. DISCUSSION

To use sequential Bayesian inference in combination with Monte Carlo sampling, we are restricted to using samples from a first inference as prior for a second inference. This can be done by directly reweighting the samples, but this is typically inaccurate. By approximating a functional form of the posterior distribution from the Monte Carlo samples, sequential inference can be performed more accurately. We have explored the use of several methods that could produce such an approximation, and we see that such approximations can indeed allow accurate sequential inference.

The approximation methods have different strengths and weaknesses. We find that Gaussian processes are highly efficient in low dimensionality, but they deteriorate in higher dimensions, at least when using isotropic kernels. Both Gaussian mixtures and vine copulas can give good approximations also in higher dimensions. Vine copulas do not work well for multimodal distributions however. Kernel density estimation appears to be less efficient than the other methods, in a multivariate setting. Finally, none of the approximation methods are adequate in the far tails, although this is to be expected.

Further extensions to the posterior approximation methods can be considered. Using mixtures of t-distributions could improve upon Gaussian mixtures [26, 27]. For vine copulas, the approximation of the marginal distributions can have a strong effect on the accuracy. Further improvements for marginals using Pareto tails could be achieved by estimating an optimal Pareto tail threshold instead of using a fixed value, and estimating the bulk and tail distributions together [28, 29]. Given the good performance of Gaussian process regression in lower dimensions, it will be interesting to explore how this can be better extended to higher dimensions. Using anisotropic kernels will likely be beneficial, but this introduces additional parameters that need to be optimized. To make this computationally feasible it will be necessary to use approximations to the GP, see e.g. [13, 30]. For kernel density estimates, sparse covariance matrices merits exploration as well, such as the method proposed by Liu *et al.* [31]. Finally, it would be interesting to explore combinations of any of the methods, for example by using a Gaussian mixture as prior mean function for Gaussian process regression.

There can be various reasons to use sequential inference. It can be conceptually appealing: all information relevant for the model is stored in the posterior distribution, allowing us to discard a dataset after the inference. Additionally, sequential inference allows us to update an existing model when additional data or samples become available, even when the initial data is no longer available. This can also be useful when an inference task was computationally demanding, and it would be impractical to redo a joint inference when additional data becomes available. Sequential inference can also allow faster inference when a subset of variables is not relevant for one of the datasets. Sequential inference using posterior approximations is an approximation to the joint inference however, and it will depend on the application whether the trade-off between speed and accuracy is reasonable. This approach may be most useful when a significant dimensionality reduction can be obtained, or when the calculation of the model can be

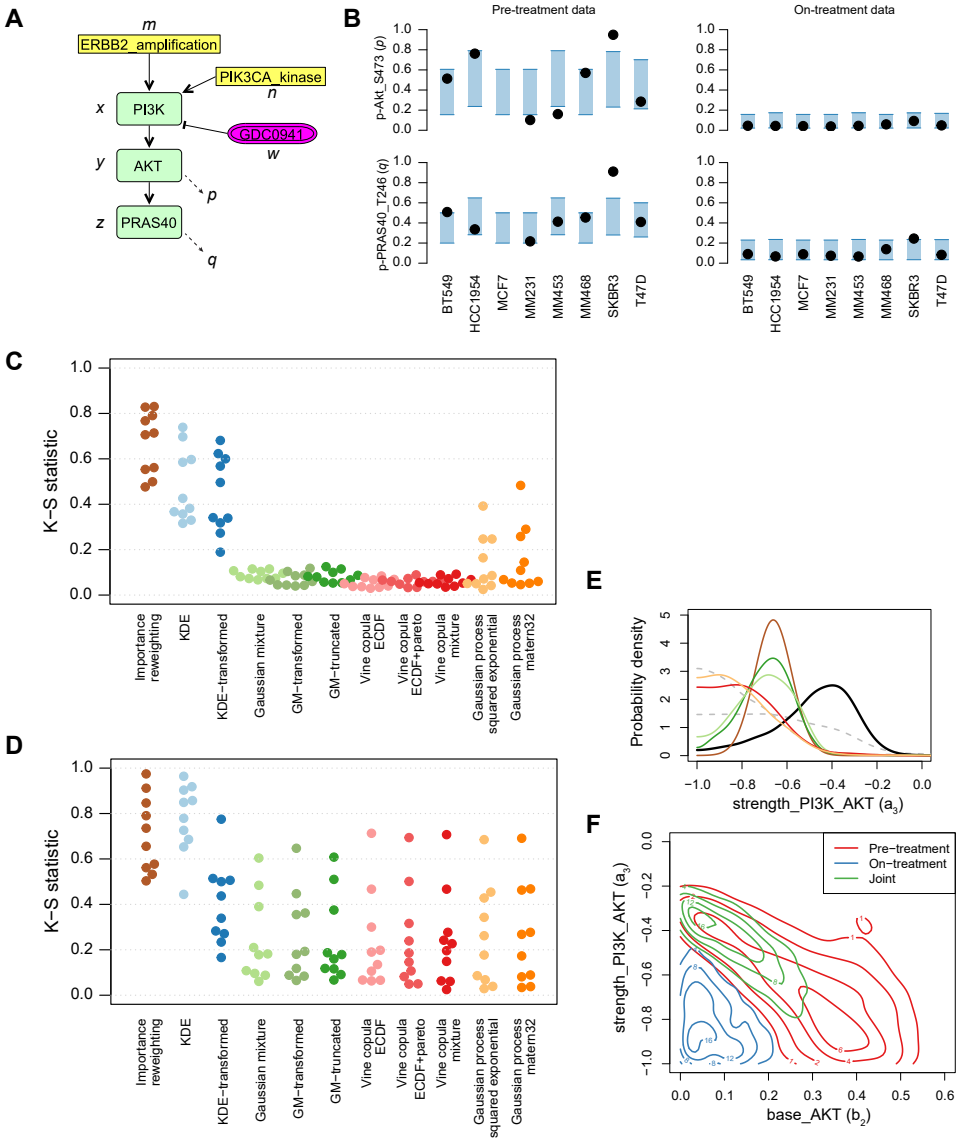


Figure 6.8: **Sequential inference in the breast cancer signaling model.** (A) Signaling model in Systems Biology Graphical Notation format. (B) Data and posterior predictive distributions. Black dots indicate the data and the blue shaded area is the 90% confidence interval of the predictive mean. "p-Akt_S473" is the measurement of p and "p-PRAS40_T246" is the measurement of q. (C) Performance in sequential inference when the data is split by first using the measurement of p and then q (i.e. first use the data shown in the top two graphs in (B), and then the bottom two). (D) Performance in sequential inference when the data is split by pre-treatment and on-treatment (i.e. first use the data shown in the left two graphs shown in (B), and then the right two). (E) Density of one of the parameters as inferred by joint inference (black line) and through sequential approximation split by treatment (colored lines). (F) Contour plot of the bivariate posterior density of two of the parameters obtained from either dataset alone or the true joint.

simplified when considering only part of the data.

APPENDIX

A INTEGRAL OF GAUSSIAN PROCESS PREDICTIVE DISTRIBUTION

In order to normalize the Gaussian process predictive distribution such that it integrates to 1, it is necessary to calculate the integral:

$$Z = \int_{-\infty}^{\infty} K(\mathbf{x}^*, X)K(X, X)^{-1}\mathbf{p}d\mathbf{x}^*.$$

Solving $K(X, X)^{-1}\mathbf{p} = \boldsymbol{\alpha}$ (using e.g. Cholesky decomposition), we have

$$Z = \int_{-\infty}^{\infty} K(\mathbf{x}^*, X)\boldsymbol{\alpha}d\mathbf{x}^*.$$

Both $K(\mathbf{x}^*, X)$ and $\boldsymbol{\alpha}$ are vectors, and we can expand the dot product between them to get

$$Z = \int_{-\infty}^{\infty} \sum_{i=1}^{N} k(\mathbf{x}^*, \mathbf{x_i})\alpha_i d\mathbf{x}^*.$$

Since $\int (f(x) + g(x))dx = \int f(x)dx + \int g(x)dx$, and $\boldsymbol{\alpha}$ is independent of $\mathbf{x}^*$:

$$Z = \sum_{i=1}^{N} \alpha_i \int_{-\infty}^{\infty} k(\mathbf{x}^*, \mathbf{x_i})d\mathbf{x}^*.$$

