University of Stuttgart
Germany

SimTech
Cluster of Excellence

Methodological Concepts for Data-integrated Modeling of Biological Systems with Applications in Cancer Biology

Antje Jensch

Methodological Concepts for Data-integrated Modeling of Biological Systems with Applications in Cancer Biology

Von der Fakultät Konstruktions-, Produktions-, und Fahrzeugtechnik und dem Stuttgarter Zentrum für Simulationswissenschaft (SC SimTech) der Universität Stuttgart zur Erlangung der Würde eines Doktors der Ingenieurwissenschaften (Dr.-Ing.) genehmigte Abhandlung

vorgelegt von

Antje Jensch

aus Berlin

Hauptberichterin:	Prof. Dr. rer. nat. Nicole Radde
Mitberichterin:	Associate Prof. Susana Vinga, Ph.D.
Tag der mündlichen Prüfung:	25.03.2022

Institut für Systemtheorie und Regelungstechnik,
Universität Stuttgart

2023

Contents

Index of notation

Bis-Tris	2-[bis(2-hydroxyethyl)amino]-2-(hydroxymethyl)propane-1,3-diol
ca	constitutively active
CBA	circuit-breaking algorithm
CERT	ceramide transport protein
CFP	cyan fluorescent protein
DAG	diacylglycerol
DLC1	deleted in liver cancer 1
dn	dominant negative
EGF	epidermal growth factor
EGFP	enhanced green fluorescent protein
enetLTS	Robust and sparse logistic regression with elastic net penalty
FA	focal adhesion
FDR	False Discovery Rate
FP	false positive
FPR	false positive rate
GAP	GTPase-activating protein
GEF	guanine nucleotide exchange factor
HER2	human epidermal growth factor receptor 2
HEK	human embryonic kidney
HRP	horseradish peroxidase

IHC	immunohistochemical testing
kb-NB	kb NB 142-70
LDA	linear discriminant analysis
MAPK	mitogen-activated protein kinase
MCMC	Markov Chain Monte Carlo
MLE	maximum-likelihood estimation
NGF	neural growth factor
ODE	ordinary differential equation
OSBP	oxysterol-binding protein
pAb	polyclonal antibody
PDBu	phorbol 12,13-dibutyrate
PI4K	phosphatidylinositol 4-kinase
PKC	protein kinase C
PKD	protein kinase D
PPD	posterior predictive distribution
ppERK	phosphorylation level of ERK
pPKD	autophosphorylated kinase
ppMEK	phosphorylation level of MEK
pRaf	phosphorylation level of Raf
RNA-Seq	RNA sequencing
ROC	receiver operating characteristic
ROCK	Rho-associated protein kinase
ROSIE	RObust Sparse ensemble for outlIEr detection and feature selection
RP	Rank Product
RSK-means	parse robust discriminant analysis with sparse partial robust M regression

siDLC1	DLC1 siRNA
siNT	control siRNA
SPRM/SPRM-DA	Sparse robust discriminant analysis with sparse partial robust M regression
ssHRP	soluble secreted variant of horseradish peroxidase
START	StAR-related lipid transfer
SVM	support vector machine
TCGA	The Cancer Genome Atlas
TGN	trans-Golgi network
TN	true negative
TNBC	Triple-Negative Breast Cancer
TP	true positive
TNR	true negative rate
TPR	true positive rate

Abstract

Biological systems are complex and diverse. Learning about and understanding these systems is nowadays not only based on experimental observations but often also involves mathematical modeling.

In this thesis I describe a workflow for data-based modeling in the context of cancer biology, with a particular focus on sparse data. Data pre-processing, system modeling, model calibration, model validation, and model analysis constitute the five workflow steps. This workflow is applied to three different biological systems. While the first project investigates a feedback mechanism of the known MAPK pathway, the second project analyzes the role of the tumor suppressor protein DLC1 in regulating PKD activity at the Golgi. Finally, the third project gives insight into the genetic composition characterizing triple-negative breast cancer in contrast to other breast cancer types. In the first two projects I employ systems biology approaches whereas the last is based on a classification approach.

All systems studied here are confronted with sparse data. In the first two systems, the sparsity is characterized by a low time resolution, measurements of only a subset of the components, large variability between replicates, and relative measurements. This generates uncertainty in the model parameters when calibrating the model. The sparsity in the third system manifests in a large feature space compared to the number of samples. As most non-sparse methods assume that the number of samples exceeds the number of features they may overfit the training data or fail completely. In addition, outliers can strongly influence the results.

Here, I address the sparsity problem encountered in the first two systems with a combination of statistical methods, which allow the propagation of uncertainty in the model parameters to the model predictions. In order to control the sparsity problem in the third

project, an ensemble integrating sparse and robust methods for feature selection and outlier identification is proposed.

I present a novel quantitative model for the interplay of DLC1 and PKD. Combining biological experiments and mathematical modeling allowed us to generate and validate hypotheses about the role of DLC1 for PKD activation at the Golgi and for Golgi secretory activity. DLC1 is a known tumor suppressor protein that plays a role in cell migration and invasion. Expression of DLC1 is down-regulated in several types of human cancer including liver, breast, and lung cancer.

Standard methods for model validation may not be applicable in sparse data settings. In this thesis new bootstrap-based methodology for model validation in this sparse setting is introduced. I propose a bootstrap-based validation approach for the DLC1 model with the aim to detect overfitting and underfitting. Moreover, for the classification project, a bootstrap approach was developed to verify the robustness of outlier detection and feature selection results regarding variations in the data.

In addition, I introduce a new ensemble approach that combines feature selection and outlier detection results from independent methods. This approach yields robust results for a broad range of data settings. The method was validated using four different approaches. In addition to the above-mentioned bootstrap approach, I performed a simulation study. Moreover, results of our ensemble approach were compared with other studies and methods. Finally, I investigated the biological relevance of the findings for medical data. The results have led to new hypotheses about potential biomarkers.

In conclusion, by discussing challenges and their solutions for the presented projects, I provide a guideline for a variety of biological studies. I also present novel insight into biological signaling pathways, gained by following this guideline. Finally, I introduce new methods that complement and enrich the pool of available methods, making decisions for the modeling of biological systems easier.

Deutsche Kurzfassung

Methodische Konzepte für die datenintegrierte Modellierung biologischer Systeme mit Anwendungen in der Tumorbiologie

Biologische Systeme sind komplex und vielfältig. Um mehr über diese Systeme zu erfahren und ihren Aufbau zu verstehen, werden heutzutage häufig nicht nur experimentelle Beobachtungen, sondern auch mathematische Modelle verwendet.

In dieser Arbeit beschreibe ich einen Workflow für die datenbasierte Modellierung im Kontext der Tumorbiologie. Ein besonderer Schwerpunkt liegt hierbei auf der Verwendung spärlicher Daten. Datenvorbereitung, Systemmodellierung, Modellkalibrierung, Modellvalidierung und Modellanalyse bilden die fünf Arbeitsschritte meines Workflows, welcher auf drei verschiedene biologische Systeme angewandt wird. Während das erste Projekt ein Feedback im bekannten MAPK-Weg untersucht, analysiert das zweite Projekt die Rolle des Tumorsuppressorproteins DLC1 bei der Regulierung der PKD-Aktivität am Golgi. Das dritte Projekt gibt Einblicke in die genetische Charakteristik von dreifach-negativem Brustkrebs im Vergleich zu anderen Brustkrebsarten. In den ersten beiden Projekten verwende ich systembiologische Ansätze, während das letzte Projekt auf einem Klassifizierungsansatz beruht.

Alle hier untersuchten Systeme sind mit spärlichen Daten konfrontiert. Bei den ersten beiden Systemen ist die Spärlichkeit durch eine geringe zeitliche Auflösung, Messungen nur einer Teilmenge der Komponenten, große Variabilität zwischen Replikaten und relative Messungen gekennzeichnet. Dies führt bei der Kalibrierung des Modells zu Unsicherheiten in den Modellparametern. Die Da-

tenspärlichkeit im dritten System manifestiert sich in einem großen Merkmalsraum im Vergleich zur Anzahl der Stichproben. Da die meisten nicht speziell angepassten Methoden davon ausgehen, dass die Anzahl der Stichproben die Anzahl der Merkmale übersteigt, kann es zu einer Überanpassung der Trainingsdaten oder zu einem vollständigen Versagen des Klassifikators kommen. Darüber hinaus können Ausreißer die Ergebnisse stark beeinflussen.

Hier gehe ich das Problem der Datenspärlichkeit der ersten beiden Systemen mit einer Kombination statistischer Methoden an, welche die Übertragung der Unsicherheit der Modellparameter auf die Modellvorhersagen ermöglichen. Um das Spärlichkeitsproblem im dritten Projekt zu lösen, wird ein Ensemble aus spärlichen und robusten Methoden zur Merkmalsauswahl und Ausreißeridentifizierung vorgeschlagen.

Ich stelle ein neues quantitatives Modell für das Zusammenspiel von DLC1 und PKD vor. Durch die Kombination von biologischen Experimenten und mathematischer Modellierung konnten Hypothesen über die Rolle von DLC1 bei der PKD-Aktivierung am Golgi und für die sekretorische Aktivität des Golgi entwickelt und validiert werden. DLC1 ist ein bekanntes Tumorsuppressorprotein, das eine Rolle bei der Zellmigration und -invasion spielt. Die Expression von DLC1 ist bei verschiedenen Arten von menschlichem Krebs, darunter Leber-, Brust- und Lungenkrebs, herabreguliert.

Standardmethoden für die Modellvalidierung sind in spärlichen Datenumgebungen möglicherweise nicht anwendbar. In dieser Arbeit wird eine neue bootstrap-basierte Methodik für die Modellvalidierung in diesem spärlichen Kontext vorgestellt. Ich schlage einen bootstrap-basierten Validierungsansatz für das DLC1-Modell vor, um Over- und Underfitting zu erkennen. Darüber hinaus wurde für das Klassifikationsprojekt ein Bootstrap-Ansatz entwickelt, um die Robustheit der Ergebnisse der Ausreißererkennung und der Merkmalsauswahl in Bezug auf Datenschwankungen zu überprüfen.

Darüber hinaus stelle ich einen neuen Ensemble-Ansatz vor, der die Ergebnisse der Merkmalsauswahl und der Ausreißererkennung aus unabhängigen Methoden kombiniert. Dieser Ansatz liefert robuste Ergebnisse für ein breites Spektrum von Daten. Die Methode wur-

de anhand von vier verschiedenen Ansätzen validiert. Zusätzlich zu dem oben erwähnten Bootstrap-Ansatz habe ich eine Simulationsstudie durchgeführt. Außerdem wurden die Ergebnisse des Ensemble-Ansatzes mit anderen Studien und Methoden verglichen. Schließlich habe ich die biologische Relevanz der Ergebnisse für medizinische Daten untersucht. Die Ergebnisse haben zu neuen Hypothesen über potenzielle Biomarker geführt.

Zusammengefasst, stelle ich einen Leitfaden für eine Vielzahl biologischer Studien zur Verfügung, indem ich die Herausforderungen und ihre Lösungen für die vorgestellten Projekte erörtere. Ich präsentiere auch neue Einblicke in biologische Signalwege, die durch die Umsetzung dieses Leitfadens gewonnen wurden. Abschließend stelle ich neue Methoden vor, die den Pool der verfügbaren Methoden bereichern und die Entscheidungen bei der Modellierung biologischer Systeme erleichtern.

Chapter 1.

Introduction

Nature is fascinating. Millions of years of evolution constructed an intricate network of species on a population level, where each species has developed to fit a specific niche for feeding and breeding, and to defend or hide from enemies. Looking at a smaller scale, for example a single organism composed of different cell types, these cells communicate in astonishing ways in order to heal a wound or memorize our experiences. Zooming in to even smaller elements, there exist regulatory mechanisms such as signaling pathways that can decide a cell's fate by triggering proliferation or cell death through molecule interactions. While we are aware of the existence of these systems, much is still to be uncovered about their components and how these components interact.

Disturbances to a system, by removal of components or changes in the environment, can negatively influence the whole system, ultimately even causing system collapse. The influence single components can have on a system are easier to observe in large scale systems. Taking insects as an example, these species are a direct food source for many animals while also pollinating plants. The currently observed declining rate of insect numbers is thus threatening the existence of many plants and animals ([59, 85, 198] and sources therein). In order to preserve insect numbers and their diversity, it is important to understand what factors contribute to their decline. Considering the more specific case of wild bee species, reasons include the usage of certain pesticides such as glyphosate [48, 77], habitat loss, climate change and the import of foreign competitors or predators [198].

Another example on a much smaller scale and a different field of

science is cancer. Cancer is a life threatening disease that is characterized by an uncontrolled cell growth and the expansion to other areas of the body. Understanding the causes of cancer growth provides valuable information in order to develop new treatment strategies. A simple answer on cancer causation would be that mutations in the genome can influence cell functions such as growth and division [242]. However, a single mutation is in general insufficient to cause cancer growth, as body cells contain several mechanisms to identify and destroy malfunctioning cells. Instead, [86] suggest six hallmarks that different cancers generally have in common. In particular, there are six cell functions that have to become deregulated or dysfunctional to allow malignant cell growth. These include the ability to circumvent programmed cell death, to stimulate angiogenesis, and to metastasize.

What these two examples have in common, is the complexity of the systems, involving a complicated network of many interacting components and even several organizational levels. Thus, a fundamental understanding of these systems in terms of system components, underlying processes, and interaction mechanisms is imperative in order to be able to prevent system collapse. In the case of endangered species like bees this may encase problem-specific protective programs, while in the case of cancer the development of drugs targeted to disease-specific receptors can support the system's own regulatory mechanisms.

Gaining knowledge starts from direct or experimental observations in order to determine which components contribute to specific system behavior and in which ways components interact. Looking back at the example of the bee ecosystem, the system is currently characterized by a continuing decrease in bee numbers. As mentioned before, there are multiple stress factors that contribute to the decline of bee populations. Depending on the time and location, several of these accumulate. However, the complex interplay of different stress factors as well as the difficulty to control conditions of free-flying bees impede the experimental investigation of interacting stress factors on populations [76].

Reexamining the cancer example from above, the cancer-inflicted organism as a system is afflicted by declining health and in the worst

case death. Similar to the insect example, cancer is highly complex involving many factors. In this context, the hallmarks of cancer constitute a summary of common biological properties of cancer cells. However, these properties can be acquired through various mutations in the cancer cells. As there are more than 100 cancer types, which can in turn be divided into subtypes, the exact mutations causing for example apoptosis can differ. Possible genetic changes can be found in the form of upregulated anti-apoptotic proteins or downregulated pro-apoptotic proteins [208]. Complexity is also added by the fact that cell signaling pathways are not isolated modules but may share components or interact with other pathways. Thus, research for treatment strategies requires understanding of not only components and networks on the scale of a single module but also of the "big picture" of pathway interactions and differences in the gene profiles of different cancer types. [103] point out the difficulty of understanding how specific mutations influence the system and how to identify good points of intervention for treatment.

As the examples above illustrate, biological systems are complex in their composition and relational networks, while a full observation under all conditions of interest may not be possible. Only by systematically analyzing the data obtained from observing and measuring parts of a system, can we hope to gain enough understanding to influence system behavior in a desired way. Nowadays, many computational methods and approaches have been developed for data analysis. These methods also take into account different system aspects that can be of interest, for example key components, network structures or system responses to perturbations. Additionally, the availability of data has improved with the rise of diverse public online databases. The Protein Data Bank [20] providing structural information on biological macromolecules, The Cancer Genome Atlas [231] containing more than 10,000 samples of molecular data from cancer patients, and the Global Biodiversity Information Facility [70] combining and providing access to biodiversity databases around the globe are just a few examples. These databases can provide large scale datasets, which can be exploited in addition or as an alternative to conducting experimental work. Nevertheless, the task of properly

extracting information from data remains challenging. Biological research topics span a wide spectrum from small scale molecules to large scale populations.

In order to simplify terminology and specify methodology, this thesis focuses on the study of biological systems related to health science, including different methodological and thematic approaches to increase the knowledge on diseases such as cancer.

Understanding biological systems builds the basis for correcting or preventing negative developments, such as cancer and other diseases. This process can lead down to learning about system components interacting on a molecular level. With the increase of computational power and the development of specialized algorithms, the study of biological systems is now strongly connected to computational tools. Hereby, several fields jointly advance the integration and development of software tools in the study of biological systems. These different fields include for example systems biology, sequence analysis, as well as gene and protein expression analysis.

The approach of combining biological research with computational methods is difficult from several points of view. Firstly, it requires knowledge of different study fields, such as biology, chemistry, physics, mathematics and informatics. Though, the exact requirement may depend on the topic of the project at hand. Secondly, advances in measurement techniques raise the need for new or adjusted methods for data handling, and statistical analysis increases as well. Apart from the measurement technique the data was gathered with, data can also be divided according to its sample size. Commonly, data is classified in sparse, medium and large datasets in regard to the available dataset.