In the case of the squared exponential kernel $k(\mathbf{x}^*, \mathbf{x}) = \exp(-\frac{|\mathbf{x}^* - \mathbf{x}|}{2l^2})$, we have

$$\int_{-\infty}^{\infty} k(\mathbf{x}^*, \mathbf{x})d\mathbf{x}^* = (\sqrt{2\pi l^2})^D$$

and

$$Z = (\sqrt{2\pi l^2})^D \sum_{i=1}^{N} \alpha_i.$$

For any isotropic kernel $k(\mathbf{x}^*, \mathbf{x}) = h(|\mathbf{x}^* - \mathbf{x}|)$ we can transform to polar coordinates to get

$$\int_{-\infty}^{\infty} h(|\mathbf{x}^* - \mathbf{x}|)d\mathbf{x}^* = \omega_{D-1} \int_0^{\infty} h(r)r^{D-1}dr,$$

where $r = |\mathbf{x}^* - \mathbf{x}|$ and ω_{D-1} is the surface area of a $(D-1)$-sphere with unit radius, which can be calculated as

$$\omega_{D-1} = \frac{2\pi^{D/2}}{\Gamma(\frac{D}{2})}.$$

For the Matérn kernel with $\nu = 3/2$ this gives

$$\int_{-\infty}^{\infty} h(|\mathbf{x}^* - \mathbf{x}|)d\mathbf{x}^* = \frac{2\pi^{D/2}}{\Gamma(\frac{D}{2})}(\frac{l}{\sqrt{3}})^D(1 + D)\Gamma(D).$$

REFERENCES

[1] D. Schmidl, C. Czado, S. Hug, and F. J. Theis, *A vine-copula based adaptive MCMC sampler for efficient inference of dynamical systems,* Bayesian Analysis **8**, 1 (2013).

[2] R. Adams, I. Murray, and D. MacKay, *The Gaussian Process Density Sampler,* Advances in Neural Information Processing Systems **21**, 9 (2008).

[3] W. Neiswanger, C. Wang, and E. Xing, *Asymptotically Exact, Embarrassingly Parallel MCMC,* UAI'14 Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence , 623 (2014).

[4] S. Särkkä, *Bayesian Filtering and Smoothing* (Cambridge University Press, 2013).

[5] M. Wand and M. Jones, *Multivariate plug-in bandwidth selection,* Computational Statistics , 97 (1994).

[6] S. J. Sheather and M. C. Jones, *A reliable data-based bandwidth selection method for kernel density estimation,* Journal of the Royal Statistical Society Series B (Statistical Methodology) **53**, 683 (1991).

[7] D. W. Scott, *Kernel Density Estimation,* in *Multivariate Density Estimation, second edition* (John Wiley & Sons, Hoboken, New Jersey, 2015) pp. 137–213.

[8] U. Schepsmeier, J. Stoeber, E. C. Brechmann, B. Graeler, T. Nagler, and T. Erhardt, *VineCopula: Statistical Inference of Vine Copulas* (2017), r package version 2.1.2.

[9] T. Bedford and R. M. Cooke, *Probability density decomposition for conditionally dependent random variables modeled by vines,* Annals of Mathematics and Artificial Intelligence **32**, 245 (2001).

[10] J. Dißmann, E. C. Brechmann, C. Czado, and D. Kurowicka, *Selecting and estimating regular vine copulae and application to financial returns,* Computational Statistics and Data Analysis **59**, 52 (2013).

[11] C. Scarrott and A. MacDonald, *A review of Extreme Value Threshold Estimation and Uncertainty Quantification,* REVSTAT – Statistical Journal **10**, 33 (2012).

[12] W. H. DuMouchel, *Estimating the Stable Index α in order to Measure Tail Thickness: A Critique,* The Annals of Statistics **11**, 1019 (1983).

[13] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, 2006).

[14] R. D. Wilkinson, *Accelerating ABC methods using Gaussian processes,* Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics **33**, 1015 (2014).

[15] C. J. Geyer, *Markov Chain Monte Carlo Maximum Likelihood,* in *Proceedings of the 23rd Symposium Interface*, 1 (1991) pp. 156–163.

[16] D. Turek, P. de Valpine, C. J. Paciorek, and C. Anderson-Bergman, *Automated Parameter Blocking for Efficient Markov-Chain Monte Carlo Sampling,* Bayesian Analysis **12**, 465 (2017), 1503.05621 .

[17] P. Del Moral, A. Doucet, and A. Jasra, *Sequential Monte Carlo samplers,* Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**, 411 (2006).

[18] J. Skilling, *Nested sampling for general Bayesian computation,* Bayesian Analysis **1**, 833 (2006).

[19] B. Thijssen, T. M. H. Dijkstra, T. Heskes, and L. F. A. Wessels, *BCM: toolkit for Bayesian analysis*

of Computational Models using samplers, BMC Systems Biology **10**, 100 (2016).

[20] C. J. Krebs, R. Boonstra, S. Boutin, and A. Sinclair, *What drives the 10-year cycle of snowshoe hares?* BioScience **51**, 25 (2001).

[21] M. J. Sheriff, C. J. Krebs, and R. Boonstra, *The sensitive hare: Sublethal effects of predator stress on reproduction in snowshoe hares,* Journal of Animal Ecology **78**, 1249 (2009).

[22] C. Elton, M. Nicholson, B. Y. C. Elton, and M. Nicholson, *The ten-year cycle in numbers of the lynx in Canada,* Journal of Animal Ecology **11**, 215 (1942).

[23] J. R. Cary and L. B. Keith, *Reproductive change in the 10-year cycle of snowshoe hares,* Canadian Journal of Zoology **57**, 375 (1979).

[24] K. Jastrzebski, B. Thijssen, R. J. C. Kluin, K. de Lint, I. J. Majewski, R. L. Beijersbergen, and L. F. A. Wessels, *Integrative modeling identifies key determinants of inhibitor sensitivity in breast cancer cell lines,* Cancer Research (2008).

[25] B. Thijssen, K. Jastrzebski, R. L. Beijersbergen, and L. F. A. Wessels, *Delineating feedback activity in the MAPK and AKT pathways using feedback-enabled Inference of Signaling Activity,* bioRxiv (2018).

[26] F. Greselin and S. Ingrassia, *Constrained monotone EM algorithms for mixtures of multivariate t distributions,* Statistics and Computing **20**, 9 (2010).

[27] K. Lo and R. Gottardo, *Flexible mixture modeling via the multivariate t distribution with the Box-Cox transformation: An alternative to the skew-t distribution,* Statistics and Computing **22**, 33 (2012).

[28] A. Tancredi, C. Anderson, and A. O'Hagan, *Accounting for threshold uncertainty in extreme value estimation,* Extremes **9**, 87 (2006).

[29] A. MacDonald, C. J. Scarrott, D. Lee, B. Darlow, M. Reale, and G. Russell, *A flexible extreme value mixture model,* Computational Statistics and Data Analysis **55**, 2137 (2011).

[30] K. Chalupka, C. K. I. Williams, and I. Murray, *A Framework for Evaluating Approximation Methods for Gaussian Process Regression,* Journal of Machine Learning Research **14**, 333 (2013).

[31] H. Liu, J. Lafferty, and L. Wasserman, *Sparse Nonparametric Density Estimation in High Dimensions Using the Rodeo,* Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-07) **2**, 283 (2007).

7

Discussion

7.1. EXPLAINING VARIABILITY IN DRUG RESPONSE

IN order to provide the optimal treatment for cancer patients, it is crucial to understand why each patient responds differently to a given treatment. In this dissertation, we approached this topic by studying kinase inhibitor sensitivity in breast cancer cell lines, using computational modeling of cellular signaling networks. If such models can be used to describe the variability in drug response in cell lines, they should provide a useful starting point for creating predictive models of patient response in the future. Along the way, models can also help us to consolidate our knowledge of drug response and understand what the important contributors to drug sensitivity are.

To address this, we analyzed the response of thirty different breast cancer cell lines to seven different kinase inhibitors. Using relative viability after three days of treatment as a measure of response, we found that a steady-state, knowledge-based model can indeed describe a large part of the variability in response (see Chapter 4). For example, for the EGFR/HER2 inhibitor lapatinib, the model can largely describe the sensitivity or resistance of all cell lines, with only minor differences between the model fit and the observed responses.

Drug response is a complex process, with many different mutations and mechanisms influencing whether a cancer cell is sensitive or resistant to a particular drug. Computational modeling allows us to investigate how these different processes work together in each cell line, and to investigate the contribution of each oncogenic mutation and resistance mechanism. For example, the model indicates that mutations in *PIK3CA* cause resistance to the EGFR/HER2 inhibitor lapatinib, but confer sensitivity to mTOR inhibitor AZD8055. Amplification of *MYC*, on the other hand, confers resistance to mTOR inhibitors. As these factors were already known, they provided a first validation of the computational model.

While the variability in response to lapatinib was described well, the model was at first not able to explain the variability in response to mTOR inhibitors to the same extent. Using additional data-driven analysis, we found that amplification of *EIF4EBP1* was associated with mTOR inhibitor sensitivity. Adding this mechanism to the model greatly increased the goodness of fit, and the model emphasized the importance of 4E-BP1 expression levels in explaining mTOR inhibitor response. Subsequent experimental testing indeed showed that ectopic overexpression of 4EBP1 increased sensitivity to mTOR inhibitors.

Nevertheless, despite all additional data analysis and model revisions, not all responses are explained equally well by our model. There are still several cell lines for which the response to mTOR inhibitors is not explained precisely, while for the PI3K inhibitor GDC0941, and the VEGFR/MET/FGFR2 inhibitor foretinib, there are also several cell lines for which the model cannot accurately explain their observed responses. For example, the cell line HCC38 is significantly more resistant to GDC0941 than what would be expected according to the model (Chapter 4). We were unable to find additional leads in our data which might explain these discrepancies. However, given that we could not yet include all known biology in our models, it is possible that a key mechanism or cellular process has not been incorporated in the model. For example, a detailed description of PI3K subunits was not included, which may be important to accurately describe response to PI3K inhibitors [1, 2]. Other mechanisms such as signaling through RSK3 and

RSK4 [3] may also be involved. Alternatively, the failure to recapitulate the observed response exactly may be a limitation of the steady-state description, as time-dependent effects may also be important [4].

These additional hypotheses regarding factors influencing PI3K inhibitor response highlight a central challenge in the construction of models of cellular signaling: it is not trivial to decide which aspects of cell biology should be included in the model, and at what level of detail. The computational cost of parameter inference poses constraints on the size and complexity of the models which can be considered (discussed further in the section on computational efficiency). This prevents us from simply including all possible molecular interactions which have been described in the literature. We have addressed this challenge by using an iterative procedure to construct our models. In each case (Chapters 3, 4 and 5), we started with a small, simple model, and tested how well this model can describe the available data. Using posterior predictive distributions, we can test in detail which data can and cannot be described by the model at hand. We then searched the literature, and used data-driven analyses, to suggest additional mechanisms which should be incorporated in the model. The models can subsequently be extended with these mechanisms and re-fitted to the data. Although this process can be time-consuming, it provides us with sparse models of the biological system of interest.

In Chapter 5, we extended the modeling framework of steady-state cellular signaling to include feedback signaling. This allowed us to delineate the strengths for several feedback mechanisms in the MAPK and AKT pathways, taking into account four different datasets. This analysis indicated that, even though the hyperphosphorylation of AKT upon AKT inhibitor treatment is likely to be an unintended effect of the inhibitor, a feedback through IRS1 is also still likely to be active. The analysis of uncertainty further showed that we cannot entirely resolve this feedback with the available data, as the S6K inhibitor had weak and uncertain efficacy. This shows the importance of analyzing uncertainty in parameter estimates. As before, these models again highlighted several data points which could not be explained, such as the increased S6 phosphorylation upon treatment with the IGF1R inhibitor NT-157 in the HCC1954 cell line. This provides a basis for additional rounds of model adaptation and specific measurements to arrive at increasingly detailed models of how cells respond to kinase inhibitors.

With these results, we can now continue the process of iterative model development to obtain increasingly accurate descriptions of drug response, although this will require methodological advances to allow larger models to be considered (discussed later). We have seen that combining measurements of relative viability after kinase inhibitor treatment with data on mutations, copy number aberrations, mRNA and (phospho)protein levels provides useful insights on how cancer cells are sensitive or resistant to these kinase inhibitors. Independently, measurements of protein phosphorylation after inhibitor treatment also provided information on the likely signaling flows in cancer cells. It will be particularly interesting to combine the on-treatment phosphorylation measurements with the dose-response relative viability data to further refine the models of inhibitor response. It has not yet been feasible to address this combination of data due to the many model evaluations required to simulate all of the datasets together, as well as a compounded increase in the number of parameters.

Apart from incorporating additional details of the signaling pathways and further integration of datasets, various other extensions will be interesting to consider as well, discussed in more detail in the section 'Extending scope and detail'. Two other important next steps are to further test the predictive performance of such models (discussed below), and to start translating the models from cell lines to patients (discussed in the section 'Models of patient response').

7.2. PREDICTIVE MODELS

Now that our computational models of response to various kinase inhibitors are approaching an accurate description of the available data, we can start to explore their predictive value. There are two types of predictions which can be made: we can either use the models to predict the effect of new treatments on the same cell lines, or to predict the effect of the same treatments on new cell lines.

Since there is only a limited number of cell lines available, and it is furthermore difficult and time-consuming to establish new lines, creating an entirely new cell line panel for validation would be difficult. Establishing organoid cultures may have higher success rates than establishing new cell lines [5], nevertheless the subsequent culturing and drug response screens are more laborious and expensive in the organoid setting. Given this, we can estimate the predictive performance using cross-validation instead. Preliminary tests indicated fairly good performance, however, the computational cost of cross-validation within a Bayesian framework is extremely high and we cannot yet confidently give estimates of predictive performance. Leave-one-out cross-validation in a panel of thirty cell lines takes approximately thirty times as long as the single inference. Given a typical inference time of 24 hours, a cross-validation run requires a month of computation time. Although this is feasible for a single model with one drug, we would need to test multiple models, and multiple kinase inhibitors. At present it is therefore only practical to comprehensively test the performance of predicting the effects of the same treatments on new cell lines for small models, although computational and technological advances may alleviate this limitation (discussed later).

Given the difficulties with predicting sensitivity for new cell lines, it may be more feasible to test the predictive value of the model by predicting new treatments on the same cell lines. The present models can in principle be used directly to predict the effect of new treatments on the same set of cell lines. Most interestingly, new treatments could also encompass combinations of inhibitors. Given the vast amount of possible combinations of drugs, it would be a major challenge to test all these combinations experimentally, especially when different concentrations are also considered. A computational model, however, could quickly evaluate the effectiveness of many different inhibitor combinations, which can be used to generate a list of promising combinations for further experimental testing.

We have begun investigating combination treatments by sequentially treating cells with an IGF1R inhibitor and an AKT inhibitor, discussed in Chapter 5. The goal here was to constrain model parameters, but we can also use the model to predict what combination of inhibitors would disrupt the proliferation of one cell line, but not another. To make this most relevant for the clinical setting, we would like to select combinations of inhibitors that selectively kill cancer cells, while leaving normal cells unaffected [6]. A

challenge with this is the difficulty in profiling the response of normal cells, since normal breast epithelial cells can only be cultured in vitro for a limited period of time. These cells can be transformed such that they can be cultured for much longer, as has been done with the cell line MCF-10A [7], although these cells no longer precisely represent normal breast epithelial cells. In addition, it appears that these transformed, normal-like cell lines are generally much more sensitive to most treatments than cancer cell lines, even though we know that in patients there is a therapeutic window where the treatment disrupts the cancer with manageable side-effects on other tissues. For now, we are therefore restricted to finding new inhibitors or combinations of inhibitors that kill specific cancer cell lines without affecting other cancer cell lines. Such selective combinations may also leave normal human cells unaffected.

In conclusion, while we have developed models that are capable of reproducing the observed data to a large extent, there is still a significant amount of work to be done to further validate these models, especially in a clinical setting.

7.3. EXTENDING SCOPE AND DETAIL

In the first section, I mentioned several mechanisms which could be important for explaining response to PI3K inhibitors, including differences between PI3K subunit classes, signaling by two RSK isoforms and time-dependent effects. Indeed, for each of the drugs various additional mechanisms of sensitivity or resistance have been described. For example, for lapatinib, signaling by Src-family kinases [8] and by the receptor tyrosine kinase AXL [9] have been reported to influence drug sensitivity. It would be interesting to explore the contributions of all of these signaling molecules and drug sensitivity mechanisms in the larger context of signaling pathways using our computational models.