Western blotting and RNA sequencing (RNA-Seq) are two prevalent, yet very different measuring techniques. Western blotting is used to detect specific proteins by employing corresponding antibodies that target the proteins of interest. However, as the price of antibodies can be high and the experimental procedure is time-consuming [161], the number of replicates of an experiment and overall data points can be very low. On the other hand, RNA-Seq employs high-throughput sequencing methods to investigate the transcriptome of cells. As

RNA-Seq methods allow transcriptome-wide measurements of gene expression, up to thousands of targets can be assessed in a single sequencing experiment. The differences in data compositions pose different challenges, with each data type requiring individual attention for information extraction. The analysis of sparse data as well as large datasets can limit the methodological and computational options as standard methods may be unsuited.

In this thesis, systems biology and classification are considered as two widely used data-integrating approaches to improve system knowledge. They represent suitable complementary approaches, as classification is often used to generate hypotheses while systems biology approaches mainly focus on investigating hypotheses.

1.1. Systems biology and classification - two approaches for the study of biological systems

In the following, I will shortly provide an overview of the general approaches and challenges in systems biology and classification.

What is systems biology?

In the preface of [127], systems biology is described as "the scientific discipline that studies the systemic properties and dynamic interactions in a biological object, be it a cell, an organism, a virus, or an infected host, in a qualitative and quantitative manner and by combining experimental studies with mathematical modeling". This concise characterization summarizes the important aspects and also hints towards challenges that systems biology is concerned with. Systems biology aims to gain a system-level comprehension of biological systems, including the structural and dynamical properties that underlie the biological function and allow to predict system behavior under certain stimulations and perturbations [126]. Furthermore, the biological object of interest can comprise different structural levels and scales [127, Chapter 1]. In a simplified way, you could consider a

human as consisting of different tissues, which are comprised of cells that include different network and pathway structures of interacting proteins, that in turn are encoded in the DNA. In case of a human, the size of the different levels ranges from meters to nanometers. Similarly, time scales could be in years for developmental processes, hours for the sleep cycle or micro- to nanoseconds for protein interactions in pathways. This indicates the diverse possibilities for research topics as well as the complexity faced in systems biology and the resulting necessity for the development of advanced technologies and methods to tackle challenges arising from these research topics. While the refinement of measuring techniques to acquire data for multiple scales is highly favorable for system modeling, the main focus of this thesis is inhered in the development of techniques for data analysis, such as data processing or system modeling. The different facets concerned and the methodologies presented in literature will be examined in more detail later in this chapter.

Returning to the description of systems biology in [127], the interdisciplinary nature of this scientific field is brought up. Systems biology requires an understanding of a wide range of disciplines in order to arrive at accurate and useful results. [175] mentions the need to be familiar with the fundamentals of life sciences such as molecular biology, genetics or cell biology. As systems biology integrates data, basic knowledge on typical features of measuring techniques is necessary. Furthermore, with the rise of high-throughput 'omics data, such as transcriptomics, proteomics, and genomics, a certain understanding of bioinformatics and big data analytics has become crucial to handle large datasets. In order to derive and simulate a mathematical model, the scientist also needs to have a good grasp on the mathematical concepts of linear algebra and optimization.

For a detailed introduction to the field of systems biology, the reader is referred to [175] and [127] for more recent books on the topic, or [126] as one of the first books about systems biology.

Sample	Feature 1	Feature 2	Feature 3	Label/Class
1	0	0	3	A
2	1	0	2	B
3	1	0	3	B
4	0	0	8	A
5	1	1	1	C

Table 1.1.: Example dataset consisting of five sample with corresponding label and three features.

What is Classification?

A classification problem considers a dataset consisting of different samples, which each have an allocated class. Classes can hereby for example refer to the disease status (healthy or sick), disease treatment (drug A, drug B, drug C), or gender (male or female). The class, also called label, is determined by all or a subset of the sample features given some classification rules. As a simple example consider Table 1.1, that lists a small dataset consisting of five samples, three features and a label for class membership of each sample. Here, class membership was determined by the sum of feature 1 and feature 2. Thus, the sum defines the classification rule and the subset is given by feature 1 and feature 2.

In the general application scenario using real data, both the subset and the classification rules are unknown while only the class membership of each sample is given [139]. For the example above, this implies that only Table 1.1 is given without further information. To deduce the classification rules, algorithms, so-called supervised classifiers, learn functional relationships between the features and the samples' class labels. In the following, the learned classification rules can then be applied to classify new samples without labels given. On the contrary, methods performing unsupervised classification, learn to group samples according to the differences and similarities of the feature values in order to deduce the class membership of the samples

without prior knowledge. This group of algorithms is also referred to as clustering approaches.

Classification approaches have been applied to a wide range of research topics that are not limited to biology and health sciences. To name a few examples, classification approaches are applied in handwriting and speech recognition, internet search engines, anomaly detection, and astronomy. In the fields of life and health science, classification approaches have been adopted for disease classification of Schizophrenia (e.g. [186]) and Parkinson's disease (e.g. [116]), biomarker discovery for muscle aging (e.g. [153]), cancer diagnosis, prognosis, or treatment (e.g. [149, 238]), and protein structure prediction (e.g. [74]).

Similar to systems biology, classification can also be considered as an interdisciplinary approach. It requires at least a basic understanding of the data acquisition techniques to ensure adequate data handling. Knowledge about different classification methods can be useful to identify the most appropriate approach for the given data and research question. Additionally, awareness of the biological background is valuable for a meaningful assessment of the classification results.

A general introduction to classification can be found in [1, Chapters 10, 11] and [195, Chapter 8]. [106] review classification methods specifically for RNA-Seq data.

1.2. The process steps of studying biological systems

Despite the differences in systems biology and classification, the process for integrating data in a modeling framework is very similar and follows the same steps. The modeling process can roughly be divided in five steps:

1. **Data pre-processing**,
2. **System modeling**,
3. **Model calibration**,

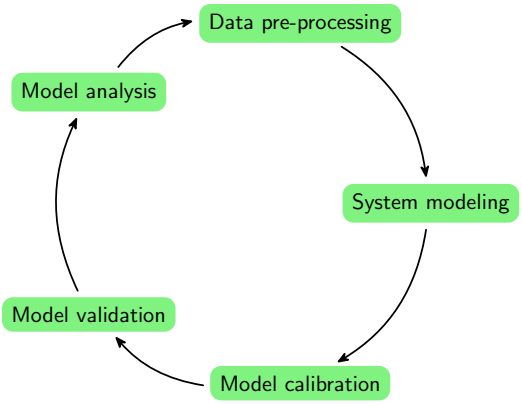


Figure 1.1.: Graphical representation of modeling steps for studying biological systems.

4. **Model validation**,
5. **Model analysis**.

Data is prepared in the first step as a foundation for the following steps. This includes the processing, evaluation and analysis of the available data. Then, a mathematical model description of the studied system is derived from the data and possibly literature, followed by parameter estimation that aims at optimally fitting the model to the data. In the next step, the calibrated model is tested in terms of its goodness of fit and predictive power. Finally, the validated model can be used to perform different analyses in order to investigate system properties of interest. This may lead to new research questions and thus back to the data preparation step. Figure 1.1 visualizes the workflow of the introduced modeling process. Despite the ordered description, decisions for one process step can influence any of the other steps. This can arise when decisions or results in those steps require an update or change in earlier steps. The specific situations that can influence other process steps will be discussed in the descriptions of the corresponding modeling steps.

Within the starting step of data preparation, the data acquisition

deliberately does not contain a point for data acquisition, as biological experiments, clinical data and other data sources are out of the scope of this work and the overall modeling procedure. In general, this work assumes a separation of the data acquisition and modeling tasks. Consequently, data is considered to be provided by or collected from an external source, such as collaboration partners, publications and data repositories. Nevertheless, as mentioned before, knowledge about the experimental setup and data composition is important for most parts of the modeling process. In the following subsections, each modeling step will be closely examined in terms of the tasks that are performed, developed methods and challenges addressed.

Data pre-processing

Data pre-processing (or preparation) encompasses the collection, cleaning and transformation of data from one or more sources prior to its usage. It can also include a first analysis in order to provide additional input for following project steps. Processing tasks may differ depending on the type of data and the intended use case.

As mentioned before, data can be collected experimentally, from publications or from repositories and should be selected with the research question in mind. Considering biological data related to health science, data used for modeling studies can be divided in four groups depending on their source: sequencing data (like of the genomics, transcriptomics, and metabolomics), sensor data (like electrocardiogram or electroencephalogram signals), health care data (for instance electronic health records with data about diagnosis, treatment and discharge) and experimental data (for instance from western blotting, mass spectroscopy and cell cultures). This restricted summary of data sources already suggests one challenge of data preparation. Each data source provides different data types and each dataset can have different properties. Additionally, preparation may also depend on the model analysis tools that will be applied later on.

As preparation tasks can be numerous, commonly encountered problems of this step are stated as follows:

- Outliers,

- Measurement noise,
- Dimensionality reduction and feature selection.

Outliers

Outliers are mostly distinct data points lying, as the name suggests, outside the remaining data points. Identifying and dealing with outliers can be a crucial part of data preparation. Independent of the data type, the identification and handling of outliers hidden in the data can be important for the calibration performance. Outlying data points may impede the accuracy of the calibration process in systems biology as well as classification problems. In [94] an outlier is described as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". Consequently, it may not be surprising that outliers can impede the process of fitting parameters to the data and learning about the underlying system. Taking basic linear regression as an example, an outlying data point can have a substantial influence on the result of model calibration. Consider six data points drawn from a linear model $y = \theta_1 x + \theta_2 + \varepsilon$ with $\theta_1 = 0.5, \theta_2 = 1$ and $\varepsilon \sim \mathcal{N}(0,\, 0.05^2)$ (black dots in Figure 1.2). A least squares optimization then provides $\hat{\theta}_1 \approx 0.54$ and $\hat{\theta}_2 \approx 0.84$ as optimal estimates of the real parameters θ_1 and θ_2 when trying to relearn the original model. However, replacing the fifth data tuple with an outlying data point with a much smaller y-coordinate, the estimation accuracy deteriorates with estimates $\hat{\theta}_1 \approx 0.81$ and $\hat{\theta}_2 \approx 0.54$. The corresponding regression lines for both cases are shown in Figure 1.2. The change in the fifth data point is marked with a red dashed line towards the outlying value. Representing the outlier-free fit, the black regression line matches the data well. On the other hand, the red regression line, calculated for the outlier-infected dataset, fails to capture the underlying model.

In order to be able to detect outlying data points, it is important to be aware of the different types of outliers and outlier sources. Outliers can hereby originate from a multitude of sources. Common examples are errors during data acquisition, corresponding to a human error, instrumental errors influencing the measurement, or biological

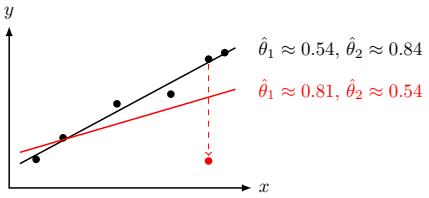


Figure 1.2.: Least squares regression is influenced by outliers in the dataset. Least squares regression for data points sampled from $0.5x + 1 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,\, 0.05^2)$ (black dots and line). Red line shows regression result including one outlying data point (red dot).

variability. Importantly, the last type of outliers does not represent an error. Instead, this kind of outlier, caused by the heterogeneity of biological systems, provides new insight in the data. Taking disease classification as an example, an outlying sample could suggest the existence of an additional subgroup of patients. Consequently, awareness and identification of outliers can be especially crucial for classification tasks.

Outliers can be grouped in the univariate and multivariate type. Univariate outliers can be identified by examining single features of the dataset. A common visual approach to find these outliers is the boxplot, which calculates the interquartile range to identify outliers. Values that extend beyond 1.5 times the interquartile range from the first or third quartile are considered outliers. In classification, a special case of this type of outlier is the instance of a switched class label of a data point (or sample). In this case, the data point should have common values for all features of a class other than the one it is labeled with. Multivariate outliers can have more than one extreme value or may not be recognizable in a single feature. For the first type, values of several features may present extreme values with respect to the average class values. Furthermore, outliers may become apparent only when analyzing higher dimensional feature combinations. Values of these outliers are inconspicuous in each variable, but do not follow

higher dimensional patterns. The outlier in the regression example in Figure 1.2 represents such an outlier. While the outlying point matches well with the other data points when considering the x- and y-values individually, it clearly deviates from the correlated behavior of the other data points. If the origin of the outlying values is natural, this may indicate a need for further analysis to explain the emergence of such observations.

[138] summarize three approaches for dealing with outliers. In order to handle outliers, it is possible to trim the dataset, which implies an exclusion of outlying values. In the second approach outliers are either weighted or replaced with normal values in order to control their influence on model calibration. The third option requires a specialized method in order to estimate model parameters robustly with respect to outliers. Removing outliers is considered inappropriate as these values still constitute observations. In turn, it is suggested to employ robust methods [138], which evaluate and weight or exclude outlying samples during the model calibration. As this approach influences the choice of the model calibration method, available methods will be introduced in the Section Model calibration.

Measurement noise

Experimental data, as basic module for mathematical model calibration, generally contain measurement noise. Depending on the model calibration approach chosen in the Section Model calibration, an assessment of this noise may be required. Especially in systems biology, consideration of noise as influence on the modeling process can be crucial. [189], for example, emphasize the importance of a suitable quantification of the measurement noise for model calibration.

A commonly applied model for measurement noise is the additive normally distributed error model. For n experimental data points, the measured data y_i, $i = 1, \ldots, n$ can then be described by

$$y_i = x_i(\theta) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), \tag{1.1}$$

where $x_i(\theta)$, $i = 1, \ldots, n$ is the model output dependent on the

parameterization θ. The variable ε_i, $i = 1, \ldots, n$ represents a normally distributed noise with a zero mean and variance σ_i^2. For simplicity reasons, Equation 1.1 summarizes n data points. However, depending on the scenario, multiple variables, time points and experimental conditions can be included. In many cases, values for σ_i^2 are unknown and cannot be extracted from literature. Consequently, they have to be either estimated from experimental replicates or included in the model calibration as unknown parameters.

Estimating variance values directly from the experimental data is a standard approach to infer noise parameters. However, this requires a suitable amount of replicates to enable reliable estimation results [189]. Additionally, if only single replicates are available, this approach is infeasible.

When including noise parameters in the model calibration, a prior selection of the noise distribution of ε in Equation 1.1 is necessary. Standard choices for error models are either an additive normal distribution or a multiplicative log-normal distribution. Although several studies favor the multiplicative error model especially for specific data types (e.g. [135, 145] and sources therein), applying a Gaussian noise model remains a widely accepted strategy applied in many scenarios (applied for example in [19, 56, 185]). In recent years, alternative distributions have been suggested in order to improve parameter estimation in the presence of outliers. A study by [151] compared the Laplace, Huber, Cauchy and Student's t distribution as possible alternatives to the Gaussian distribution to assess the noise distribution of data containing outlying values. The author's, after evaluating the error models on an ordinary differential equation model of the JAK/STAT signaling pathway, recommend the usage of a Laplace or Huber distribution for outlier-corrupted datasets.

The studies by [135] and [151] evaluate error models on exemplary system models and data. As such, results may differ depending on the different biological systems and data settings. However, the computational cost for evaluating several error models by repeating the parameter estimation for each scenario may be too high to include in a standard framework. Additionally, further settings can arise when also considering the number of variance parameters. The simplest

assumption assigns one variance parameter for all measurement points, independent of species, time or experimental condition. [5] choose this setting when comparing two model hypotheses for the propagation of methylation states of the histones H3K27 and H3K36 together with two noise model options. In contrast, in a modeling study by [158] a noise parameter was introduced for each measurement technique, observed variable and cell type.

Selecting a fitting error model with the optimal granularity with respect to the variance parameter is a computationally expensive task when included in the model calibration step. Therefore, in Chapter 3, a data pre-processing step is introduced that allows for a time-efficient and easily adaptable error model selection.

Dimensionality reduction and feature selection

In the context of this thesis, this topic is mainly related to classification problems. Consequently, the following review of methodological approaches will be focused on this problem class. However, basic concepts from the field of classification are also applicable in systems biology.

Large datasets, as often seen in omics studies, may require excessive computation time while containing uninformative variables. Therefore, selecting a priori a subset of the data space can speed up computations and provide better interpretability of the results. There are two ways to reduce the dimensionality: (i) reduce the number of samples or (ii) reduce the number of features. As biomedical datasets tend to have an imbalance of a huge number of features with a comparatively small sample number, I will focus on point (ii) for dimensionality reduction. The curse of dimensionality is a well-known problem when working with high-dimensional datasets. High dimensionality can not only impede computation speed but also the accuracy of learning algorithms and the usefulness of the calibrated model in terms of its interpretability [147].

Removing unimportant or redundant features constitutes a first step to reduce the dimensionality. For a classification task this can for example consist of features with a constant value across all

samples, which do not provide additional knowledge about the class separation. Detection and deletion of such features can be quickly executed without specialized tools. Subsequently, extracting features which play an important role for the class separation, is in general computationally more demanding and a variety of methods exists.

There are different approaches for feature selection in classification problems that also vary in the modeling step at which the selection is performed. Based on these differences, methods can be divided into three categories: filter, wrapper and embedded approaches [147]. Filter methods employ a performance measure in order to identify important features in a pre-processing step before calibrating the model [117]. Methods have been developed based on several different measures, such as distance, consistency, or similarity. Accordingly, a variety of filter methods exists. Examples include Correlation-based feature selection, the Fisher score, Chi-square, and ReliefF, which are described in [84, 148, 246], and [133] respectively.

While filter methods are applied as a pre-processing step, independent of the classification model, wrapper methods utilize the classifiers performance metric to evaluate groups of selected features. As the classification task needs to be repeated for each feature subset, wrappers are computationally expensive to use [117].

Finally, embedded methods perform feature selection as an integrated part of the classification task. Consequently, this approach is related to the choice of the model calibration approach and not a pre-processing step. Further explanations are therefore presented in the later Section Model calibration.

The above list of data preparation tasks is not exhaustive, but presents key aspects of this step that will be revisited in the following chapters. In Table 1.2 a list of further common data preparation tasks is presented together with a few references for further reading.

System modeling

The second process step, namely the system modeling, defines the model structure and the mathematical description. Therein, the connections and relations of the system are characterized according

Data preparation task	potential subtasks	further reading
Missing values	identify and handle missing data entries	[138]
Feature engineering	create new features from original data	[51]
Labeling	prepare labels for data samples, format labels in accordance to algorithm requirements	[166]
Data partitioning	appropriately split data for model training and testing	[157], [241]
Transformation	adjust data type (e.g. continuous to categorical), normalization (e.g. adjusting for measurement specifications)	[1, Chapter 2]

Table 1.2.: Data preparation tasks

to the chosen modeling approach. As modeling approaches differ widely for systems biology and classification tasks, methods will be considered separately in the following.

Systems Biology

In systems biology, the model provides a simplified description of the studied biological system. While a model may not be an exact representation of the true mechanism, good mathematical models can be employed for various tasks. Such tasks may include a hypothesis test for the network structure of a system, as in [12], the comparison of different potential mechanisms, as in [55], or the exploration of system behavior under changing conditions, as in [40]. The research question will hereby influence the choice of a modeling approach.

Besides the selection of a mathematical model to describe the system as good as possible, one important question remains. What exactly is a *good* model? The famous quote by statistician George E. P. Box:

> *"Essentially, all models are wrong, but some are useful."*
>
> Box and Draper [33]

has become an aphorism for the difficulty of developing appropriate models in all fields of science.

As already mentioned earlier, the interdisciplinary nature of systems biology widens the range of possible methods to infer knowledge about biological systems. Nevertheless, it also proposes a challenge, as researchers from different disciplines and with different perspectives on the research subject need to cooperate. In this context, [22] identifies communication problems between experimentalists and modelers as an obstacle for the development of useful models. The author suggests, that while experimentalists may not be fully aware of the ability of mathematical modeling, modelers may tend to apply simplifications without proper explanation. This highlights the necessity and difficulty of good communication in multi-disciplinary projects.

The model structure is determined by the given data and, if available, information from literature. It describes the relevant system components and their relationships. Given the model structure, a mathematical model is defined. Hereby, different modeling formalisms can be applied that determine model properties. Models vary in their temporal (static/dynamic) and spatial (discrete/continuous) description. Moreover, depending on the applied model, the results are able to make qualitative statements about the process or give quantitative values stemming from the model. Additionally, models can be categorized in deterministic or stochastic models leading to one concrete statement or a probability of the occurrence of the result. In particular, systems biology favors the description of the model via ordinary differential equations (ODE). Therefore, the presented methods and tools will be mainly based on the assumption that ODE models are deployed. Nevertheless, many other formalisms exist, for example directed graphs, boolean networks and stochastic master equations, stating just a few. A review of different modeling methods can be found in [115].

The choice of the optimal method depends on the system under study, the available data as well as the goal of the study. A typical model within the scope of systems biology can be given in the form of:

$$\dot{\mathbf{z}}(t) = \frac{\mathrm{d}\mathbf{z}}{\mathrm{d}t} = f(\mathbf{z}(t), \mathbf{u}(t), \boldsymbol{\theta}), \tag{1.2a}$$

$$\mathbf{x}(t) = h(\mathbf{z}), \tag{1.2b}$$

where $f(\mathbf{z}(t), \mathbf{u}(t), \boldsymbol{\theta})$ describes the changes of the n state variables $\mathbf{z}(t) \in \mathbb{R}^n$ over time t depending on, potentially time-dependent, inputs $\mathbf{u}(t) \in \mathbb{R}^m$, and a parameter vector $\boldsymbol{\theta}$. The function h maps the state variables $\mathbf{z}(t)$ to the model output $\mathbf{x}(t)$. Each state variable hereby represents a system component and inputs refer to stimuli from outside the system. The set of parameters $\boldsymbol{\theta}$ can comprise three different types: 1. Dynamic parameters used for process description in the network, 2. observation parameters which define a connection between model variables and the output, and 3. error

parameters from the error model [89]. While dynamic parameters are typically unknown with limited prior knowledge of the system, observation parameters, such as scaling factors, might be known from the experimental setup.

The Equation 1.2 characterizes a reaction dynamic which can be defined in different ways. In order to illustrate the basic concepts of biological modeling, three modeling strategies for an irreversible reaction will be examined. Thus, consider a reaction where a substrate A is converted to a product B without the possibility to change back:

$$\mathrm{A} \xrightarrow{k} \mathrm{B},$$

where k describes the conversion rate of the reaction. In order to mathematically describe the above reaction, several modeling strategies, which represent different assumptions about the system behavior, are possible. In a reaction network, these different assumptions can be modeled adapting various reaction rates.

The most basic approach is given by mass-action kinetics. Here, a linear activating relation between the amount or mass of variable A and variable B is assumed. The reaction velocity v describes the change of the amount of compound B over time t. Applying mass-action kinetics yields the reaction velocity

$$v = \frac{\mathrm{d}[B]}{\mathrm{d}t} = k[A], \tag{1.3}$$

where compound amounts for A and B are marked with square brackets $[A]$ and $[B]$, respectively. Analogously, the change in compound A can be described by $\mathrm{d}[A]/\mathrm{d}t = -k[A]$.

While this linear activation of B is inherently not bounded, the other options provide a maximal reaction rate $r_{\max}(x)$. Michaelis–Menten and Hill-type functions are thus used to describe saturation events approaching a maximum rate.

Michaelis–Menten kinetics were derived for enzymatic reactions. In this setting, an enzyme E and a substrate S bind and build a complex ES in a reversible reaction. The enzyme-substrate complex then disintegrates in an enzyme E and a product P, which can be summarized as

$$\mathrm{E} + \mathrm{S} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{ES} \xrightarrow{k_2} \mathrm{E} + \mathrm{P},$$

with rate constants k_1, k_{-1} and k_2. The original work by Michaelis and Menten ([160], translated to English in [113]) assumed an equilibrium for the process of the formation of the complex ES. From that, the assumption of constant amount of total enzyme $[E]_{\mathrm{tot}} = [E] + [ES]$, and the application of mass action kinetics, the reaction velocity for product formation can be derived as

$$v = \frac{\mathrm{d}[P]}{\mathrm{d}t} = k_2[E]_{\mathrm{tot}} \frac{[S]}{k_{-1}/k_1 + [S]}. \tag{1.4}$$

As the constant total enzyme amount can be considered as a parameter, the product formation is solely dependent on the substrate variable $[S]$. It describes a saturation event, where product formation increases with the amount of substrate available but cannot surpass a certain threshold that is given by $k_2[E]_{\mathrm{tot}}$. As Equation 1.4 only depends on parameters and the substrate amount, this equation can also be applied to the above example reaction with A and B, if the reaction velocity is assumed to correspond to a Michaelis–Menten equation.

Hill-type equations as discussed in [43, 73] can be used to describe sigmoidal reaction velocities as shown in Figure 1.3. Hill-type equations also describe a saturation event with a maximum velocity $v_{\max}$ and present a more general version of the Michaelis–Menten equation:

$$v = v_{\max} \frac{[S]^n}{K_A^n + [S]^n}. \tag{1.5}$$

The parameter $v_{\max}$ depicts again the maximum velocity that can be reached, K_A is the substrate concentration at which half the maximum velocity, $v_{\max}/2$, is reached, and K_A^n is the equilibrium dissociation constant. Finally, n is the Hill coefficient and defines the steepness of the slope for $n \geq 1$. The special case of $n = 1$ corresponds to a Michaelis–Menten equation. Hill-type equations are used in biochemistry to model the binding of a ligand to a molecule. Here, it is assumed that bound ligands increase the probability of

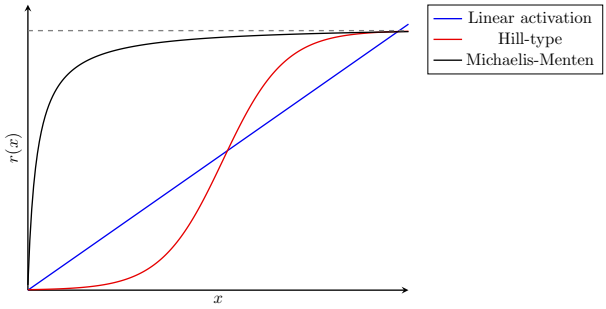


Figure 1.3.: Different mechanisms of activation.

additional ligand binding.

Figure 1.3 sketches the three possible modeling scenarios for reaction rates for an activation from A to B.

Mass-action kinetics, Michaelis–Menten and Hill-type equations constitute important building blocks for the mathematical modeling of a multitude of different biological systems. To name some examples, mass-action kinetics have been used to model histone tail methylations [5], protein phosphorylation [158], and pathways of signaling lipids [171]. Hill-type equations have found application in the modeling of gene expression dynamics under p53 pulsing [209], of infection rates in an HIV model [15], and of muscle forces during different activities [83].

The simple reaction example of substrate A reacting to product B depicts a single reaction, whereas real models are much more complex. Models applied to realistic problems include more variables and interactions, depending on the number of components considered. Additionally, experimental conditions also need to be incorporated in the model, adding to the number of unknown parameters. Experimental conditions hereby describe different experimental settings, for example the unperturbed versus a perturbed system. In a perturbation experiment, the total amount of a compound may be up- or downregulated artificially. Another possibility are inhibition experiments, where special inhibitors are used to target specific compounds

and their activation. Alternatively, activating signal compounds can be supplied as an input u, as given in Equation 1.2.

Including more variables, interactions, and experimental conditions allows for more insight in the biological system, but at the same time proposes a multitude of additional parameters which also increase the complexity of the model. Additionally, increasing the model size and complexity naturally increases the computational cost for evaluating the differential equation model.

Another important point of note is given by the danger of overfitting. Overfitting refers to models which are too adjusted to a given dataset, including extreme values and outliers. Thus, the model looses its usefulness, as it does not describe the actual system mechanism but one specific dataset. Evaluating whether a model is overfitting the experimental data is an important step of model validation and will be discussed in the corresponding Section Model validation. Nevertheless, a careful choice of model granularity can provide a good approach to avoid overfitting.

Consequently, while the model should describe the biological mechanisms, it needs to remain as simple as possible. Simplifications need to be balanced to on the one hand reduce the parameter space and allow for better model interpretability while on the other hand maintain the ability to correctly describe the system dynamics. Mass conservation of compounds provides a simple approach to reduce the number of variables under the assumption that the total compound amount is constant over the duration of the experiment.

Classification

Constructing models for classification tasks differs from systems biology approaches. Similarly to the systems biological approach, classification defines a model that aims to approximate a mapping function from input variables to output variables. However, the output variable takes the form of labels (representing different classes). Given the class membership, the classifier hereby learns the underlying data structure that generates the labels. The learned model can then be used to predict the class membership of new data points whose labels

are unknown. Standard models are for example Logistic Regression, Artificial Neural Networks, Decision Trees, Support Vector Machines (SVM) and Naive Bayes. Through ongoing research, there exist several variants of these approaches, such as regularized logistic regression [232], principal weighted logistic regression [124], random support vector machine cluster [23], and Information Gain-Support Vector Machine [66]. Consequently, a modeler has to choose from an increasing number of methods. Additionally, even though comparative reviews exist, there is no generic optimal method as the best approach may differ for different datasets and project objectives.

Many of these classifiers are directly able to handle binary as well as non-binary labels, for example Naive Bayes and Decision Trees. For other methods, such as SVM, that generally learn binary labels, adapted versions have been developed for multiclass learning [1]. As these methods vary fundamentally in their approach to classification and working mechanisms of classifiers are not in the scope of this thesis, the reader is referred to [1] for an overview of different classifiers and an explanation of their classification mechanisms.

Model calibration

Independent of the model characteristic, whether it is a systems biology or classification model, in order to simulate the model derived in the previous step, the numerical values for all parameters need to be determined. The number of these parameters usually increases with the complexity of the model. For systems biology models, assigning appropriate values may be difficult in biological settings as even potentially measurable parameters, like rate constants, might only be known with high uncertainty [165]. For classification models this approach is generally not feasible, as the parameters, in contrast to for example rate constants, do not represent biological functions.

Computational approaches to parameter fitting are primarily based on optimization techniques which minimize an objective function. This objective function usually takes the form of a cost function which describes the difference between the real measurement data and the parameterized model output. Common choices for the cost function

are the Least Squares method

$$L_{\mathrm{LS}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y_i - x_i(\boldsymbol{\theta})\right)^2 \tag{1.6}$$

and Maximum Likelihood (ML) method

$$L_{\mathrm{ML}}(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} p\left(y_i|\boldsymbol{\theta}\right), \tag{1.7}$$

where $p\left(y_i|\theta\right)$ is the probability density of observing the measured data point y_i given the model parameterized with a parameter set $\boldsymbol{\theta}$ defined by $x_i(\boldsymbol{\theta})$. For a systems biology model, x_i is a state variable and may also depend on time and inputs, as in Equation 1.2. For a classification model, x_i refers instead to the model output used to predict the class membership. Applying an ML cost function requires the selection of an error model for the measurement data. The cost function as written above describes a simple case of n unspecified data points. However, depending on the modeling setting, the cost function may include factors for different time points, variables and experimental conditions.

A great variety of calibration methods exists, both for network and classification models. Algorithms to optimize the cost function can be divided according to different criteria, such as local and global methods or deterministic and stochastic methods. Local search methods can be further divided in gradient based methods, for example Newton's algorithm and the Broyden–Fletcher–Goldfarb–Shanno method, and non-gradient based methods, such as Powell's method and generalized pattern search [234].

Local methods converge in general faster to a local or global optimum than global methods. In turn, they may not reach the global optimum depending on the starting point. In order to find the global optimum in the presence of multiple optima, a simple procedure is to use a multistart approach. By selecting initial parameter sets for example through Latin hypercube sampling [155], different regions of the search space can be covered.

Alternatively, global methods search the whole parameter space in order to identify the global or a near global optimum. Global optimization methods can roughly be divided in stochastic and deterministic approaches. Common stochastic optimizer are Genetic Algorithm [100], Particle Swarm Optimization [121] and simulated annealing [125]. These strategies apply probabilistic approaches and are motivated from natural phenomena such as natural selection. They apply randomness in the optimization algorithm to increase the chance of finding the global optima. There are also many deterministic global search methods, the DIRECT algorithm [114] which employs Lipschitzian optimization to find suitable parameter regions for further local investigation.

Local and global optimization techniques are shortly reviewed in [234]. [192] additionally present a broad overview of non-gradient based algorithms.

A special approach to the parameter fitting problem is given by Markov Chain Monte Carlo (MCMC) methods, which are applied in systems biology studies. MCMC methods employ a Bayesian approach by considering all system components in terms of probabilities. In contrast to previously discussed optimization techniques, MCMC methods generate samples from the probability distribution of the parameter space. This allows a global investigation of the parameter values and their uncertainties. Hereby, a Markov chain is constructed that converges towards the probability distribution. The chain starts in general from a set of random points in the parameter space. The algorithm then performs a random walk through the parameter space, while searching for areas with parameter sets improving the data fit. As a result, a global investigation of the parameter values and especially their uncertainties is performed. Various implementations of the MCMC approach exist, of which Metropolis-Hastings [90, 159] is one of the most popular methods. MCMC methods allow for uncertainty analysis of different attributes by sampling from the posterior distribution of the parameter space. The sampling of multiple parameters across probability distributions implies high computational costs, as the cost function needs to be evaluated for each parameter sample drawn.