Apart from such additional details of signaling directly downstream of growth factor receptors, it will be important to consider the signaling and regulation in entirely different processes as well. The regulation of cell death, senescence and cellular differentiation are important processes that can determine whether cells are killed by a particular drug, or merely temporarily arrested in growth. Given the type of drug response data which was used here (relative viability), it is difficult to model these additional processes in significant detail. However, technological developments in biological assays and measurement devices have made it feasible to perform measurements of apoptosis or more general cell death in a relatively high-throughput fashion. An example of this is automated systems for live cell imaging, such as the IncuCyte platform. In addition to measuring confluence, fluorescent markers can be used to not only count individual cells, but also identify apoptotic cells can, thus providing the additional information needed to model additional cellular processes.

The importance of apoptosis is highlighted by a computational model of signaling by BCL-2 family members [10]. It was found that this model can be predictive of response to chemotherapy in colorectal cancer. As with kinase inhibitors, many other factors affecting response to chemotherapy have been described, and it will be interesting to merge such models of apoptosis regulation and chemotherapy response with models of growth-factor signaling pathways to elucidate the relative importance of each of these mechanisms.

In addition to modeling pathways in more detail and extending the scope with addi-

tional biological processes, it will be important to explore a more detailed mathematical description of the cellular processes. More specifically, the steady state description used in Chapters 4 and 5 is not able to handle processes occurring on different timescales. This limitation precludes inclusion of both fast post-translational signaling and slower transcriptional regulation in the same model. We focused on post-translational signaling, but such models may be poor at describing long-term drug response. For instance, the important negative transcriptional feedback of AKT signaling to RTK expression [11, 12] could not be included in our models. Accommodating such mechanisms would require the calculation of steady state levels with feedback on multiple time scales. Alternatively, it may be beneficial to switch to dynamic models in such cases.

It is unclear which of these aspects – that is, details of signaling pathways, additional cellular processes or time scales – are most important to consider. Classically, colony formation assays are seen as the gold standard for determining the clinical potential of drug treatment using cell lines. These assays determine whether individual cells can grow out into colonies of at least 50 cells after drug treatment [13]. Although this assay has limitations and is still an artificial situation, it may be a reasonable reflection of what occurs in vivo [14]. Nevertheless, it still only measures the net effect of drug treatment, and cannot dissect the contributions of the diverse cell biological processes affecting response. In line with this dissertation, the best way to approach this may be to construct increasingly detailed mechanistic computational models with each of the possible extensions. Once these models can describe drug response in a number of situations, and when they are found to be predictive, we can start to gain confidence that they describe reality, and can be used to dissect the various contributors to drug response.

7.4. MODELS OF PATIENT RESPONSE

Apart from obtaining a more complete understanding of kinase inhibitor response, the ultimate goal of these computational models is to use them to guide treatment decisions for cancer patients. Before such models can be used for this purpose, they first need to be translated from describing cell line response to describing patient response. Subsequently, the predictive power of the model needs to be tested and validated. Testing the predictive performance in this case can first be done by dividing a cohort into a training and a test set. Given good performance in the test set, validation in an independent cohort will also be necessary to test the generalizability of the model.

In principle, the present model can be translated to describing patient response in a simple way, by introducing a logistic function linking relative viability to a binary response variable such as pathological complete response after neoadjuvant therapy. This would assume that patient response is determined only by the signaling pathways included in the model, and is an strong simplification. However, it does reflect the reasoning that is often taken when considering the efficacy of targeted inhibitors.

Alternative ways of translating models to patients can be envisioned as well, including simple evolutionary dynamics models of the tumor, where growth and death rates observed in cell lines can be used as prior information. Given a matching dataset of cell lines and patients, including molecular profiling and drug response in both systems, we can again use an iterative model development procedure to find a good description of patient response. We can also use the approximation methods described in Chapter 6

to leverage the posterior probability distributions obtained from cell line data as prior information for models of patient response. In this way, knowledge of kinase inhibitor response in cell lines can be used for constraining the parameters of patient response, a step which may be necessary for smaller patient cohorts that provide only limited information.

If this approach delivers computational models that can retrospectively predict patient response to treatment in independent datasets, a prospective trial would be needed to confirm the predictive value. In such a trial, molecular profiling of a pre-treatment biopsy would provide the input for the computational model, which could then be used to simulate the effect of one or more treatments. If the model also passes this test, it could then be used to guide decisions on which treatment is given. The Bayesian framework of the computational models would also make it feasible to develop an adaptive trial, comparable to the approach used in the I-SPY2 trial [15, 16], where cohorts are expanded or shut down based on observed responses. This ensures that drug candidates, or in our case treatment decision protocols, could be graduated into clinical practice as quickly as possible.

7.5. ENABLING EXTENDED SIGNALING MODELS

One of the challenges with the modeling approach used in this thesis is the computational cost of the inference. To obtain robust models, it is necessary to test various different types of models, including not only variations in the network topology, but also in the mathematical specification, such as different types of activation functions, error models or data transformations. Given this requirement of testing numerous different model versions, each individual model should be computable in a manageable amount of time. We generally restricted the size of the computational model such that the inference required at most 24 hours to converge. With our inference software package (Chapter 2) and simulation methods (Chapters 4 and 5), we could consider models with up to 20-30 signaling nodes and 50-150 unknown parameters, depending on whether feedback is included and how many model conditions have to be simulated for the datasets used in the inference.

Such models of medium size provided useful insights, but significantly more biological knowledge is available, and additional mechanistic detail may be needed to fully describe drug response, as described earlier. To enable further extensions of the models, improvements in both parameter inference and model simulation should be considered. Improvements in sampling algorithms will be discussed in more detail in the section on Bayesian computation below. There are however opportunities for improvement of the model simulation methods as well.

For the models in Chapter 4, which excluded feedback, we used BCM's code generation feature. This feature parses the model specification, and generates C++ code that can be compiled to executable code using optimizing compilers. The models in Chapter 5 however, which included feedback mechanisms, are more complex to simulate due to the heuristic Newton-Raphson iteration. In addition, these models were specified in the more easily reusable SBML format. Although the simulation code for these models was also written in C++, the implementation does not generate model-specific code. Code generation for the feedback models could improve their simulation efficiency, as it

would eliminate various loops and model traversing during simulation, as well as leverage automatic compiler optimization.

Another optimization which should be considered is the use of vectorization. For the models in Chapter 4, the compiler may be able to introduce some vectorization based on BCM-generated code, however, the extent of vectorization is limited, and it also currently does not apply for models including feedback. Given that a model needs to be simulated many times, using different parameter values as well as different model conditions, it would be worthwhile to vectorize the code across either of these dimensions. For the feedback models, the benefit may be somewhat limited as the Newton-Raphson iteration may require a different number of iterations in each model calculation. Nevertheless, given the wide registers and dedicated execution units in the latest processors, vectorization may provide up to an order of magnitude faster simulation.

Apart from vectorization, the model simulation can also be further parallelized across compute nodes. Although BCM is multi-threaded, it is restricted to a single compute node. Inference algorithms based on sequential Monte Carlo sampling scale very well to many threads, and although efficient parallelization for MCMC algorithms is not trivial, work in this area is underway [17–19]. As with vectorization, it will be necessary to test whether parallelization could best be done across parameter values or across model conditions. With the availability of national and international computing clusters, inference time could be reduced by up to two orders of magnitude, providing the turn-around time needed for iterative model development, although a large amount of total CPU time would still be required.

At the outset, time investment into these more advanced computing techniques was not justified, but given that our models are now starting to be robust and have provided useful biological results, further investments in improving the efficiency of model simulation are becoming justifiable. In another area of computational science, molecular dynamics simulation, three decades of performance optimization have provided highly efficient simulation software [20, 21], which in turn allows for assembly of increasingly detailed models leading to new biological insights [22]. Similar advances in computational modeling of cellular signaling networks may also lead to an improved understanding of the variability in drug response. Indeed, in the simulation of deterministic ordinary differential equation models, specialized integrators can provide improved efficiencies for simulating models of biological systems [23].