While MCMC methods can be used to gain further insight in systems biology models, similar concepts also exist for classification problems. As mentioned in the Section Data pre-processing, several extensions with additional functionality of classifiers exist. Robust methods have been developed in order to improve classification results where datasets are infused with outliers by reducing their impact on the learning process. Despite their influence on the classifier accuracy, outlying samples can provide additional and novel insight into data characteristics. Consequently, applying a robust classification method when data might contain outliers is advisable.

There exist several approaches to robust classification, that are often based on the underlying classification method. Additionally, several different strategies can exist for a single classifier. A survey of robust SVMs [211] from 2020 summarized seven different approaches to robustify SVMs, such as a fuzzy or weighted SVM. Similarly, different alterations to induce robustness exist for other classifiers as well and ongoing research adds new variants regularly.

In a similar fashion to robust classification methods, classifiers have also been extended to allow for feature selection. These so-called sparse classifiers learn a subset of features that is sufficient to differentiate between classes. This approach is of interest, when the number of features exceeds the number of samples and the features are of further research interest. Sparse classification is applied to infer a set of important features which can provide further insight in the studied systems. Furthermore, sparse classification methods are imperative in data settings with a large amount of features but comparatively few samples. Sparse classification approaches have been applied to a variety of types of datasets and problems: [187] classify Alzheimer's Disease from data of magnetic resonance imaging, [154] apply a sparse classification approach to flow cytometry data of protein marker expression to predict leukemia and [3] use microRNA data of endometrial cancer patients to find molecular markers for the prediction of lymph node metastasis.

As mentioned in the Section Data pre-processing, sparse classification methods include embedded feature selection methods. The methodological approach to include feature selection in the classifi-

cation algorithm generally depends on the classifier of interest [1]. A valid approach in sparse methods is the application of weights as penalty on the features in the optimization step in order to force feature coefficients towards zero while fitting the model to the data [117]. Therein, feature coefficients describe the influence of a feature on the classification and penalties are referred to as regularization terms. Such regularization terms are generally applied to linear classifiers, for example linear regression or SVMs. Standard approaches are Ridge Regression, Lasso and Elastic Net. Considering a linear classification model

$$y_i = f(\mathbf{X}_i, \boldsymbol{\beta}) + \varepsilon_i, \tag{1.8}$$

where y_i, $i \in \{1, \ldots, N\}$ represents the class membership of sample i, $f(\mathbf{X}_i, \boldsymbol{\beta})$ is a function that relates the measurement matrix $\mathbf{X}$ and the feature coefficients $\boldsymbol{\beta}$ to the class labels $\mathbf{y}$ and ε_i denotes a noise term. In multivariable linear regression, $f(\mathbf{X}_i, \boldsymbol{\beta})$ takes the form $\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ with p features. In order to find the optimal set of parameters $\boldsymbol{\beta}$, the error between all labels $\mathbf{y}$ and the mapping function $f(\mathbf{X}, \boldsymbol{\beta})$ is minimized. Least Squares optimization with its variants and ML estimation constitute common approaches to parameter inference. For a Ridge Regression regularization [97], an l_2-norm penalty $\lambda\|\boldsymbol{\beta}\|_2^2$ is added to the cost function that enforces the coefficients to shrink toward zero. λ hereby defines the strength of the penalty term. A Lasso penalty [230] is incorporated by adding an l_1-norm term $\lambda\|\boldsymbol{\beta}\|_1$ instead. In contrast to the l_2-norm, the l_1-norm penalty allows several coefficients to become zero. In both cases, λ defines the strength of the penalty term. Combining both approaches, the Elastic Net [251] defines the penalty term as $\lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2$. These approaches can usually be used to enhance linear classification approaches such as different variants of regression and SVM. Further implementations of extensions have been developed for linear as well as other classifiers, for example sparse Naive Bayes classifiers [14, 27], sparse Decision Trees [104, 213], and sparse Neural Networks [6, 215].

As outliers as well as an imbalance of the feature to sample ratio can appear in the same dataset, classifiers that combine both robustness and sparsity have been developed as well. And again, a variety of

strategies can exist for a single classification model. For example, robust and sparse versions of SVM have been proposed by [146] and [108]. Likewise for logistic regression, [31] and [137] introduced new methods that combine robustness and sparsity.

The great variety of classification methods and their different implementations signify a big challenge to a modeler trying to select a classifier. Determining the best algorithm requires knowledge about the available methods and the dataset. Considerations towards the data size to determine whether a sparse method is necessary or whether the data is possibly outlier contaminated can help narrow down the number of appropriate methods. If outliers are expected to be present in the dataset, methods may also be chosen differently depending on the type of outliers. In [31], the robust part of their robust and sparse logistic regression method is concentrated on mislabeled microarrays. Accordingly, their proposed method includes a label flipping probability to detect potential outliers to robustify the classification. However, in other scenarios with outliers in the feature space, where only certain feature values of a sample are outlying, this approach may not be able to detect the outliers.

In summary, as no classification method performs best on all datasets and specifics, such as the presence, type, and amount of outliers, may not be known, selecting the optimal classifier presents a difficult task.

Model validation

The next step after model calibration is the assessment of the results. A first impression can be gained by considering the quality of the data fit when comparing model simulations with real data. However, on one hand this is a subjective way of judging the derived model and on the other hand it does not provide information on the predictive power of the model. For this reason, several model validation techniques have been developed. These methods are generally based on a separation of the available data, such that one part is used for model calibration and the remaining part to test the predictive power of the calibrated model.

[1] separate the classifier evaluation in two parts which they term "methodological issues" and "quantification issues". Even though [1] describe methods for classification problems, the fundamental approaches are also valid in systems biology. The methodological part represents hereby the choice of a data partitioning scheme to create a training set for model calibration and a test set for model evaluation. The quantification part contains the choice of a numerical measure to describe the performance of the model given a specific data partitioning. Both parts will be inspected in the following.

Methodological part

Cross validation is an approach widely used for classification problems [21], while hold-out validation constitutes a basic approach for ODE based systems biology models [87]. Nevertheless, both techniques can be applied to both scenarios.

Cross validation In cross validation, the dataset is partitioned in m equally sized subsets. $m-1$ such subsets are used as training set for model calibration and the remaining subset constitutes the test set for evaluation. Repeating the calibration and evaluation step such that each of the m subsets is used once as the test set, an average performance metric can be calculated over all test sets. A typical choice for the number of subsets m is ten. Choosing m equal to the number of data points results in the special case of leave-one-out cross validation. A short introduction to cross validation is given in [1, 87] and more information on its variants can be found in [21].

As each data point is contained exactly once in the m test scenarios, the overall accuracy of the cross validation approach may give a pessimistic estimate of the model accuracy. Additionally, this validation method is computationally expensive for large datasets, as the calibration procedure needs to be repeated m times. However, even for small datasets, this method may not be realizable. In systems biology, where only few replicates per experiment may be available, a partitioning can be difficult as the training data must be able to represent all model features such as experimental conditions.

Hold out validation Hold out validation is fundamentally similar to cross validation. However the data is divided in two sets, one training and one test dataset and usually more data points are assigned to the training set than the test set. The model is then calibrated using the training set and evaluated based on the performance on the test dataset. This procedure can be repeated to calculate an average performance metric. Further information on this method can be found in [21] and [87].

For large datasets, applying repeated hold out validation results again in a high computational cost. Conversely, the results of using a single set of training and test data can be highly dependent on the specific partitioning. Considering small data, similar to cross validation, all biological features included in the model need to be represented in the training and test data. In this context, [87] demonstrated the dependence of the validation decision on the chosen partitioning for systems biology models.

Bootstrap Another approach to model validation is given by the bootstrap method. Hereby, a training set is constructed by sampling with replacement from the original dataset. As the training set contains as many data points as the original dataset, it may contain duplicates. The model is calibrated on the training set and evaluated on the test set given by the original full dataset. A short introduction to bootstrap validation is given in chapter 10 in [1].

As a consequence of the sampling procedure, it is possible that even for repeated bootstrap sampling, data points may never be selected for the training or the test dataset. Another drawback of the bootstrap approach is that the classifier will always achieve a hundred percent accuracy for the data points in the training set. Consequently, the performance estimate will be optimistic with regard to the true performance.

In general, as highlighted in [80], model validation in systems biology is highly dependent on the purpose of the modeling study. It was already mentioned previously that a model can be used for different purposes. As this holds true for systems biology as well as

classification models, validation techniques may need to be carefully selected depending on the specific application and available data.

Quantification part

In the next step, in order to quantify the result of the validation methods, different measures can be considered to evaluate the model performance. As methods for systems biology and classification models differ, approaches will be discussed separately.

Systems Biology In systems biology, the quantification can be conducted using the metric applied for parameter optimization, for example the mean squared error. If the error of the calibrated model is also low for the test data, this supports the model hypothesis, as the model is able to reflect also previously unseen data. In turn, if the error in the test data is large, the model may suffer from overfitting and a revision of the model setup might become necessary. As this validation approach is time-consuming and requires careful selection of training and test data, other approaches are commonly preferred. Nevertheless, in [164] a bootstrap approach was used to evaluate the robustness of a transition graph calibrated from a boolean network model. Also, in [141] leave one out cross validation was performed for an ODE model.

As modeling studies in systems biology can have a variety of purposes, validation approaches can take different forms [80]. In case of hypothesis testing, the most common approach is most likely validation via prediction. Hereby, the calibrated model is used to predict system behavior under new conditions. The simulation results are then compared to corresponding experimental data that was not used for model calibration. Aiming to predict system behavior under new conditions, careful selection of the experiment and its implementation in the model is important, as the model needs to be able to mirror the new conditions. In [185], the authors apply two prediction scenarios to validate their model of drug response in gastric cancer. The model combines three drug specific signaling pathways that describe intracellular signaling in response to receptor signaling. Therein, the two

validation scenarios nicely demonstrate the possibilities and limits of validation via prediction. For the first scenario, the authors utilize published experimental data involving receptor inhibition. This experimental condition was not originally part of the model and the authors had to extend their model and utilize published parameters for simulations. Even though this approach allows for new insight in the predictive power of a model, it requires the availability and compatibility of parameter values for the additional experimental setups. For the second scenario, long time cell behavior was considered. This approach requires no model adjustment. However, depending on the system under study and the examined time frame, additional mechanisms may become involved and influence prediction accuracy. This shows the complexity of validating a model by predicting new experimental data.

Classification In classification, quantification methods are generally based on the classifier performance with respect to a specified class, for example assigned with label 1. The other class or classes are summarized in one group with label 0, simplifying the analysis to that of binary classification. In disease classification the group of interest could be the group of sick patients, or the group with a specific medical intervention when considering drug response. Classification results are then separated in classifications that are

- true positive (TP): correctly assigned samples of class 1,
- true negative (TN): correctly assigned samples of class 0,
- false positive (FP): samples falsely assigned to class 1, or
- false negative (FN): samples falsely assigned to class 0.

A summary of these results can be presented in a contingency table as shown in Figure 1.4. This contingency table summarizes how well predicted class memberships match the actual class.

Separating the assessment for the two classes is important, as false assignments of either type can have different significance and impact. For example, falsely classifying a sick patient as healthy, an FN, impedes timely treatment and may propose a risk to the

		Predicted Class: 1	Predicted Class: 0
Actual Class	1	TP	FN
	0	FP	TN

Figure 1.4.: Contingency table summarizing the number of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) classifications.

patient, whereas an FP, in this case classifying a healthy subject as sick, subsequent examinations might lead to potentially harmful and costly treatment.

Prevalent statistics that assess different aspects of the classification results are

- Sensitivity or Recall or True Positive Rate (TPR):

$$TPR = \frac{TP}{TP + FN},$$

- Specificity or True Negative Rate (TNR):

$$TNR = \frac{TN}{TN + FP},$$

- False Positive Rate (FPR):

$$FPR = 1 - TNR = \frac{FP}{TN + FP},$$

- Precision or Positive Predictive Value:

$$PPV = \frac{TP}{TP + FP},$$

- Accuracy:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}.$$

A detailed explanation of these and additional statistics can be found in [225]. In general, these criteria are assessed in pairs as single statistics may be misleading [195]. A perfect TPR is for example possible by assigning the class label 1 to all samples. While this results in $TPR = 1$, the specificity is at the same time reduced to zero ($TNR = 0$).

The Receiver Operating Characteristic (ROC) is a common evaluation criterion combining two of the above statistics. It is presented as a scatter plot of TPR against FPR. One point in this plot depicts the performance of one classifier given one specific dataset and threshold. A ROC curve is obtained by varying parameters of the classifier. ROC curves can be used to assess the classifier performance in comparison to other classifiers, under changing parameter settings and given different datasets [195]. Hereby, an optimal classifier would achieve $TPR = 1$ and $FPR = 0$. Figure 1.5A shows an example for a ROC curve as well as the position of an optimal classifier (red dot). The closer a classifier is to this optimal point, the better is its performance.

An alternative method to assess the classifier performance using two classifier statistics is the precision recall diagram. Here, the precision is plotted against the TPR. This method thus relates the number of TP labels to the total number of samples with real class 1 and with assigned class 1. A high TPR can in general be achieved for a low precision as an approach which favors assigning class 1 creates high numbers of FP labels (lowering the precision) and low numbers of FN labels (increasing the TPR). Analogously, a high precision can usually be attained for a low TPR. A good classification method

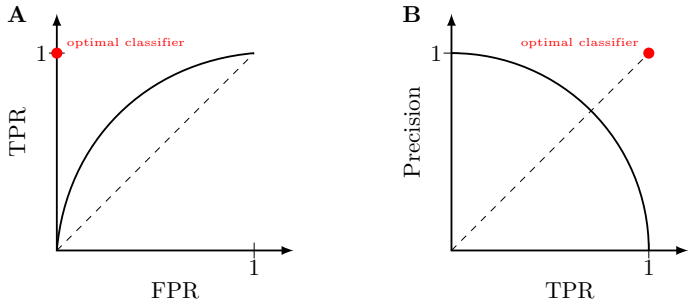


Figure 1.5.: Example results of a ROC curve (A) and a precision recall curve (B) constructed by assessing the classifier performance for different settings (curved black line) and an optimal classifier (red dot).

achieves a high TPR and a high precision which is assessed from the point where the precision recall curve intersects the main diagonal (Figure 1.5B). More information on ROC and precision recall curves as well as their connection can be found in [49].

As mentioned before, the best choice of evaluation methods depends on the data and the purpose of the study. Using standard models, for example to analyze outliers, thus requires other validation approaches than a new classification model.

Model analysis

There are a multitude of different analyses for any given model. Which ones are relevant for a specific system is determined by what questions we want to answer with the model. Whether we can then get our answers depends on the capabilities and predictive power of the model, available analysis techniques and our own inventive talent. As the goals of systems biology and classification modeling studies are different in general, the discussion on model analysis will be divided.

Systems Biology

In the simplest case, the goal of a systems biological study is to test a model hypothesis. The question whether a model hypothesis must be rejected, can in most cases be answered after the validation step. In the next step, the model can then be used to make predictions of system behavior under varying conditions that may be difficult or expensive to test experimentally. Through this, it is possible to analyze the role of model components, such as feedback connections, as was performed in the study described in chapter 2.

As predictions are dependent on the calibrated parameter set, their reliability also depends on the identifiability of the model parameters. However, how well parameters can be determined is influenced by several factors, such as the possibility to measure all components of the model or the amount of available data in general. This may lead to non-identifiable parameters, which roughly speaking means that different parameter values can result in the same model output. Consequently, this also implies that the objective function value does not differ for these values. [188] proposed an approach involving the profile likelihood in order to examine parameter identifiability. The profile likelihood for a parameter estimate $\hat{\theta}_i$ is calculated by re-optimizing the objective function for increased and decreased fixed values of $\hat{\theta}_i$. This is repeated for all model parameters. The results can then be visualized by plotting the objective function values against the fixed parameter values for each parameter. From the form of the profile likelihood, [188] differentiate three types of parameters:

- Identifiable parameters: the profile likelihood surpasses a specified threshold for decreasing as well as for increasing values of the parameter θ_i.
- Structurally non-identifiable parameters: the profile likelihood remains flat for increasing and decreasing parameter values of θ_i. Structural non-identifiability arises from functional relations between parameters. Changes in the parameter value of θ_i can thus be compensated by related parameters in order to achieve the same objective function value.

- Practical non-identifiable parameters: the profile likelihood pertains an optimum, but the specified threshold cannot be surpassed for either increasing or decreasing values of θ_i. Practical non-identifiability arises when the amount and quality of the experimental data for model calibration is deficient.

Analyzing and characterizing identifiability issues with this method can give information on opportunities to improve the accuracy of model predictions. Identifying correlated parameters that induce structural non-identifiability can be utilized to perform model reduction by replacing a non-identifiable parameter with its functional relation to other parameters. On the other hand, investigating the trajectories along a practically non-identifiable parameter can be used for experimental planning. Trajectory phases with high uncertainty can be used to suggest additional measurements in order to improve parameter identification. Further information on non-identifiability, specifics on the choice of the threshold and details on the profile likelihood method are give in [188].