At the root of the computational difficulty is the large amount of unknown parameters which have to be inferred from data. Rather than increasing the computational efficiency of simulation and inference, a more direct way to address this would be to constrain parameters separately using additional measurements. This would allow us to fix those parameters, or at least reduce the size of the prior distribution, thereby reducing the parameter space which has to be searched during an inference. It is not trivial to measure these parameters however, as they are abstract representations of signaling strength which do not correspond to real physical constants that could be measured directly. Nevertheless, we may again derive inspiration from the field of molecular dynamics simulation: in this field the computational models also heavily depend on abstract parameters which cannot be measured directly, but fitting these parameters to diverse datasets has produced increasingly accurate sets of parameter values [24, 25]. Similar

efforts may be necessary to obtain re-usable parameter values in models of cellular signaling networks. If successful, this would in turn allow larger, more detailed models to be considered.

7.6. BAYESIAN COMPUTATION

Throughout this dissertation, we have employed Bayesian statistics to infer probability distributions describing the uncertainty in model parameters. The advantage of this approach is that it characterizes the full, joint uncertainty of all model parameters, and Chapters 5 and 6 have illustrated that this provides additional information on the behavior of the model. For example, the uncertainty of feedback strength between S6K, mTOR and IRS was correlated with uncertainties in the efficacy of the S6K inhibitor (Chapter 5). The Bayesian approach also allows the inclusion of prior information, which we have used to incorporate biochemical measurements of kinase inhibitor affinities. Two main alternatives for the characterization of uncertainty are the use of bootstrapping in the maximum likelihood context, or using profile likelihoods. Profile likelihoods can be said to combine advantages of frequentist and Bayesian formalisms [26]. The step of global optimization may be computationally more efficient than fully traversing the posterior with Monte Carlo sampling, while the profile likelihood still characterizes the entire distribution of single parameters. However, it does not provide the joint uncertainty between parameters. In this dissertation, we persevered with the fully Bayesian treatment, despite the potentially higher computational cost.

To ensure that implementation of the sampling algorithms was not a bottleneck, we developed the BCM software package, which provides efficient, multi-threaded implementation of various different sampling algorithms (Chapter 2). In Chapter 6, we further explored whether the computational efficiency of the inference with multiple datasets can be improved by inferring the posterior distributions sequentially. This requires an approximation of the intermediate probability distributions. We found that Gaussian mixtures provide the best approximation in higher dimension problems, while Gaussian processes can be very accurate in lower dimensional problems. Vine copulas can also provide good approximations, as long as the distribution is unimodal. Sequential inference using these intermediate approximations can be more efficient than a joint inference with multiple datasets, although the intermediate approximations do introduce an additional error in the final posterior distribution. Our work indicated that, when performing inference with multiple datasets, it is important to first perform a separate inference with each dataset separately. The posterior distributions induced by each dataset can then be compared, and if they are found to be highly divergent, adaptations to the model may be necessary before attempting a joint or sequential inference with the datasets together.

The work-horses of Bayesian inference employed in this thesis are the Monte Carlo sampling algorithms. Efficiency of these samplers crucially depends on having a good proposal distribution for generating new candidate values. Adaptive methods attempt to learn a good proposal distribution from the samples generated so far. We used random-walk sampling with the covariance matrix of previous samples as basis for adapting the proposal distribution [27], along with parameter blocking in the case of MCMC sampling [28] and scaling to optimize the acceptance rate [29, 30]. An important limitation of this

approach is that the shape of the proposal distribution is constant across space. In unimodal problems this may be inefficient if the shape of the mode does not correspond to the shape of the proposal distribution, but generally performs well. The situation is more dire in multi-modal problems, as in this case the shape of the proposal distribution can become dominated by the location of the modes, rather than by the shape of each mode separately. This is particularly problematic at intermediate temperatures (in both parallel tempering and sequential Monte Carlo), where there is a transition from the prior mode to the posterior mode. At specific temperatures multi-modal distribution often arise, in turn resulting in highly inefficient sampling at intermediate temperatures. The parallel tempering scheme does overcome this by transitioning into the prior or posterior, but many round-trips are necessary.

A sampling scheme where modes are identified from the previous samples to make mode-specific proposal distributions could alleviate this. Such a scheme has been proposed in the PolyChord sampler [31], in a nested sampling framework. This sampler uses a k-nearest neighbor algorithm to identify clusters, and subsequently uses slice sampling within each cluster. It will be interesting to test this sampler for inference with biological computational models, particularly in higher dimensions. Clustering in high-dimensional space is non-trivial, since all points tend to become equidistant as the dimensionality increases, and it may be worthwhile to explore the performance of different clustering algorithms.

Several sampling algorithms can make use of the gradient or the Hessian of the likelihood function to improve sampling efficiency, including Hamiltonian Monte Carlo [32], the Metropolis-adjusted Langevin algorithm [33] and Riemann manifold versions of these two algorithms [34]. Whether this is beneficial depends on the relative cost of evaluating the gradient and Hessian of the likelihood function, compared to the gain in sampling efficiency. In the comparable task of global optimization, it was found that finite difference approximations of the gradient give unreliable results when they are used in the optimization procedure [35]. In order to benefit from sampling algorithms that can use the gradient information, it therefore seems necessary to evaluate the gradient directly rather than relying on a finite difference gradient approximation. In the Stan software package [36], code for evaluating the gradient is generated automatically from the model specification, and a similar scheme can be implemented for the ISA models. For models containing feedback, an extended system of equations could be solved to evaluate the gradient. In ODE models, adjoint sensitivity analysis can be an efficient way to evaluate the gradient even for large models [37].

An important trade-off in sampling algorithms is the balance between exploration and exploitation. To obtain a complete description of the posterior, the entire parameter space needs to be explored sufficiently to identify all regions of high posterior probability (i.e, the exploration) and subsequently each mode of the distribution should be characterized in detail (the exploitation). Using adaptive samplers, the proposal distribution is adapted based on previous samples [27]. Although this adaptation increases the efficiency of exploitation, it should not be done too quickly to prevent an incomplete exploration of the entire parameter space. However, it is not always clear how the parameters of the algorithm affect this trade-off, which introduces manual trial-and-error in order to optimize the sampling efficiency.

The latter is a more general issue with sampling algorithms; manual tweaking is generally needed to find efficient algorithm settings. This is in part caused by an absence of a universal method that can determine convergence of the sampler [38]. There is progress in automated optimization of the sampling algorithm; for example, the No-U-Turn sampler automatically tunes path lengths for the Hamiltonian Monte Carlo algorithm [39]. Such automation is only a partial remedy however. Even if automatically tuned, the Hamiltonian Monte Carlo sampling may not be suitable for the posterior of interest, as it does not handle multi-modal distributions well. Since we generally do not know the characteristics of the posterior distribution before starting the inference, trial-and-error is still needed to determine whether a sampling algorithm is suitable, even if it automatically tweaks all algorithm parameters.

Taken together, there are still many opportunities for improving the efficiency and practical applicability of sampling algorithms. Such improvements would in turn allow the inference of increasingly detailed and complex models. In biology, furthermore, models tend to have large numbers of unknown parameters, and posterior distributions tend to have complicated shapes [40, 41], which we have also seen throughout this dissertation. Particularly when multiple datasets are included, posterior distributions can be challenging to sample, as illustrated in Chapter 6. As such, computational models of biological systems along with the inclusion of multiple datasets provide challenging cases for parameter inference.

7.7. ALTERNATIVE MODELING APPROACHES

In addition to the models of signaling in relation to drug response already mentioned, including models based on a Boolean framework [42], using modular response analysis [43, 44], or dynamic models [10, 45], other approaches have been taken as well.

Recently, Eduati et al. [46] have reported on combining dynamic models with a discrete logical framework. In this case, fourteen colorectal cancer cell lines were used, and fourteen epitopes weremeasured in 43 conditions. They used their model to suggest potential combination experiments, highlighting several combinations which are already in clinical trials. A novel prediction, namely the combination of a MEK inhibitor with a GSK inhibitor, was also tested experimentally. The data from this experiment was stated to validate the model prediction, but important controls of GSK3 inhibition in the absence of a MEK inhibitor were not included, and no statistical evaluation of sensitivity to the combination was presented.

In a recent pre-print, Fröhlich et al [47] have reported on analyzing drug response with a very large dynamic model of cellular signaling, including 1,228 molecular species and 4,100 parameters. They used adjoint sensitivity analysis along with parallelization to speed up the parameter inference. A crucial difference to our modeling approach however, is that the parameter optimization for each inference, although run for several restarts, included only 100 iterations. This very limited parameter search was said to reduce overfitting [47], but even using gradient evaluations and multiple restarts, it seems unlikely that such few iterations of parameter values can accurately describe a 4,100 dimensional parameter space. As a result, it would be hard to interpret the specific parameter values, while this would be of most interest to understand which processes are important for determining drug response. It is also unclear how computationally demand-

ing the entire optimization task was (a wall time of both 1 week and 4,000 hours (over 5 months) is mentioned). Nevertheless, they tested the cross-validation performance, and the signaling model did outperform statistical models of drug response in predictive performance. A more detailed investigation will be needed to understand how this large model, despite the apparently coarse characterization of parameters, performs well in predicting drug response. With such additional analyses, these large models could be of great interest for obtaining a broader understanding of how cellular signaling networks affect drug response.

Each of these modeling approaches has advantages and disadvantages. Boolean models can be faster to compute and hence allow larger models to be considered, but it is more difficult to study quantitative differences in this framework. On the other end, dynamic models can provide detailed information on the kinetics of the system, but lead to much higher computational costs. Given the importance of understanding drug response, much insight can be gained from employing different approaches to create computational models of cellular signaling in the context of drug response. Each of these approaches may reveal different aspects of how signaling networks function to regulate response. As in any modeling situation, it will depend on the precise question being addressed which modeling formalism may be most appropriate.

7.8. CONCLUSION

We have shown that integrative modeling of kinase inhibitor response using steady-state, knowledge-based models can be useful in establishing whether current knowledge can explain the variability in drug response. These models are providing a more comprehensive view of how diverse oncogenic driver mutations and drug sensitivity mechanisms affect response to anticancer drugs. In conjunction with advances in computational methods and additional measurements, we can further develop the general approach described in this dissertation to obtain increasingly detailed models of response to treatment in cancer. We can also begin to translate this to a clinical setting by creating models of patient response, followed by testing the predictive performance of these models. We believe that this is a promising approach with potential to make precision medicine for cancer patients a reality.

REFERENCES

[1] S. Wee, D. Wiederschain, S.-M. Maira, A. Loo, C. Miller, *et al.*, *PTEN-deficient cancers depend on PIK3CB.* Proceedings of the National Academy of Sciences of the United States of America **105**, 13057 (2008).

[2] C. Costa, H. Ebi, M. Martini, S. A. Beausoleil, A. C. Faber, *et al.*, *Measurement of PIP3levels reveals an unexpected role for p110β in Early adaptive responses to p110α-Specific inhibitors in luminal breast cancer,* Cancer Cell **27**, 97 (2015).

[3] V. Serra, P. J. Eichhorn, C. García-García, Y. H. Ibrahim, L. Prudkin, *et al.*, *RSK3/4 mediate resistance to PI3K pathway inhibitors in breast cancer,* Journal of Clinical Investigation **123**, 2551 (2013).

[4] M. Will, A. C. R. Qin, W. Toy, Z. Yao, V. Rodrik-Outmezguine, *et al.*, *Rapid induction of apoptosis*

by PI3K inhibitors is dependent upon their transient inhibition of RAS-ERK signaling. Cancer discovery **4**, 334 (2014).

[5] F. Weeber, S. N. Ooft, K. K. Dijkstra, and E. E. Voest, *Tumor Organoids as a Pre-clinical Cancer Model for Drug Discovery,* Cell Chemical Biology **24**, 1092 (2017).

[6] J. E. Dancey and H. X. Chen, *Strategies for optimizing combinations of molecularly targeted anticancer agents,* Nature Reviews Drug Discovery **5**, 649 (2006).

[7] H. D. Soule, T. M. Maloney, S. R. Wolman, E. C. Line, W. D. Peterson, *et al.*, *Isolation and Characterization of a Spontaneously Immortalized Human Breast,* Cancer , 6075 (1990).

[8] B. N. Rexer, a.-J. L. Ham, C. Rinehart, S. Hill, N. de Matos Granja-Ingram, *et al.*, *Phosphoproteomic mass spectrometry profiling links Src family kinases to escape from HER2 tyrosine kinase inhibition,* Oncogene **30**, 4163 (2011).

[9] L. Liu, J. Greger, H. Shi, Y. Liu, J. Greshock, *et al.*, *Novel mechanism of lapatinib resistance in HER2-positive breast tumor cells: Activation of AXL,* Cancer Research **69**, 6871 (2009).

[10] A. U. Lindner, C. G. Concannon, G. J. Boukes, M. D. Cannon, F. Llambi, *et al.*, *Systems analysis of BCL2 protein family interactions establishes a model to predict responses to chemotherapy,* Cancer Research **73**, 519 (2013).

[11] N. V. Sergina, M. Rausch, D. Wang, J. Blair, B. Hann, K. M. Shokat, and M. M. Moasser, *Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3.* Nature **445**, 437 (2007).

[12] S. Chandarlapaty, A. Sawai, M. Scaltriti, V. Rodrik-Outmezguine, O. Grbovic-Huezo, *et al.*, *AKT inhibition relieves feedback suppression of receptor tyrosine kinase expression and activity,* Cancer Cell **19**, 58 (2011).

[13] N. A. P. Franken, H. M. Rodermond, J. Stap, J. Haveman, and C. van Bree, *Clonogenic assay of cells in vitro,* Nature Protocols **1**, 2315 (2006).

[14] J. M. Brown and L. D. Attardi, *The role of apoptosis in cancer development and treatment response.* Nature reviews. Cancer **5**, 231 (2005).

[15] A. D. Barker, C. C. Sigman, G. J. Kelloff, N. M. Hylton, D. a. Berry, and L. J. Esserman, *I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy.* Clinical pharmacology and therapeutics **86**, 97 (2009).

[16] J. W. Park, M. C. Liu, D. Yee, C. Yau, L. J. van 't Veer, *et al.*, *Adaptive Randomization of Neratinib in Early Breast Cancer,* New England Journal of Medicine **375**, 11 (2016).

[17] A. E. Brockwell, *Parallel Markov Chain Monte Carlo simulation by pre-fetching,* Journal of Computational and Graphical Statistics **15**, 246 (2006).

[18] W. Neiswanger, C. Wang, and E. Xing, *Asymptotically Exact, Embarrassingly Parallel MCMC,* UAI'14 Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence , 623 (2014).

[19] R. Nishihara, I. Murray, and R. P. Adams, *Parallel MCMC with Generalized Elliptical Slice Sampling,* Journal of Machine Learning Research **15**, 2087 (2014).

[20] M. Christen, P. H. Hünenberger, D. Bakowies, R. Baron, R. Bürgi, *et al.*, *The GROMOS software for biomolecular simulation: GROMOS05,* Journal of Computational Chemistry **26**, 1719 (2005).

[21] B. Brooks, C. Brooks, A. Mackerell, L. Nilsson, R. Petrella, *et al.*, *CHARMM: The Biomolecular*

Simulation Program, Journal of computational chemistry **30**, 1545 (2009).

[22] W. Toy, Y. Shen, H. Won, B. Green, R. A. Sakr, *et al., ESR1 ligand-binding domain mutations in hormone-resistant breast cancer,* Nature Genetics **45**, 1439 (2013).

[23] P. Gonnet, S. Dimopoulos, L. Widmer, and J. Stelling, *A specialized ODE integrator for the efficient computation of parameter sensitivities,* BMC Systems Biology **6** (2012).

[24] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. Van Gunsteren, *Definition and testing of the GROMOS force-field versions 54A7 and 54B7,* European Biophysics Journal **40**, 843 (2011).

[25] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, *et al., CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields,* Journal of computational chemistry **31**, 671 (2009).

[26] A. Raue, C. Kreutz, F. J. Theis, and J. Timmer, *Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability,* Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **371**, 20110544 (2012).

[27] H. Haario, E. Saksman, and J. Tamminen, *An Adaptive Metropolis Algorithm,* Bernoulli **7**, 223 (2001).

[28] D. Turek, P. de Valpine, C. J. Paciorek, and C. Anderson-Bergman, *Automated Parameter Blocking for Efficient Markov-Chain Monte Carlo Sampling,* Bayesian Analysis **12**, 465 (2017), 1503.05621 .

[29] G. O. Roberts, A. Gelman, and W. R. Gilks, *Weak convergence and optimal scaling of random walk Metropolis algorithms,* Annals of Applied Probability **7**, 110 (1997).