Another method to evaluate the influence of parameter selection on model simulations is sensitivity analysis. While several methods exist, the common idea is to observe the model output under varying parameter values. Sensitivity based approaches can be divided in local and global methods. Local methods provide sensitivity coefficients for perturbations of one parameter. On the other hand, in global approaches all parameters are varied simultaneously. Different approaches for sensitivity analysis are reviewed in [222, chapter 2].

Depending on the chosen calibration method, direct analysis of parameter uncertainty is possible. If a Monte Carlo scheme has been applied to sample the parameter space, in general a large number of parameter sets is available that represent samples of the posterior distribution of the parameters. These parameter sets can be employed to analyze 2D correlations by exploiting correlation coefficients or graphical approaches such as scatter plots. Additionally, by simulating the model for these samples, uncertainties for further quantities of interest can be investigated. This may include the analysis of system behavior under varying constraints. The calibrated model can

be simulated under different experimental conditions, for example varying inputs, silencing experiments, or inhibiting factors by manipulating corresponding parameters or variables in the model. This allows for the investigation of system response under these various conditions. As biological experiments are time-consuming, potentially expensive, and not all variables of interest may be measurable, simulation studies provide a good alternative allowing for relatively quick implementation and execution of research questions. Nevertheless, while this approach is capable of reflecting perturbations to model variables, investigating the influence of factors not directly included in the model is a difficult task and requires careful consideration. This problem is similar to the difficulties encountered when validating a model by predicting new experiments. Aside from varying model conditions, model simulations can also support the examination of model components and mechanisms of system behavior. For example, in [40] the authors develop a model for the crosstalk between two survival signaling pathways in human breast cancer cells. Breast cancer cells may switch between the two survival signaling pathways as a response to cancer therapy and thus develop a drug resistance. The analysis of the model in [40] provides deeper insight in the dynamic behavior of breast cancer cells with focus on the mechanisms underlying the switch of the survival signaling pathway.

Considering the different approaches and purposes of model analysis, its influence on other modeling steps becomes evident. Depending on the requirements of the planned analysis, a suitable calibration method needs to be selected. After performing the intended analysis, results can be used for experiment design, thus influencing the data preparation and consequently the modeling step if new experimental conditions are involved. Finally, evaluation of parameter correlations can be used for model reduction, which again concerns the modeling step.

Classification

Similar to systems biology, there are different ways that a calibrated classification model can be used. A straight forward application is to

apply the classifier in order to predict labels for new samples with unknown class membership. However, classification models can do more which will be the topic in the next paragraphs.

Outlier detection The topic of outliers has been considered in previous sections, first as a possible investigation target in the preprocessing step and again in the calibration step when selecting an appropriate method. As mentioned before, outliers can be a research target on their own. Samples identified as outliers could contain a false label or show conspicuous values in one or more features. The second group of outliers may hint to the existence of another class and compel further analysis. Applying a robust classification method allows to capitalize on the intrinsic outlier selection. However, there is no single method that can reliably identify outliers in any dataset. Given an arbitrary dataset, it is difficult to know which of the numerous methods will perform best. In this context, an ensemble approach has been proposed in [149] that combines different classifier results. The authors train three sparse classifiers on an RNA-Seq dataset. Outliers are determined by a consensus approach that summarizes and evaluates the residual measures of each classifier to rank samples in order of their outlierness. By aggregating information from different classifiers, this approach can compensate shortcomings of a single method for a specific dataset. On the other hand, a study by [223] concludes that the ensemble approach proposed by [149] yields high accuracy in outlier discovery for small percentages of outliers, but could not detect many outliers in data settings with a high outlier percentage. Furthermore, the methods applied in the ensemble in [149] are similar in their mechanisms, as two of the methods are based on partial least squares optimization. In turn, these methods may have similar advantages and disadvantages with regard to their classification performance and subsequent outlier analysis. Thus, employing more diverse and additionally robust classifiers in the ensemble approach may improve results.

Feature selection In the same way as outlier detection, feature selection has been discussed in previous sections. The presence, type,

and amount of outliers as well as the size of the dataset in general or correlations among the features may influence the classification results and subsequently the feature selection. As many sparse classification methods exist and knowledge on data specifics such as the amount of outliers may be limited, it is difficult to choose the most appropriate method for a given dataset. Additionally, depending on the sparse classifier and its settings, a large number of features may still be selected which may make further analysis of all of these selected features infeasible. Under these constraints, [149] utilize the ensemble approach once more. The final group of features of interest is hereby obtained by selecting features marked as important by all classifiers. Thus, the number of selected features can never surpass the number of features selected by any of the methods while also increasing the confidence in the selection. In the study by [223] it was shown that the ensemble approach had the best feature selection accuracy among the three tested methods. However, the sensitivity decreased for larger portions of outliers. As results in the study could be improved by pre-processing the data and removing outliers, the application of a robust ensemble may prove beneficial.

Further analysis Given the detected outliers and selected features, further examination can support the validity of the results and provide additional insight into the data. Visualization of different aspects of the data is therefore an important tool. This includes for example the comparison of feature values of outliers with their nominal class or the evaluation of correlations between features.

In the absence of standards, the choice of methods is determined by the modeler and the dataset under study. Thus, an overview over all approaches is not feasible. For example, the studies of both [249] and [206] investigate potential gene markers that differentiate triple from non-triple negative breast cancer. [249] examined the prognostic value of their findings by analyzing the survival time of patients with respect to the gene expression of the selected genes. For this approach, [249] employed Kaplan-Meier curves which are estimators of the probability that a certain event does not take place in a specific time frame. In most cases, these events are the

onset of a disease or death and the Kaplan-Meier curve describes the cumulative probability to remain disease-free or the survival probability, respectively. Analysis of the Kaplan-Meier curves in [249] indeed found a subset of genes to be associated with overall survival. In contrast, [206] examined the correlations among selected genes for the two classes. By visualizing a network based on gene correlations, they inferred differences between gene correlations for the two groups. As these two studies show, analysis techniques are dependent on the research question and the modelers' preferences.

In summary, similar to systems biology, the analysis of classification models can also influence other steps of the modeling procedure. An interest in outliers and important features may determine the choice of the calibration method. At the same time, certain analysis techniques may require additional information, such as survival times for Kaplan-Meier curves. These need to be prepared in the data pre-processing step. The complete modeling process starting from data preparation to the final model analysis still remains in the hands of the modeler. Decisions for certain methods within each of the introduced modeling steps always need to be adapted to the current available data to achieve a *good* model.

1.3. Layout of subsequent chapters

Each step requires the choice of an appropriate method, depending on the possible problems described above. In the following three chapters, I present three projects and their approaches to the five steps. The projects cover a range of aforementioned problems under different study aims and data conditions. Chapter 2 and 3 pursue a systems biology approach, whereas Chapter 4 applies a classification approach.

The first project, presented in Chapter 2, considers a well studied signaling pathway, namely the mitogen-activated protein kinase (MAPK) pathway. As the model structure has been studied before, focus of this project lay on the analysis of the feedback mechanism involved in MAPK signaling. Even though the construction of a

satisfactory mathematical model constituted a large portion of the work in this project, the main focus is on model analysis and the corresponding choice of the model calibration method.

The second project, presented in Chapter 3, considers the regulation mechanisms of the protein deleted in liver cancer 1 (DLC1). Prior work on this signaling pathway is limited and components as well as their interaction with DLC1 have been mostly unknown. Consequently, the project focused on the modeling step and the model validation.

In the third project, presented in Chapter 4, the differentiation of a specific type of breast cancer from other breast cancer types is studied. The work is focused on the identification of genes that can be used as potential biomarkers as well as the extraction of outlying samples. As such, at the center of this project is the model analysis. Much importance is also given to the validation of results using different approaches.

Each project constitutes one chapter consisting of two subchapters. The first subchapter includes the project results as published or submitted to a peer-reviewed journal as well as a description of my contribution to the work. The second subchapter then elaborates about the five steps of modeling under the given conditions such as the aim of the study and the available data. These subchapters focus on project specific circumstances and the resulting choices or development of methods required to reach the project goal.

Finally in Chapter 5, the methodological choices and advances are summarized and discussed.

1.4. Notes on the cumulative part

The manuscript content, figures and tables are presented as published or submitted. Following small modifications were performed to embed the manuscripts in the thesis:

- The formatting style of the thesis has been applied to all articles.
- The bibliographies of all articles have been combined and are listed at the end of the thesis.

- Supplementary material and additional information of all articles are presented in the Appendix. References in the articles have been adjusted to link to the appendix.
- The contents list of the Supporting Information of Chapter 3 was integrated in the contents list of this thesis.
- In Chapter 3, the header *Introduction* was added for consistency and readability.
- A few typos have been corrected.

Chapter 2.

Sampling-based Bayesian approaches reveal the importance of quasi-bistable behavior in cellular decision processes on the example of the MAPK signaling pathway in PC-12 cell lines

2.1. Published manuscript and contributions

This chapter corresponds to the following contribution:

> *A. Jensch, C. Thomaseth, and N. E. Radde. "Sampling-based Bayesian approaches reveal the importance of quasi-bistable behavior in cellular decision processes on the example of the MAPK signaling pathway in PC-12 cell lines". In:* BMC Syst Biol *11.1 (2017), p. 11*

2.1.1. Abstract

Background: Positive and negative feedback loops are ubiquitous motifs in biochemical signaling pathways. The mitogen-activated protein kinase (MAPK) pathway module is part of many distinct signaling networks and comprises several of these motifs, whose functioning depends on the cell line at hand and on the particular context.

The maintenance of specificity of the response of the MAPK module to distinct stimuli has become a key paradigm especially in PC-12 cells, where the same module leads to different cell fates, depending on the stimulating growth factor.

This cell fate is regulated by differences in the ERK (MAPK) activation profile, which shows a transient response upon stimulation with EGF, while the response is sustained in case of NGF. This behavior was explained by different effective network topologies. It is widely believed that this sustained response requires a bistable system.

Results: In this study we present a sampling-based Bayesian model analysis on a dataset, in which PC-12 cells have been stimulated with different growth factors. This is combined with novel analysis methods to investigate the role of feedback interconnections to shape ERK response. Results strongly suggest that, besides bistability, an additional effect called quasi-bistability can contribute to explain the observed responses of the system to different stimuli. Quasi-bistability is the ability of a monostable system to maintain two distinct states over a long time period upon a transient signal, which is also related to positive feedback, but cannot be detected by standard steady state analysis methods.

Conclusions: Although applied on a specific example, our framework is generic enough to be also relevant for other regulatory network modeling studies that comprise positive feedback to explain cellular decision-making processes. Overall, this study advices to focus not only on steady states, but also to take transient behavior into account in the analysis.

2.1.2. Background

Feedback regulations are ubiquitous network motifs in all kinds of molecular interaction networks, such as for example metabolic networks, regulatory modules or signaling networks [7]. The role of single positive and negative feedback is well-characterized also from a theoretical point of view. Negative feedback, which counteracts external perturbations, can cause oscillating behavior, but also has

a stabilizing effect, implies robustness of cell states to internal and external perturbations [142], and plays a major role in maintaining homeostasis (see e.g. [61, 78, 226]). Furthermore, it can accelerate the response to a transient signal. By contrast, positive feedback amplifies an external perturbation or signal, which can cause multi-stability, hysteresis and memory effects or switch-like behavior. Positive feedback is omnipresent in cellular decision processes, in which these phenomena arise. It can also produce ultrasensitivity and prolong the response to a transient external signal [8, 61, 201].

In this study we investigate the role of feedback regulation for proper signal processing by a case study on the well-known mitogen-activated protein kinase (MAPK) signaling pathway. This pathway is an evolutionary conserved signaling module, which is involved in many essential cellular processes such as proliferation, survival or differentiation [79, 129, 130, 172]. It is de-regulated in various diseases and represents an important drug target [172]. The pathway module consists of a cascade of phosphorylation events, leading to the activation of ERK, which targets more than 80 substrates in the nucleus and the cytosol. It is integrated into multiple signaling pathways and shows a variety of different responses depending on the stimulus and the cell-type specific context [129, 172, 199]. Specificity of the cellular response is tightly related to distinct time courses of active ERK upon different stimuli, in particular amplitude and duration of the signal response [172, 196, 199]. A well-studied paradigm for such a context-specific response is the different behaviors of PC-12 cells upon stimulation with epidermal growth factors (EGF) and neural growth factors (NGF) [36, 199]. Cells stimulated with NGF show sustained activation of ERK, accompanied by a translocation of ERK into the nucleus, which eventually initiates cell differentiation. In contrast, ERK activity is transient and mainly restricted to the cytosol upon stimulation with EGF, which in turn triggers proliferation.

The pathway module is well-characterized experimentally and from a modeling point of view (for reviews see e.g. [122, 129, 130, 140, 172, 233]). Starting with the early work of Huang and Ferrell [105], many models of different complexity and with different foci have been suggested in the meantime [2, 62, 150, 172, 210]. In particular,

quite a number of studies focus on modeling and understanding the mechanisms behind the distinct responses upon EGF and NGF stimulation in PC-12 cells [13, 25, 36, 60, 196, 199, 200].

It is commonly believed and well-described that a system which shows a sustained response to a transient signal, such as PC-12 cells upon NGF stimulation, is a bistable system [13, 25, 60, 122, 144, 180, 210, 212, 245]. Hence modeling of this phenomenon usually focuses on the investigation of the bistability properties of respective models, and advanced methods have been developed tailored to the investigation of steady states in these models (see e.g. [13, 26, 184]).

In this study we turn our attention to a phenomenon called quasi-bistability and its role in the regulation of the MAPK module for cellular decision-making. Quasi-bistability is the ability of a monostable system to maintain a second steady state for a long period of time upon a transient stimulus [162]. It is also related to positive feedback, but less well investigated and understood. Using a dynamic modeling approach and a dataset of the MAPK module in PC-12 cell lines, the system is analyzed via Bayesian sampling techniques. Mechanisms behind sustained ERK responses are investigated by a combination of steady state analysis methods and novel methods that also allow to investigate time scales of transient behavior.

2.1.3. Methods

2.1.3.1. Experimental data used for model calibration

For our modeling study we used a dataset described in [199], where PC-12 cell lines were stimulated with EGF and NGF, and phosphorylation of the proteins in the cascade was measured via Western blotting and flow cytometry. For model calibration we used the data shown in Figs 1 and S1b in [199]. This dataset contains data from control experiments, in which cells were stimulated with 100 ng/ml EGF or 50 ng/ml NGF, and measurements from RNA interference experiments.

In the control experiments the dynamic response of the system upon stimulation was measured in terms of phosphorylation levels of Raf (pRaf), MEK (ppMEK) and ERK (ppERK). In the following we

EGF stimulation			
Time [min]	$\tilde{v}_1(t)$	$\tilde{v}_2(t)$	$\tilde{v}_3(t)$
5	0.79	0.76	0.76
10	0.47	0.4	0.33
15	0.34	0.3	0.10
30	0.15	0.14	0.03
60	0.09	0.09	0.02

NGF stimulation			
Time [min]	$\tilde{v}_1(t)$	$\tilde{v}_2(t)$	$\tilde{v}_3(t)$
5	0.78	0.75	0.80
10	0.31	0.34	0.55
15	0.39	0.42	0.30
30	0.26	0.35	0.30
60	0.35	0.30	0.20

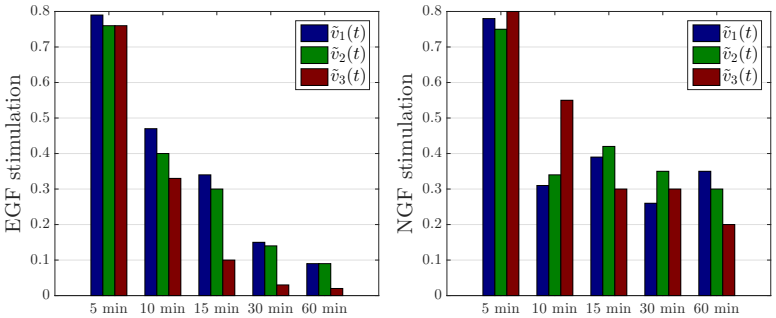


Figure 2.1.: Activities of Raf, MEK and ERK after stimulation. Scaled activities of Raf, MEK and ERK measured by polychromatic flow cytometry (by visual inspection from Fig S1b in [199]).

will refer to the active states of the proteins by using the following variables:

$$v_1 = \text{pRaf}$$
$$v_2 = \text{ppMEK}$$
$$v_3 = \text{ppERK}.$$

We used data from flow cytometry experiments (Fig S1b in [199]) as reference for model calibration, since all proteins were quantified in this experiment. Extracted values are illustrated in Fig 2.1 and show a transient signal response in case of stimulation with EGF, and a sustained response after stimulation with NGF. The quantified values are a scaled version of the quantities v_i, and are defined as $\tilde{v}_i$.

In the siRNA experiments, Raf, MEK and ERK were consecutively

downregulated. These data were used in [199] to analyze the network topology via Modular Response Analysis [123]. In this analysis, global response coefficients R_{ij}, $i, j = 1, 2, 3$ were calculated from the Western blot signals (Fig S1c in [199]) via

$$R_{ij} = 2\frac{\partial \ln(v_i)}{\partial \ln(p_j)} \approx 2\frac{\bar{v}_i^{(s_j)} - \bar{v}_i^{(c)}}{\bar{v}_i^{(s_j)} + \bar{v}_i^{(c)}}. \tag{2.1}$$

The variables $\bar{v}_i^{(c)}$ and $\bar{v}_i^{(s_j)}$ denote the steady state concentrations of variable v_i before and after perturbation p_j, i.e. silencing of component j, respectively.

Equation (2.1) can be resolved for $\bar{v}_i^{(s_j)}/\bar{v}_i^{(c)}$,

$$\frac{\bar{v}_i^{(s_j)}}{\bar{v}_i^{(c)}} = \frac{2 + R_{ij}}{2 - R_{ij}}, \tag{2.2}$$

which gives the concentration change of component i relative to the control experiment in response to silencing of component j.

Values of the response coefficients of four replicates for silencing of each protein are provided in Table 1 in Fig S1d in [199]. These data were used to calculate empirical means and standard deviations, as illustrated in Fig 2.2, together with the respective relative changes of protein concentrations after silencing.

Time points were set to 5 min after EGF stimulation, the time about which the maximum of the signal response is reached in the control experiments, which is assumed to be close to a steady state condition. In case of NGF, global response coefficients are given at 5 and 15 min after stimulation. These two time points correspond to the times at which the maximum of the signal response was reached and at which the system seems to have reached the new steady state. In [199], these coefficients were used to extract the network structure based on the so-called local response coefficients. This analysis indicates positive feedback from ERK to Raf upon NGF stimulation and negative feedback when stimulated with EGF. This result will be taken into account in our modeling approach.

5 min after EGF stimulation						
j	$\widehat{\mathbb{E}}(R_{1j})$	$\widehat{\sigma}(R_{1j})$	$\widehat{\mathbb{E}}(R_{2j})$	$\widehat{\sigma}(R_{2j})$	$\widehat{\mathbb{E}}(R_{3j})$	$\widehat{\sigma}(R_{3j})$
1 (siRaf)	−0.6692	0.1913	−0.3312	0.3434	−0.4698	0.4684
2 (siMEK)	0.3727	0.3376	−0.4780	0.2923	−0.2985	0.2377
3 (siERK)	0.1525	0.1688	0.4970	0.3427	−0.7271	0.4068
5 min after NGF stimulation						
j	$\widehat{\mathbb{E}}(R_{1j})$	$\widehat{\sigma}(R_{1j})$	$\widehat{\mathbb{E}}(R_{2j})$	$\widehat{\sigma}(R_{2j})$	$\widehat{\mathbb{E}}(R_{3j})$	$\widehat{\sigma}(R_{3j})$
1 (siRaf)	−0.5600	0.0455	−0.3459	0.4273	−0.3869	0.4456
2 (siMEK)	−0.1314	0.1521	−0.2909	0.2268	−0.3295	0.3927
3 (siERK)	−0.1466	0.0500	0.2251	0.1456	−0.6345	0.2845
15 min after NGF stimulation						
j	$\widehat{\mathbb{E}}(R_{1j})$	$\widehat{\sigma}(R_{1j})$	$\widehat{\mathbb{E}}(R_{2j})$	$\widehat{\sigma}(R_{2j})$	$\widehat{\mathbb{E}}(R_{3j})$	$\widehat{\sigma}(R_{3j})$
1 (siRaf)	−0.7762	0.3307	−0.4829	0.4982	−0.8150	0.6924
2 (siMEK)	0.0787	0.2218	−0.2891	0.4811	−0.3451	0.4526
3 (siERK)	−0.4154	0.3704	0.4514	0.4218	−1.0215	0.5350

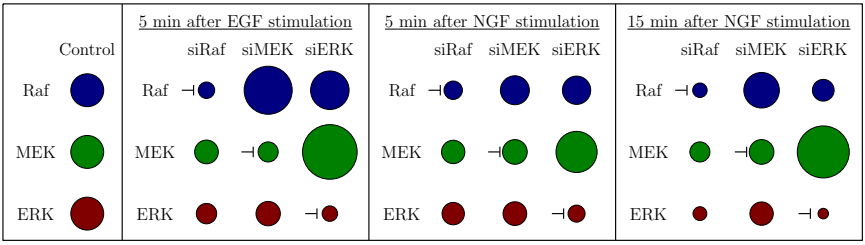
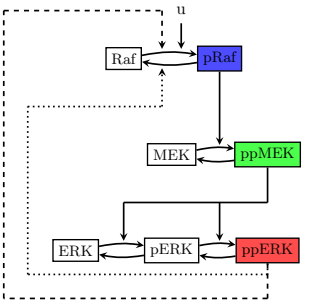


Figure 2.2.: Data from modular response analysis. Table. Means and standard deviations of the global response coefficients extracted from the silencing experiments via modular response analysis. These were calculated from replicates in Table S1d in [199]. Figure. Illustration of respective changes in protein concentrations in response to silencing relative to the control experiments (without silencing).



B

$$\dot{\text{pRaf}} = \mathbf{k_1^+} \cdot (\text{Raf}_{\text{TOT}} \cdot s_1 - \text{pRaf}) \cdot u - \mathbf{k_1^-} \cdot \text{pRaf} + \\ + fn \cdot [-\mathbf{k_{Fn}} \cdot \text{ppERK} \cdot \text{pRaf}] + \\ + fp \cdot \left[\mathbf{k_{Fp}} \cdot \frac{\text{ppERK}^5}{\text{ppERK}^5 + \mathbf{g}^5} \cdot (\text{Raf}_{\text{TOT}} \cdot s_1 - \text{pRaf})\right]$$

$$\dot{\text{ppMEK}} = \mathbf{k_2^+} \cdot (\text{MEK}_{\text{TOT}} \cdot s_2 - \text{ppMEK}) \cdot \text{pRaf} - \mathbf{k_2^-} \cdot \text{ppMEK}$$

$$\dot{\text{pERK}} = \mathbf{k_3^+} \cdot (\text{ERK}_{\text{TOT}} \cdot s_3 - \text{pERK} - \text{ppERK}) \cdot \text{ppMEK} + \mathbf{k_4^-} \cdot \text{ppERK} \\ - \mathbf{k_3^-} \cdot \text{pERK} - \mathbf{k_4^+} \cdot \text{pERK} \cdot \text{ppMEK}$$

$$\dot{\text{ppERK}} = \mathbf{k_4^+} \cdot \text{pERK} \cdot \text{ppMEK} - \mathbf{k_4^-} \cdot \text{ppERK}$$

$$u(t) = 1 - \frac{t^3}{(t^3 + \mathbf{K}^3)}$$

$$\theta = (k_1^+, k_2^+, k_3^+, k_4^+, k_1^-, k_2^-, k_3^-, k_4^-, k_{Fn}, k_{Fp}, g, K)$$

Figure 2.3.: Model structure of the MAPK module. A. Reaction scheme of the MAPK module. Upon addition of growth factors, Raf, MEK and ERK are successively activated in a phosphorylation cascade. Different feedback topologies are assumed to shape context dependent ERK response: Effective negative feedback from ERK to Raf upon EGF stimulation (dotted line from ppERK to dephosphorylation of pRaf), and positive feedback in case of NGF stimulation (dashed line from ppERK to phosphorylation of Raf). B. Differential equation model of the MAPK cascade. Bold parameters are the unknown constants, collected in the parameter vector θ, while gray parameters define the specific experimental condition for the simulation.

2.1.3.2. Sampling-based Bayesian approach for model calibration

Data-driven modeling approach Based on the experimental data available for model calibration and on existing modeling studies for the MAPK module [13, 105, 130], we formulated a differential equation model based on mass action kinetics for the three-tiered phosphorylation cascade (Fig 2.3).

In this cascade, both MEK and ERK require dual phosphorylation to become fully active. Double phosphorylation makes the cascade behave in an ultrasensitive way, which is advantageous for noise filtering [129, 172, 203]. MEK phosphorylation is processive, i.e. both sites

are phosphorylated in a single step, whereas ERK phosphorylation is distributive and requires two interactions [129, 203]. We have taken this into account by modeling MEK double phosphorylation as a single reaction, while full activation of ERK is obtained in a two-step reaction.

Furthermore, we exploited conservation of total protein concentrations,

$$\mathrm{Raf}_{\mathrm{TOT}} \cdot s_1 = \mathrm{Raf} + \mathrm{pRaf} \tag{2.3a}$$

$$\mathrm{MEK}_{\mathrm{TOT}} \cdot s_2 = \mathrm{MEK} + \mathrm{ppMEK} \tag{2.3b}$$

$$\mathrm{ERK}_{\mathrm{TOT}} \cdot s_3 = \mathrm{ERK} + \mathrm{pERK} + \mathrm{ppERK} \tag{2.3c}$$

to end up with a four variable model, shown in Fig 2.3B. Rate constants are denoted by $k_i^{+/-}$, $i = 1, \ldots, 4$, reduction of total protein amounts in the siRNA perturbation experiments are described by the silencing factors $s_i \in (0, 1]$ $(i = 1, 2, 3)$. These factors were extracted from quantification of the proteins in the control and the silencing experiments, as reported in Fig 1c of [199]. Their values were set to $s_1 = 0.72$, $s_2 = 0.7$ and $s_3 = 0.65$ when simulating silencing of Raf, MEK or ERK, respectively.

The input $u(t)$, which mimics signal initiation after addition of growth factor and summarizes all upstream processes, was described via a sigmoidally decreasing function, whose parameter K was also included in the optimization procedure,

$$u(t) = \begin{cases} 0 & t < 0 \\ 1 - \dfrac{t^3}{t^3 + K^3} & t \geq 0. \end{cases} \tag{2.4}$$

Thus, $u(t)$ jumps from 0 to 1 at time $t = 0$, which mimics addition of ligand, and subsequently decreases sigmoidally, reflecting observations of transient Ras activity, which is upstream of Raf and returns to its inactive Ras-GDP state within five minutes [210]. This implies that our model has a trivial steady state in which all variables are equal to 0 for $u = 0$, which is also a simplification, since proteins usually have minimal basal activities. However, since we do not have data for

$t = 0$, which would reflect these basal activities, and since these are anyway assumed to be very low compared to the stimulated case [24], we consider this simplification not a crucial one.

In our model the input $u(t)$ is not directly coupled to the network structure, which is clearly a simplification, since EGF and NGF trigger different receptor systems. However, exactly the same model structure has been used in other studies as well (see e.g. [245]) and was shown to display a rich variety of different behaviors, including ultrasensitivity and bistability and, as we will demonstrate, is also sufficient to capture various observed responses. Moreover, we follow here the argumentation in [36], according to which the different ERK responses are unlikely to be caused by different receptor systems. The Boolean variables f_p and f_n account for the experimental condition and act as switches between the two network structures, depending on the growth factor.

The positive feedback from ERK to Raf that was postulated from the modular response analysis in [199] was described by a sigmoidal function in order to facilitate bistability. Although this feedback is not necessarily required for bistability in the MAPK signaling pathway [144, 150, 180, 212], it has been shown to enhance the range of bistable behavior and to make the occurrence of bistability less sensitive to stochastic fluctuations and parameter variations [212].

Model calibration procedure In the next step we inferred the unknown model parameters

$$\theta = (k_1^+, k_2^+, k_3^+, k_4^+, k_1^-, k_2^-, k_3^-, k_4^-, k_{Fn}, k_{Fp}, g, K) \tag{2.5}$$

by using the described set of data y. For this model calibration procedure we used a sampling-based Bayesian approach, which provides a consistent statistical description for all quantities-of-interest. In a Bayesian approach, parameters θ and measurements y are interpreted as random variables that are characterized by probability distributions. Hence such an approach offers full information about uncertainties in terms of underlying distributions. A short explanation of the Bayesian idea is provided in Additional file A.1.

In our Bayesian framework the ODE model is stochastically embedded by defining the underlying stochastic process from which the experimental data are assumed to be generated. This is sometimes also referred to as noise model (see Additional file A.1 for more details). Here we exploit log-normal error models for protein concentrations, using the same standard deviation of 0.2 for the logarithmic transformation of the experimental data, which by definition are normally distributed.

These are translated into respective error models for the global response coefficients via transformation of probability distributions. Altogether, this defines the likelihood function $l_y(\theta) = p(y|\theta)$, which is a measure of how likely it is to see the experimental data given a particular model.

In a Bayesian framework, the objective function of interest is the posterior distribution $p(\theta|y)$, which is a distribution of parameters conditional on the given dataset. According to the Bayes Theorem, the posterior distribution is proportional to the product of the prior distribution $p(\theta)$ of the parameters and of the likelihood function,

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \tag{2.6}$$

Since the light signals of the Western blot data require appropriate rescaling and normalization to a reference experiment for a comparison across different experimental conditions, the ODE model in Fig 2.3B also had to be rescaled and normalized in order to enable a comparison with these data. This procedure is described in Additional file A.2. Moreover, a detailed formulation of the posterior distribution is given in Additional file A.3.

We investigate the posterior distribution by generating samples $\{\theta_i\}_{i=1,\ldots,N}$ via Markov Chain Monte Carlo (MCMC) sampling. These samples are subsequently used for Monte Carlo estimates of other quantities-of-interest. For example, the posterior predictive distribution (PPD) to see new data $\tilde{y}$ in any experimental scenario is given by

$$p(\tilde{y}|y) = \int_\Theta p(\theta, \tilde{y}|y)d\theta \quad \text{Marginalization} \tag{2.7a}$$

$$= \int_\Theta p(\tilde{y}|\theta, y)p(\theta|y)d\theta \quad \text{Factorization} \tag{2.7b}$$

$$= \int_\Theta p(\tilde{y}|\theta)p(\theta|y)d\theta \quad \tilde{y} \text{ is independent of } y \text{ given } \theta \tag{2.7c}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} p(\tilde{y}|\theta_i) \quad \theta_i \sim p(\theta|y) \quad \text{Monte Carlo estimate} \tag{2.7d}$$

If not stated otherwise, model predictions are consistently given in terms of these PPDs in this work.

2.1.4. Results

2.1.4.1. Calibrated model describes experimental data

We generated samples $\{\theta_i\}_{i=1,\dots,N}$ from the posterior distribution as described (see also Additional file A.4 for implementation details). Kernel density estimates of the marginal parameter distributions and 2D scatter plots for the two-dimensional parameter marginals are shown in Additional file A.5 and Additional file A.6. Most of the parameters show a large variance. The only exceptions are the dephosphorylation rates of pRaf and ppMEK, which mainly determine the speed of the decay of the signal. Moreover, the threshold parameter K of the input signal can be extracted from the data. There are also almost no correlations visible in the 2D scatter plots except a strong positive correlation between k_{Fp} and k_1^+.

Fig 2.4 shows the result of the Bayesian model calibration in the prediction space. Depicted are the Monte Carlo estimates of the PPDs in comparison with experimental data. Fig 2.4A and B show the dynamic responses of the observables pRaf, ppMEK and ppERK in the control experiments after stimulation with EGF (A) and NGF (B). The model captures the EGF scenario very well, with low variances in the PPDs. In case of NGF some data points are slightly overestimated, but the data are still within the predicted confidence intervals, which are larger here compared to the EGF scenario. The colors chosen for pRaf (blue), ppMEK (green) and

ppERK (red) are maintained for all simulation results throughout the paper.

A comparison of the global response coefficients is depicted at the bottom (Fig 2.4C). The sign structure is preserved for almost all silencing experiments, the only exception being MEK in the siERK experiments with NGF stimulation. This is due to the fact that we did not include the direct negative feedback from ERK to MEK that was postulated from the modular response analysis in [199] in our model, since the signal-to-noise ratio was rather low for this interaction, and we wanted to keep the model simple. At first glance the fits seem to be reasonable, which is however hard to judge solely from visual inspection, since error bars are large for most of these values. This is also mirrored by the variances of the PPDs. Thus we decided to validate the model via predictions of further experiments with the same cell line that were not used for model calibration.