[30] G. O. Roberts and J. S. Rosenthal, *Optimal scaling for various Metropolis-Hastings algorithms,* Statistical Science **16**, 351 (2001).

[31] W. J. Handley, M. P. Hobson, and A. N. Lasenby, *POLYCHORD: Next-generation nested sampling,* Monthly Notices of the Royal Astronomical Society **453**, 4384 (2015).

[32] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, *Hybrid Monte Carlo,* Physics Letters B **55**, 2774 (1987).

[33] G. O. Roberts and O. Stramer, *Langevin Diffusions and Metropolis-Hastings Algorithms,* Methodology and computing in applied probability **4**, 337 (2002).

[34] M. Girolami and B. Calderhead, *Riemann manifold Langevin and Hamiltonian Monte Carlo methods,* Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**, 123 (2011).

[35] A. Raue, M. Schilling, J. Bachmann, A. Matteson, M. Schelke, *et al., Lessons learned from quantitative dynamical modeling in systems biology.* PloS one **8**, e74335 (2013).

[36] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, *et al., Stan: A Probabilistic Programming Language,* Journal of Statistical Software **76** (2017).

[37] F. Fröhlich, B. Kaltenbacher, F. J. Theis, and J. Hasenauer, *Scalable Parameter Estimation for Genome-Scale Biochemical Reaction Networks,* PLoS Computational Biology **13**, 1 (2017).

[38] M. K. Cowles and B. P. Carlin, *Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review,* Journal of the American Statistical Association **91**, 883 (1996).

[39] M. D. Hoffman and A. Gelman, *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo,* Journal of Machine Learning Research **15**, 1593 (2014).

[40] M. Girolami, *Bayesian inference for differential equations,* Theoretical Computer Science **408**, 4 (2008).

[41] S. Hug, A. Raue, J. Hasenauer, J. Bachmann, U. Klingmüller, J. Timmer, and F. Theis, *High-Dimensional Bayesian Parameter Estimation: Case Study for a Model of JAK2/STAT5 Signaling,* Mathematical Biosciences **246**, 293 (2013).

[42] J. Saez-Rodriguez, L. Alexopoulos, M. Zhang, M. K. Morris, D. A. Lauffenburger, and P. K. Sorger, *Comparing signaling networks between normal and transformed hepatocytes using discrete logical models,* Cancer research **71**, 5400 (2011).

[43] I. Stelniec-Klotz, S. Legewie, O. Tchernitsa, F. Witzel, B. Klinger, *et al.*, *Reverse engineering a hierarchical regulatory network downstream of oncogenic KRAS.* Molecular systems biology **8**, 601 (2012).

[44] B. Klinger, A. Sieber, R. Fritsche-Guenther, F. Witzel, L. Berry, *et al.*, *Network quantification of EGFR signaling unveils potential for targeted combination therapy,* Molecular Systems Biology **9** (2013).

[45] D. C. Kirouac, J. Y. Du, J. Lahdenranta, R. Overland, D. Yarar, *et al.*, *Computational modeling of ERBB2-amplified breast cancer identifies combined ErbB2/3 blockade as superior to the combination of MEK and AKT inhibitors.* Science signaling **6**, ra68 (2013).

[46] F. Eduati, V. Doldàn-Martelli, B. Klinger, T. Cokelaer, A. Sieber, *et al.*, *Drug resistance mechanisms in colorectal cancer dissected with cell type-specific dynamic logic models,* Cancer Research **77**, 3364 (2017).

[47] F. Froehlich, T. Kessler, D. Weindl, A. Shadrin, L. Schmiester, *et al.*, *Efficient parameterization of large-scale mechanistic models enables drug response prediction for cancer cell lines,* bioRxiv , 174094 (2017).

SUMMARY

CANCER patients often respond very differently to any given drug. Some patients respond very well, while others do not respond at all, leaving the cancer to grow unimpeded. If we have a good understanding of how this variability in response arises, we will be better able to choose the optimal treatment strategy for each patient.

The variability in drug response observed in patients is also seen in cancer cell lines when they are cultured in vitro. Detailed cell-biological studies have revealed many different mechanisms which affect the response of cancer cells to anticancer drugs. Certain mutations can render cells sensitive to a certain drug, while other mutations, or changes in gene expression, can cause resistance. However, since any combination of these drug sensitivity mechanisms can be operating in a particular cell line, it is difficult to predict whether it will be sensitive or resistant to a particular drug.

Computational modeling can be used to better understand this complexity. In this dissertation, we developed a novel method, which we call Inference of Signaling Activity, that can be used to infer the contributions of different drug sensitivity- and resistance mechanisms. We used the available knowledge of signal transduction in cells, and integrated multiple data types including mutations, gene amplifications and deletions, gene expression levels, protein phosphorylation, growth rates and drug response data to infer the signaling activities in each cell line. After an extensive characterization of thirty different breast cancer cell lines, we developed a model that can explain a large part of the variability in the response of these cell lines to seven different kinase inhibitors. At the same time, the response of some cell lines was not recapitulated exactly. Using further data-driven analysis, we found a novel determinant of mTOR inhibitor sensitivity. Overexpression of 4EBP1 in breast cancer cells renders them more sensitive to these inhibitors. This modeling approach can now be further developed to determine whether it can also be used to explain and predict the response of cancer patients.

Initially this modeling framework did not permit the inclusion of feedback signaling mechanisms, even though we know feedback control to be an important feature of cellular signaling networks. We therefore subsequently extended our framework such that feedback could be included, and with this extension we were able to delineate signaling activities in regulatory networks with multiple, interrelated feedback loops, again taking into account different datasets.

An important consideration in this dissertation was the quantification of uncertainty in model parameters, for which we used Bayesian statistics. If the uncertainty in parameter estimates is not taken into account, we can be lulled into a false sense of security and misinterpret which elements of the model are important. We developed a software package with efficient, multi-threaded implementations of various Monte Carlo sampling algorithms, which allowed the inference to be done in workable amounts of time. We further showed in a different biological system – cell cycle regulation in yeast – that the integration of different types of measurements can increase the identifiability of pa-

rameters. Finally, we investigated whether Bayesian inference with multiple datasets can be done sequentially using intermediate posterior approximations. Each of these contributions to Bayesian inference with multiple datasets may be used more broadly in modeling different biological systems.

Although further development and validation of the drug response models is needed, the use of integrative computational modeling appears to be a promising approach for enabling precision medicine for cancer patients in the future.

SAMENVATTING

PATIËNTEN met kanker reageren vaak sterk verschillend op een gegeven medicijn. In sommige gevallen slaat een medicijn goed aan, terwijl er in andere gevallen maar weinig profijt blijkt te zijn. Om elke patiënt een optimale therapie te geven, is het belangrijk om te begrijpen hoe deze variabiliteit in therapierespons tot stand komt.

Als we kankercellijnen in kweek laten groeien, zien we dat ze ook in deze geïsoleerde omgeving vaak verschillend reageren op medicijnen. Veel cel-biologische studies hebben inmiddels een groot aantal mechanismes ontrafeld die invloed hebben op hoe cellen reageren op medicijnen. Bepaalde mutaties zorgen ervoor dat cellen gevoelig zijn, terwijl andere mutaties, of veranderingen in genexpressie, juist resistentie veroorzaken. Echter, aangezien in elke cellijn verschillende combinaties van deze mechanismes actief kunnen zijn, is het moeilijk om te voorspellen of een cellijn gevoelig of resistent zal zijn voor een bepaald medicijn.

Computationale modellen kunnen helpen om deze complexiteit beter te begrijpen. In dit proefschrift hebben we een nieuwe modeleermethode ontwikkeld, waarmee we de contributie van verschillende sensitiviteit- en resistentiemechanismes kunnen bepalen. We gebruiken daarbij de reeds bestaande kennis van signaaltransductie, en integreren vervolgens verschillende types metingen, inclusief mutaties, genamplificaties en -deleties, genexpressie, eiwitphosphorylatie, groeisnelheden en de uiteindelijke respons van cellen op verschillende inhibitoren. Hiermee wordt dan de activiteit van de signaleringsmoleculen in elke cellijn geschat. Na een uitgebreide profilering van dertig verschillende borstkankercellijnen, hebben we met deze aanpak een model gemaakt waarmee we een groot deel van de variabiliteit in respons van deze borstkankercellijnen op zeven verschillende kinase inhibitoren kunnen verklaren. Tegelijkertijd bleken er ook cellijnen te zijn waarvan de respons minder goed verklaard kon worden door het model. Door vervolgens met data-gedreven methodes verder te zoeken, hebben we een nieuw mechanisme gevonden dat van invloed is op de sensitiviteit voor mTOR inhibitoren. Borstkankercellen die een verhoogde expressie van het 4EBP1 eiwit hebben, blijken gevoeliger te zijn voor deze inhibitoren. Deze modeleermethode zou nu verder ontwikkeld en toegepast kunnen worden om te bepalen of niet alleen de respons van kankercellijnen in kweek, maar ook de respons van patiënten verklaard en voorspeld kan worden.

Het was in eerste instantie met deze modeleermethode niet mogelijk om feedback-signalineringsroutes in de modellen mee te nemen. We weten echter dat feedback-regulatie een belangrijk onderdeel is van cellulaire signalering, en we hebben daarom de methode verder uitgebreid om ook feedbackmechanismes te kunnen verwerken. Met deze uitgebreidere methode hebben we vervolgens de activiteit van verschillende, met elkaar verbonden feedback loops geschat, waarbij we wederom meerdere sets van metingen in acht hebben genomen.

Een belangrijke overweging in dit proefschrift was het kwantiferen van de onzekerheid in modelparameters, waarvoor we Bayesiaanse statistiek hebben gebruikt. Als deze

onzekerheid niet in kaart wordt gebracht, kan een volledig verkeerd beeld ontstaan van welke aspecten van het model daadwerkelijk belangrijk zijn. Om deze analyse mogelijk te maken hebben we software ontwikkeld met efficiënte implementaties van verschillende Monte Carlo algoritmes, waardoor de inferentie in afzienbare tijd gedaan konden worden. We hebben verder ook in een ander biologisch systeem – de regulatie van de celcyclus in gist – laten zien dat het integreren van verschillende types metingen de identificeerbaarheid van parameters kan vergroten. Tot slot hebben we onderzocht of dergelijke Bayesiaanse inferentie met verschillende datasets ook goed sequentieel gedaan kan worden. Deze bijdrages aan de ontwikkeling van Bayesiaanse statistiek voor inferentie met meerdere datasets kunnen breder gebruikt worden bij het modelleren van verschillende biologische systemen.

Hoewel verdere ontwikkeling en validatie nodig is, lijkt het gebruik van integratieve computationele modellen een veelbelovende methode om gepersonalizeerde therapie voor kanker in de toekomst mogelijk te maken.

DANKWOORD/ACKNOWLEDGMENTS

Allereerst wil ik Lodewyk Wessels bedanken, voor de altijd productieve meetings, scherpe inzichten en alle begeleiding en support. Ik waardeer het zeer dat je altijd tijd maakt wanneer nodig, en we kunnen altijd snel de diepte ingaan. Ook op persoonlijk vlak sta je klaar om te helpen.

This dissertation would not have been possible without all the work and dedication of Kathy Jastrzebski. I knew I could always trust you to be thorough and careful, and I believe this has paid off. Moreover, it is just a pleasure to work with you and discuss all the details and latest results. If there are any mistakes left in the text of this dissertation, it must be in parts you haven't gone through!

Tijdens mijn master stage heb ik al veel geleerd van Tjeerd Dijkstra and Tom Heskes, en ik ben dankbaar voor jullie aanhoudende hulp met de Bayesiaans-statistische kant van deze thesis.

The computational cancer biology group which Lodewyk has assembled has been a stimulating environment and has provided a very friendly atmosphere to work in. I would like to thank everyone in this group for useful discussions and feedback, and in particular Sergio Rossell, Evert Bosdriesz and Daniël Vis for input on details of the analyses as well as for discussions on the broader goals and directions of computational cancer biology.

Ik zou Roderick Beijersbergen, Marcel Reinders, Frank Bruggeman and Marjanka Schmidt willen bedanken voor waardevolle begeleiding en feedback. En hoewel ik het toen niet altijd met jullie eens was, zie ik dat het allemaal goede suggesties waren.

Het werk van het 'HER2-project' is niet in dit proefschrift belandt, maar ik kijk ik er erg naar uit om hiermee verder te gaan. Ik wil daarom ook hier Jelle Wesseling, Esther Lips, Mette van Ramshorst, Lennart Mulder and Gabe Sonke bedanken voor het mogelijk maken hiervan, dat jullie het vertrouwen hadden om hiermee van start te gaan en alle bijdrages tot nu toe.

Jaap, we hebben al nagenoeg ons hele leven interessante discussies, en ook zowel de inhoud van als de filosofie achter dit proefschrift hebben eraan moeten geloven. Hoewel een post-hoc fallacy op de loer ligt, lijkt me dat het bijgedragen heeft aan de totstandkoming; en anders is er in ieder geval de last-minute editing en support.

Als laatste wil ik natuurlijk ook Nort, Irma en Lobke bedanken. Het is onmisbaar om zo'n fijne achterban te hebben.

CURRICULUM VITAE

Bram THIJSSEN

31-10-1985	Born in Gouda, The Netherlands.

EDUCATION

1997–2003	Gymnasium Coornhert Gymnasium Gouda
2003–2004	(unfinished) B.Sc. Liberal Arts & Sciences University College Utrecht
2007–2010	B.Sc. Biomedical Sciences Maastricht University
2010–2012	M.Sc. Computational Biology & Bioinformatics ETH Zürich & University of Zürich
2013–	Ph.D. Computational Cancer Biology Netherlands Cancer Institute & Delft University of Technology

PROFESSIONAL EXPERIENCE

2005–2006	Programmer, independent contractor Established a start-up video game development company of which the second half of 2005 was in Navi Mumbai, India.
2006–2008	Programmer, Playlogic Game Factory, Breda, The Netherlands 1 year full-time and 1 year 0.3 FTE
2008–2010	Software Architect, Vsee Labs Inc., Mountain View, CA, USA 3 years 0.4 FTE (in parallel to B.Sc. at Maastricht University).

AWARDS

2008	Maastricht University top-3% award
2018	BioSB Young Investigator Award 2018

LIST OF PUBLICATIONS

9. **Thijssen B**, Jastrzebski K, Beijersbergen RL, Wessels LFA. Delineating feedback activity in the MAPK and AKT pathways using feedback-enabled Inference of Signaling Activity. bioRxiv:268359, 2018.

8. **Thijssen B**, Wessels LFA. Approximating multivariate posterior distribution functions from Monte Carlo samples for sequential Bayesian inference. arXiv:1712.04200, 2018.

7. Jastrzebski K*, **Thijssen B***, Kluin RJC, de Lint K, Majewski IJ, Beijersbergen RL, Wessels LFA. Integrative modeling identifies key determinants of inhibitor sensitivity in breast cancer cell lines. *Cancer Research*, OnlineFirst May 29, 2018, 10.1158/0008-5472.CAN-17-2698.
* *Equal contribution*

6. **Thijssen B**, Dijkstra TMH, Heskes T, Wessels LFA. Bayesian data integration for quantifying the contribution of diverse measurements to parameter estimates. *Bioinformatics*, 34:803-811, 2018.

5. **Thijssen B**, Dijkstra TMH, Heskes T, Wessels LFA. BCM: toolkit for Bayesian analysis of Computational Models using samplers. *BMC Systems Biology*, 10:100, 2016.

4. Johnson J, **Thijssen B**, McDermott U, Garnett M, Wessels LFA, Bernards R. Targeting the RB-E2F pathway in breast cancer. *Oncogene*, 35:4829-4835, 2016.

3. Wang X*, **Thijssen B***, Yu H. Target essentiality and centrality characterize drug side effects. *PLOS Computational Biology*, 9:e1003119, 2013.
* *Equal contribution*

2. Wang X*, Wei X*, **Thijssen B***, Das J*, Lipkin SM, Yu H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature Biotechnology* 30:159-164, 2012
* *Equal contribution*

1. Riniker S, Horta BAC, **Thijssen B**, Gupta S, van Gunsteren WF, Hünenberger PH. Temperature Dependence of the Dielectric Permittivity of Acetic Acid, Propionic Acid and Their Methyl Esters: A Molecular Dynamics Simulation Study. *ChemPhysChem* 13:1182-90, 2012.