2.1.4.2. Model is able to predict various perturbation experiments

For model validation we decided to use the model to predict outcomes of a set of perturbation experiments that have not been used for model calibration. The result is shown in Fig 2.5. In particular, the following experimental setups were considered:

Dose response profiles of ERK activation. We mimicked dose-response profiles of ERK activation to increasing EGF and NGF doses measured via flow cytometry (Fig 2 in [199]). Since these datasets are single-cell measurements that represent a heterogeneous cell population, we interpreted our parameter samples to represent such a cell population, whose average is consistent with the data used for calibration, and whose distribution accounts for population heterogeneity. Increasing ligand concentration was reflected by multiplying the parameter k_1^+, which describes the input strength, by a factor k_u. Resulting mean values of ppERK are shown in Fig 2.5A. We note here that this comparison can only be done in a qualitative way, since we lack a receptor model that directly relates growth factor concentrations to the input signal for Raf activation. Thus, it is

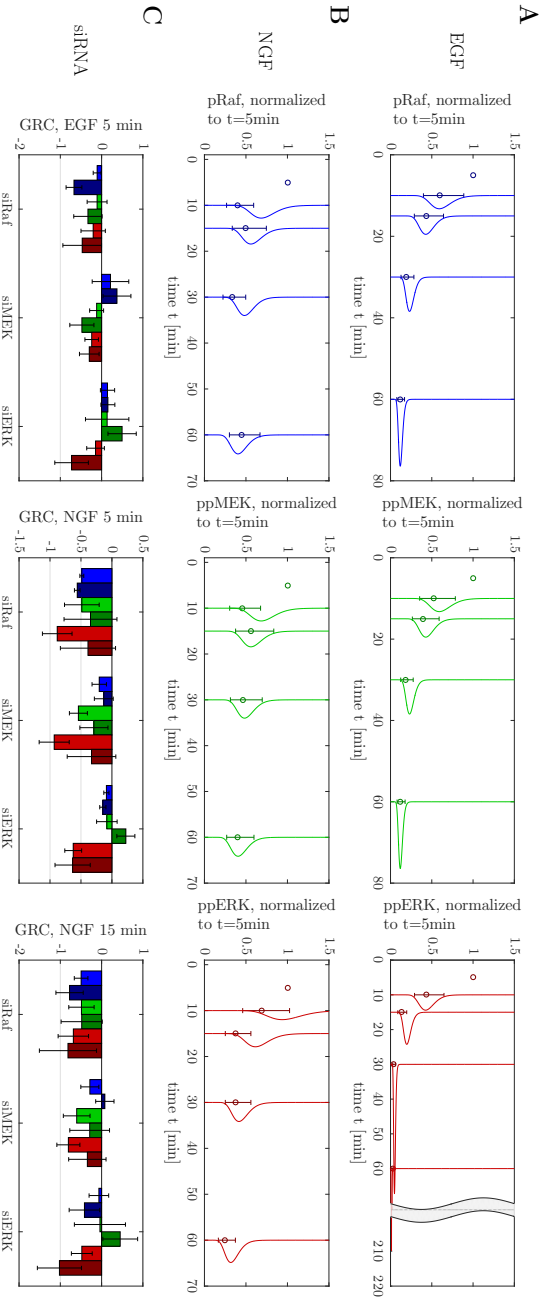

A
B
C
EGF
NGF
siRNA
pRaf, normalized to t=5min
ppMEK, normalized to t=5min
ppERK, normalized to t=5min
time t [min]
GRC, EGF 5 min
GRC, NGF 5 min
GRC, NGF 15 min
siRaf
siMEK
siERK

Figure 2.4.: Calibrated model using a Bayesian approach. Dynamic responses of pRaf, ppMEK and ppERK after stimulation with EGF (A) and NGF (B). Shown are the values acquired from flow cytometry experiments (Fig 2.1) in comparison to the respective PPDs predicted by the model. Data have been normalized to $t = 5$ min. C. Comparison of data for the global response coefficients (GRC) extracted from the siRNA perturbation experiments (Fig 2.2), and respective simulated distributions, here for clarity represented with the first and second moment. Data are taken from [199].

here not possible to include the respective experimental data directly for comparison. However, model simulations capture the observed qualitative phenomena quite well: In case of stimulation with EGF the ERK activity profile is unimodal and raises sigmoidally with increasing EGF concentration. In contrast, upon stimulation with NGF the profile becomes bimodal when NGF doses exceed a threshold. Moreover, with increasing NGF concentrations the fraction of cells with a sustained response as well as the mean ERK activities of both subpopulations increase.

Effect of feedback breaking via inhibition of MEK and PKC. We predicted the influence of MEK inhibition via the MEK inhibitor PD184352 on the temporal activity of Raf (Fig S1e in [199]), by assuming that MEK activation is completely abolished. This was realized in our model by setting the MEK phosphorylation rate k_2^+ to zero, which destroys the feedback from ERK to Raf in the simulations, and inspection of Raf activity (Fig 2.5B left). While the response is sustained in the control case (blue continuous PPDs), MEK inhibition results in the loss of sustained Raf activity, and pRaf follows the transient signal and rapidly drops within a few minutes (gray dashed PPDs). This result is in agreement with the observations in [199].

In addition, we mimicked the inhibition of PKC via Gö7874 during NGF stimulation (Fig 4a in [199]). We considered the feedback to be completely eliminated as a result and realized this by removing the feedback connections from our model (Fig 2.5B right). In the control case (red continuous PPDs) activity of ppERK was sustained, whereas the feedback deletion caused a decrease in ERK activation (gray dotted PPDs), again in accordance with experimental findings.

Irreversibility in MAPK activation. Finally, we also compared our model to experimental data on the irreversibility in MAPK network activation upon NGF stimulation, which was investigated via terminating the signal by growth factor neutralizing antibodies and TrkA inhibitors (Subfigs 3a and c in [199]). Therefore, both perturbations, i.e. addition of neutralizing antibody and TrkA inhibitor after stimulation, were mimicked via abrupt signal termination at the

respective time points. Results are shown in Fig 2.5C. While in case of stimulation with EGF, ppERK was virtually zero shortly after addition of the neutralizing antibody (gray dotted PPDs in the left Figure), the NGF inhibition profile still showed some activity after 60 min (red PPDs).

For a further comparison we simulated ppERK time courses upon stimulation with NGF and addition of TrkA inhibitor at two different time points (Fig 2.5C right). PPDs for ppERK are depicted at $t = 17$ min after stimulation when TrkA inhibitor was given at $t = 3$ min after stimulation (continuous curve) and $t = 12$ min after stimulation (dashed curve), compared to the control case (dotted curve). In agreement with experimental findings, results show that ERK activity rapidly drops in case that the stimulus terminated too early.

Overall, the results in Fig 2.5 nicely demonstrate that our model is able to predict many important features of the signaling cascade quite accurately. Since these simulation scenarios capture the responses of the system to several treatments that are quite different from the experiments which have been used for fitting, the model is validated to have predictive power.

In the next step we decided to use the model to analyze mechanisms behind sustained ERK response in case of NGF stimulation.

2.1.4.3. Mechanism behind sustained response caused by NGF

Bifurcation analysis reveals that bistability is not sufficient to explain model outcomes upon NGF stimulation In order to investigate the mechanisms behind sustained response to transient NGF signals, we combined our sampling-approach with the circuit-breaking algorithm (CBA) [184], which allows for an efficient calculation of steady states based on the topology of the signaling network, and for an automatic classification into mono- and bistable systems. Our approach is schematically illustrated in Fig 2.6. Figs 2.6A-C illustrate the steps of the CBA applied to our network model for a single parameter sample θ_i. The CBA operates on the topology of the interaction graph $G(V, E)$, which is a directed graph that shows de-

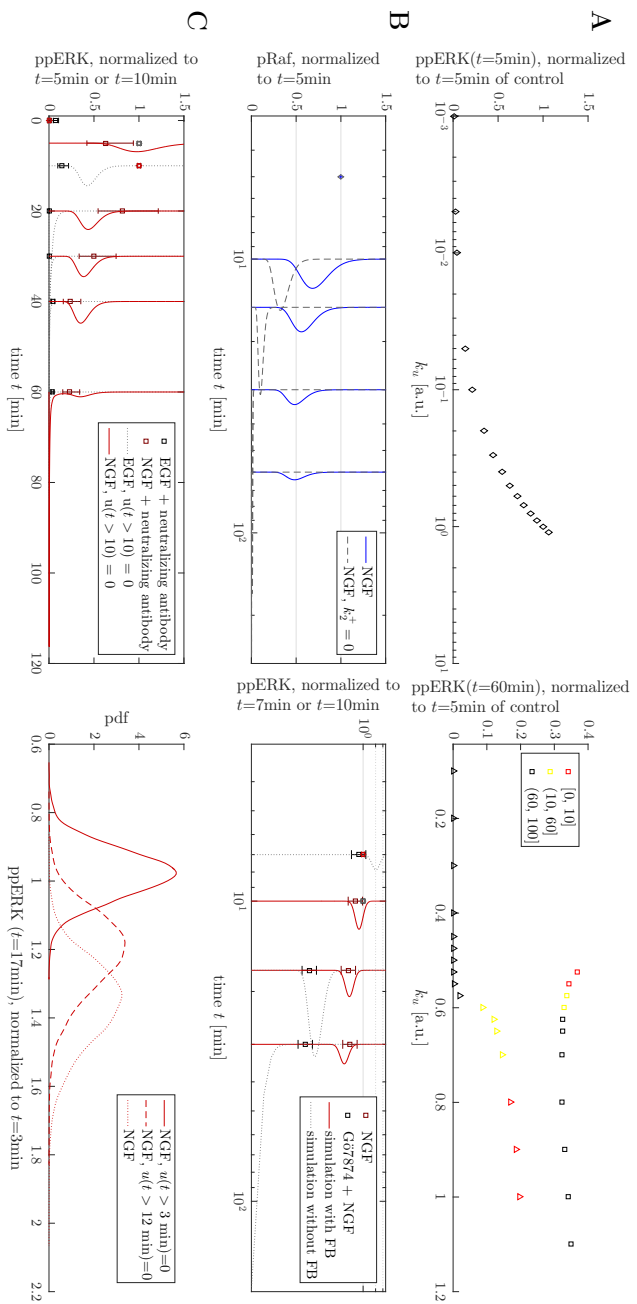

A
ppERK(t=5min), normalized to t=5min of control
k_u [a.u.]
B
pRaf, normalized to t=5min
time t [min]
NGF
NGF, $k_2^+ = 0$
C
ppERK, normalized to t=5min or t=10min
time t [min]
EGF + neutralizing antibody
NGF + neutralizing antibody
EGF, u(t > 10) = 0
NGF, u(t > 10) = 0
ppERK(t=60min), normalized to t=5min of control
k_u [a.u.]
[0, 10]
(10, 60]
(60, 100]
ppERK, normalized to t=7min or t=10min
time t [min]
NGF
G6874 + NGF
simulation with FB
simulation without FB
pdf
ppERK (t=17min), normalized to t=3min
NGF, u(t > 3 min)=0
NGF, u(t > 12 min)=0
NGF

Figure 2.5.: Model validation. A. Dose-response profiles of ERK activation were mimicked by simulating the model with increasing input strength parameter k_u for stimulation with EGF (left) and NGF (right). The system shows an unimodal and ultrasensitively increasing ppERK concentration after stimulation with EGF ($t = 5$ min after stimulation) and a bimodal distribution when stimulated with NGF exceeding a threshold concentration ($t = 60$ min after stimulation) (compare data in [199], Subfigs 2c and d). B. Inhibition of MEK (left) results in the loss of sustained Raf activation upon stimulation with NGF (gray dashed PPDs) compared to the control case (blue continuous PPDs). Inhibition of PKC via Gö7874 (right) causes the loss of sustained ERK activation upon NGF stimulation (data from [199], Fig 4a). This was simulated by switching off the feedback connection. C. Irreveversibility in MAPK activation upon NGF stimulation was investigated via mimicking treatment of the cell culture with neutralizing antibodies (left) and TrkA inhibitors (right) (compare data in [199], Subfigs 3a and c).

pendencies between variables in the model (Fig 2.6A). In the first step all feedback loops[1] are broken by deleting incoming edges for a suitably chosen subset $\tilde{V}$ of vertices and setting the respective variables to fixed values κ. The remaining vertices are collected in the set $\widehat{V}$. Here we set $\tilde{V} = \{x_4\}$, $x_4 = \kappa$ and $\widehat{V} = \{x_1, x_2, x_3\}$. The state variables $x_i, i = 1, \ldots, 4$, in the interaction graph refer to the rescaled states of our ODE model that we used for all simulations (see Additional file A.2). Then we calculated the steady state coordinates of the variables in $\widehat{V}$ in dependence of the input κ, obtaining the set $\bar{x}_{\widehat{V}}(\kappa, \theta_i)$ (Fig 2.6B). In the last step the circuits are released one after another by releasing vertices in the set $\tilde{V}$ (Fig 2.6C). Mathematically, this translates here into the calculation of the zeros of the circuit-characteristic $c(\kappa, \theta_i)$, which is given by

$$c(\kappa, \theta_i) = f_{x_4}(x_4 = \kappa, x_{\widehat{V}} \in \{\bar{x}_{\widehat{V}}(\kappa, \theta_i)\}) = 0. \tag{2.8}$$

The obtained zeros $\bar{\kappa}$ of the circuit-characteristic correspond to the steady state coordinates of the state variable x_4, from which the set of steady states of the full system can be derived. All details about the calculation of the values for $\bar{\kappa}$ and of the expressions of the steady state coordinates for the other three state values $\bar{x}_{\widehat{V}}(\kappa, \theta_i)$, as functions of the parameter sample, are given in Additional file A.7.

We applied the CBA to all parameters of the estimated posterior sample. The outcome was automatically classified, by using this analysis, according to the number of steady states of the system (Fig 2.6D). Results show an overall probability of 10% for the system to be bistable. We found this a surprisingly small number, which indicates that bistability is probably not the main mechanism behind the observed sustained ERK activation. Even worse, our analysis only provides an upper bound in two respects: First, depending on the parameters θ_i, not all trajectories of bistable systems might be pushed to the basin of attraction of the second fixed point by the transient signal. Second, this set might also contain bistable systems in which the distance of the two steady states is rather small, such that the bistability will not be visible in any real experiment.

[1] Called in the following *circuits*, for consistency with graph-theoretic terminology

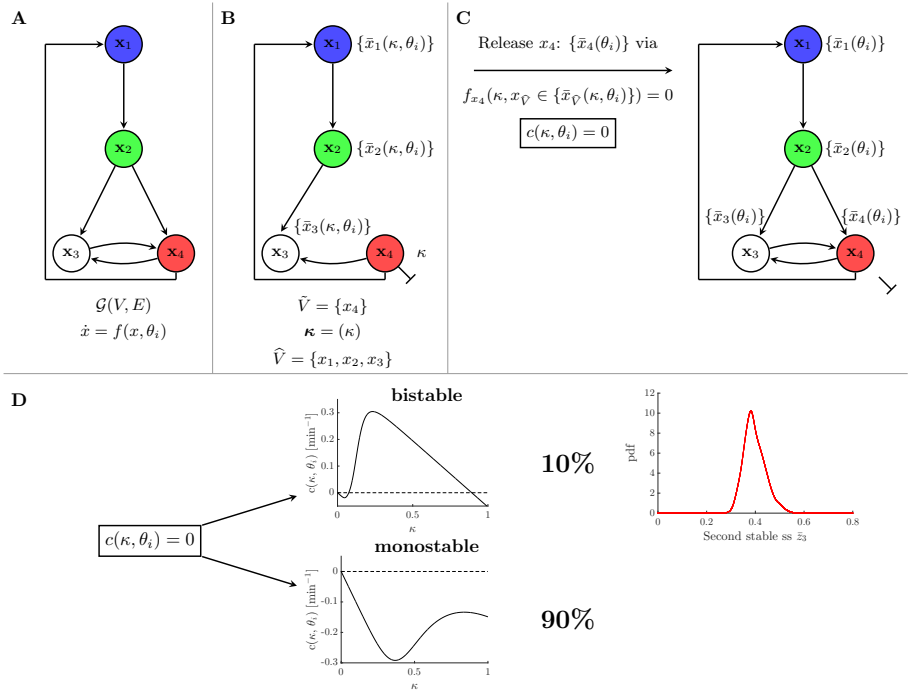


Figure 2.6.: Steady state analysis using the circuit-breaking algorithm. The CBA is used for an efficient calculation of the steady states of the system for the MCMC parameter samples and subsequent automatic classification into mono- and bistable systems (Subfigs A-C). D. Result of this classification analysis. Depicted is also the distribution of the second stable steady state $\bar{z}_3 \neq 0$ in case of a bistable system, which corresponds to the concentration of active ERK normalized to $t = 5$ min (see Additional file A.2).

Furthermore, we simulated the ODE model with the obtained subset of parameter samples θ_i leading to a bistable system, and we calculated the distribution of the second positive stable steady state $\bar{x}_4$. This is shown in Fig 2.6D on the right, by considering the normalized state variable (see Additional file A.2)

$$z_3(t) = x_4(t)/x_4(t = 5 \text{ min}).$$

Overall, this analysis suggests that bistability is not sufficient to explain the observed sustained activation of ERK after NGF stimulation.

Quasi-bistability can explain sustained ERK activation We complemented our steady state analysis by a simulation-based classification of model trajectories after NGF stimulation, as illustrated in Fig 2.7. A similar classification approach was used in [150], without explicitly investigating quasi-bistability.

We used the posterior sample to simulate model responses up to $t = 600$ min. These responses were automatically classified in a second step (Fig 2.7A): Using ERK activity at $t = 5$ min as a reference value, samples were sorted according to the following classification scheme:

- Class 1 (Bistable systems):

$$\frac{\text{ppERK}(60 \text{ min})}{\text{ppERK}(5 \text{ min})} > 0.2 \text{ and } \frac{\text{ppERK}(600 \text{ min})}{\text{ppERK}(5 \text{ min})} \geq 0.1$$

- Class 2 (Quasi-bistable systems):

$$\frac{\text{ppERK}(60 \text{ min})}{\text{ppERK}(5 \text{ min})} > 0.2 \text{ and } \frac{\text{ppERK}(600 \text{ min})}{\text{ppERK}(5 \text{ min})} < 0.1$$

- Class 3 (Monostable systems):

$$\frac{\text{ppERK}(60 \text{ min})}{\text{ppERK}(5 \text{ min})} \leq 0.2$$

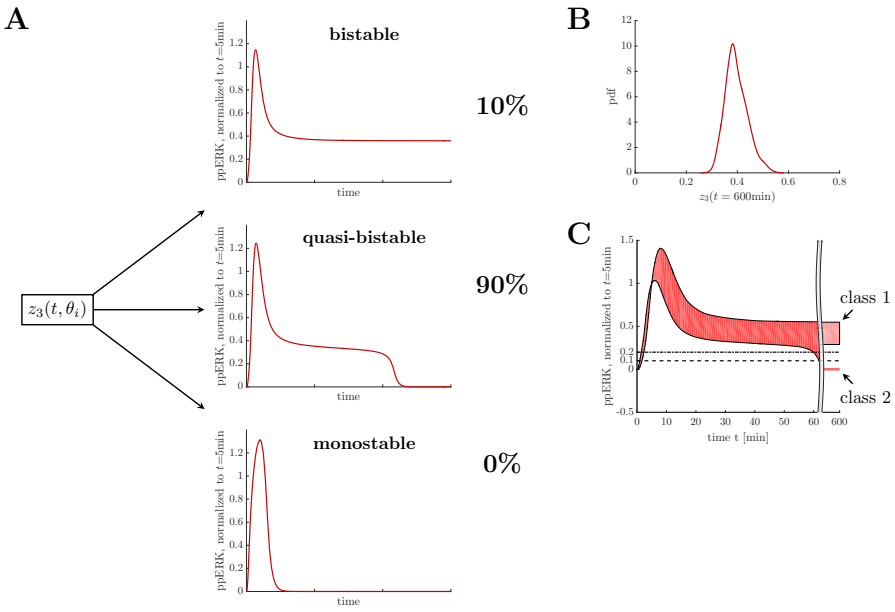


Figure 2.7.: Simulation-based analysis of the long-term behavior of the cell population. A. Trajectories are automatically classified into bistable, quasi-bistable and monostable systems, as described in the text. For stimulation with NGF in the control experiment 90% show a quasi-bistable behavior, while only 10% are really bistable. B. Distribution of the second stable steady state distinct from zero of the bistable trajectories. C. For the choice of threshold parameters that were used for the classification scheme we evaluated the 0% and 100% percentiles of ERK trajectories in the simulation of the NGF control experiment.

The threshold 0.2 for $t = 60$ min was chosen such that all trajectories are above this value. This implies that class 3 is empty, and no trajectory is classified to be simply monostable, as visualized in Fig 2.7C. This implication is reasonable, since $u(t)$ is close to 0 already a few minutes after stimulation, and hence we expect simple monostable systems to follow the input with a delay that is much smaller than 1h. Fig 2.7C also shows that the classification result is rather insensitive to fine-tuning of the second threshold at $t = 600$ min: At this late time point the quasi-bistable and bistable trajectories are already well separated and there is a clear gap between the trajectories of classes 1 and 2.

The analysis revealed a fraction of 10% belonging to class 1. The estimated distribution of the second steady state equals that from the CBA analysis, which hints to the fact that trajectories of virtually all bistable systems detected via the CBA converge to the second steady state after stimulation with NGF. The rest of the samples, which are 90%, belong to class 2, which represent monostable systems that can show a sustained response for more than 60 minutes after stimulation. However, trajectories in this class converge to their unique steady state at a later time point.

In order to understand the mechanism behind this highly prolonged response to a transient input signal, we filtered the parameter sample for monostable systems that belong to class 2 and investigated their behavior in more detail. Therefore, we used the input $u(t)$ as a bifurcation parameter and investigated the respective time-varying set $\{\bar{x}(u)\}$ of steady states of the system via the CBA. Fig 2.8 shows the temporal behavior of the set $\{\bar{z}_3(u(t))\}$ for a representative parameter sample belonging to class 2. After a fast transient phase, the system is bistable, since $u(t)$ is sufficiently large to maintain two stable steady states. However, the second stable steady state vanishes due to a rapidly decreasing $u(t)$. For the trajectory at hand the system becomes monostable already at about $t \approx 102$ min, which is fast compared to its switching time at $t \approx 440$ min. This comes from the fact that, although the system is monostable, $c(\kappa, \theta_i)$ is extremely small about the region of the former second steady state. This causes a very slow dynamic, which can be seen by tracking the normalized

state variable $z_3(t)$, as indicated in the Figure. Only at $t \approx 440$ min $z_3(t)$ reaches an area where $|c(\kappa, \theta_i)|$ becomes larger, which results in a subsequent fast convergence to the unique globally stable fixed point at the origin.

Thus, taken together, this analysis suggests that quasi-bistability is caused by traversing a region in the state space in which $\dot{x}$ is extremely small, resulting in a very slow dynamics. The system is only accelerated towards its single steady state when the state of the system leaves this region. This makes the system behave as a bistable system for a long time span. This hypothesis was confirmed by a subsequent bifurcation analysis with some representative parameter sets for classes 1 and 2 of the classification scheme, as shown in Fig 2.9. Fig 2.9A illustrates the two effects that act together to delay the response of the system upon a transient stimulus. Fig 2.9B shows the absolute value of the vector field $\|f(x(t, \theta_i))\|$ of the same trajectory as in Fig 2.8, which shows high values a few minutes after stimulation, followed by a long period where $\dot{x}(\theta_i)$ is virtually zero, and a second peak at about $t = 440$ min, where the trajectory is pushed towards the systems unique steady state. A comparison of bifurcation diagrams for representative parameter sets belonging to classes 2 (quasi-bistable) and 1 (bistable) is depicted in Fig 2.9C and shows that the difference between these two classes is actually 'smooth' in terms of changes in limit sets.

2.1.5. Discussion and conclusions

We presented a modeling study that focuses on mechanisms behind sustained responses of signaling pathways upon transient stimulation in PC-12 cells. The model is based on chemical reaction kinetics and was calibrated to a dataset of PC-12 cell lines that were stimulated with EGF and NGF in a control setting and under silencing perturbations. We used a sampling-based Bayesian approach for model calibration, and analyzed model predictions in terms of posterior predictive distributions, which provides complete information about remaining uncertainties. The model was validated by comparing model predictions of new scenarios to experimental data.

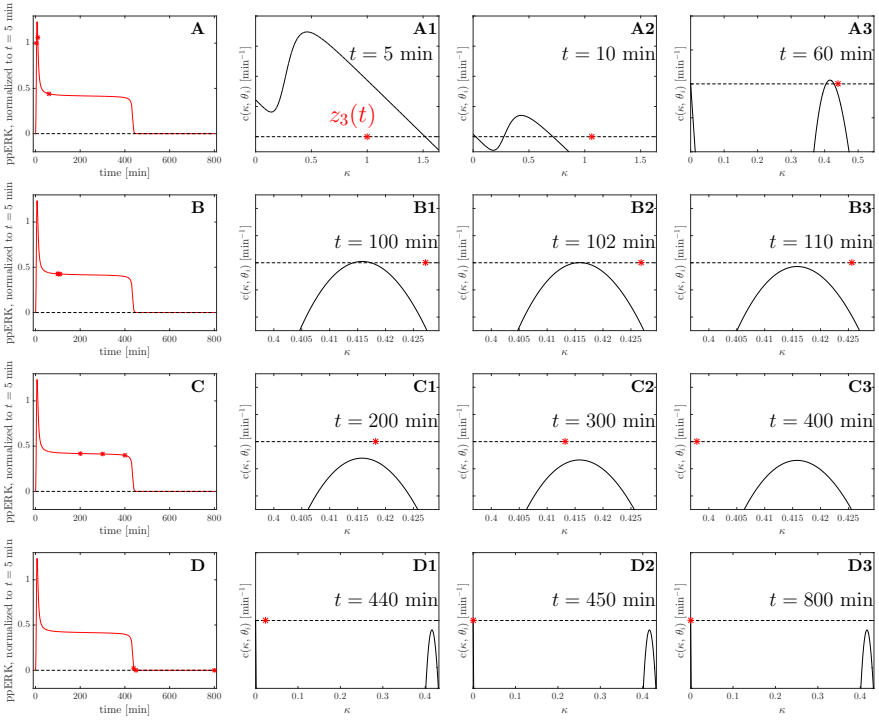
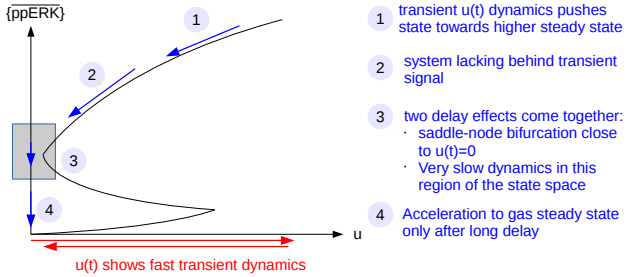
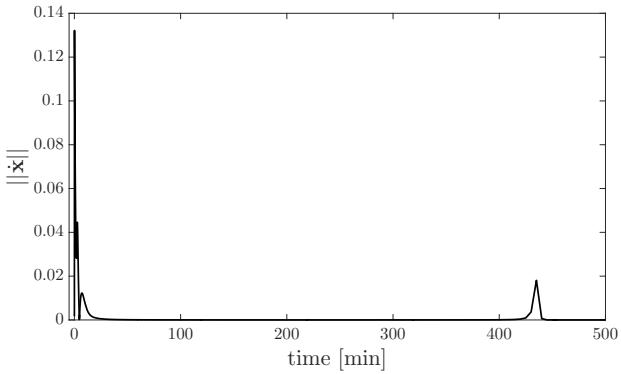

A
A1
A2
A3
B
B1
B2
B3
C
C1
C2
C3
D
D1
D2
D3
$t = 5$ min
$t = 10$ min
$t = 60$ min
$t = 100$ min
$t = 102$ min
$t = 110$ min
$t = 200$ min
$t = 300$ min
$t = 400$ min
$t = 440$ min
$t = 450$ min
$t = 800$ min
$z_3(t)$
ppERK, normalized to $t = 5$ min
time [min]
$c(\kappa, \theta_i)$ [min^{-1}]
κ

Figure 2.8.: Quasi-bistability phenomenon. The CBA is used for the investigation of the quasi-bistability phenomenon, in which the system, despite being monostable, shows a very prolonged sustained response. The first column shows the time course of normalized ppERK for a representative parameter sample from class 2 with switching time at $t_{\text{switch}} \approx 440$ min. Columns 2,3 and 4 show the circuit-characteristic $c(\kappa, u(t))$, along with the actual normalized state ppERK(t) for 12 different time points. After a fast transient dynamic (Subfig A1) the circuit-characteristic has three zeros (Subfig A2-B1), which disappear at a later time point, here $t = 102$ min (Subfig B2), via a saddle-node bifurcation. After 60 min the input is almost zero and the vector field and therefore the circuit-characteristic changes only slowly. The system state has almost approached the higher fixed point. Subfigures B1-C3 are eyeglass views on the dynamics near this second fixed point. These plots show that, even if the fixed point has disappeared, the system trajectory moves very slowly through the state space for a rather long time, since $\dot{x}$ is still small. Only after about 440 min the system has overcome this slow region of the state space, and from here on rapidly moves towards its globally asymptotically stable steady state $\bar{x} = 0$.

A


{ppERK}
1
2
3
4
u
u(t) shows fast transient dynamics
1 transient u(t) dynamics pushes state towards higher steady state
2 system lacking behind transient signal
3 two delay effects come together:
· saddle-node bifurcation close to u(t)=0
· Very slow dynamics in this region of the state space
4 Acceleration to gas steady state only after long delay

B


0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
||ẋ||
0
100
200
300
400
500
time [min]

C

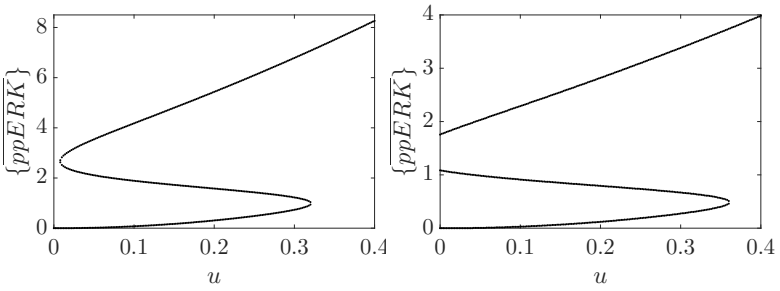

{ppERK}
u
{ppERK}
u

Figure 2.9.: Combination of two delay mechanisms in quasi-bistable systems. A. Scheme of a bifurcation diagram for a quasi-bistable system. The system is monostable for $u = 0$ and has a saddle-node bifurcation u^{SNB} close to $u = 0$, where it becomes bistable. A sufficiently strong transient signal $u(t)$ pushes the system state into the basin of attraction of the higher stable steady state (1). As long as the change in $u(t)$ is not slow compared to the dynamics of the system, the system cannot be considered in quasi-steady state, and we observe a transient dynamics (2). When $u(t)$ is almost back to 0, two delay effects lead to quasi-bistable behavior (3). First, the system remains in the upper stable steady state as long as $u(t)$ is still above the saddle-node bifurcation. Second, for $u(t) < u^{\mathrm{SNB}}$ the acceleration remains very small in this region of the state space. B. Absolute value of the vector field along the model trajectory for the same model parameters that have been used in Fig 2.8. C. Two representative bifurcation diagrams for a quasi-bistable system belonging to class 2 of the classification scheme, and a bistable system belonging to class 1.

Interestingly, the system shows a sustained ERK activity profile upon NGF stimulation, while the response was transient in case of EGF stimulation. This phenomenon has been well-investigated experimentally and theoretically, and it is well believed that the observed sustained response is caused by a bistable system. Here we combine our statistical inference approach with steady state analysis to investigate mechanisms behind this sustained ERK response. Surprisingly, our results indicate that the probability for bistable behavior is far below the observed response, and thus suggest that it is not sufficient to concentrate analysis on steady states only. A simulation-based analysis of the phenomenon revealed the importance of quasi-bistability to shape ERK response. A system is said to be quasi-bistable, if it is monostable but able to maintain a state distinct from its steady state for a long time period. It is known that positive feedback can generally cause quasi-bistability [162], it has however not been shown that this is relevant for decision-making in living systems.

Our bifurcation analysis showed that the transition between the three classes (bistable, quasi-bistable and monostable) is actually smooth in terms of locations of bifurcation points and limit sets, and we expect the range of parameters in which quasi-bistability occurs to be rather small. This expectation was confirmed by a sensitivity analysis for the outcome of the simulation-based classification scheme (Additional file A.8), in which we varied all model parameters independently one at a time about the maximum-a-posteriori estimator. The result shows that the appearance of quasi-bistability is highly sensitive to these variations for almost all parameters. Except for the phosphorylation rate of Raf (parameter k_1^+) and the threshold parameter K of the input function, which do not have any influence on the limit sets of the system for $u = 0$, small variations of parameters induce switches to mono- or bistable systems, which is due to the fact that the location of the saddle-node bifurcation and hence the delay time are very sensitive to parameter changes. The fact that most of the samples fall into this seemingly small parameter range shows that the gradient of the posterior distribution must be rather high when varying parameters individually. This is indeed the

case, since the fit quality rapidly drops at least for the NGF control experiment when switching from the quasi-bistable to the monostable range, and we observe a similar effect for the switch to the bistable range. Altogether, these results also indicate the existence of strong correlations between parameters in the posterior distribution.

For a biological system it might not make a difference at all whether the underlying system is indeed bistable or quasi-bistable, since the system probably acts as an integrator of a response, which starts to trigger further events as soon as a threshold has been reached. However, one has to be careful with the analysis of models for such mechanisms. Our results propose to consider, besides limit sets, also the transient behavior of a system when investigating processes such as switches, memory effects or decision-making.

We note here that the distribution of switching times of quasi-bistable trajectories of our inferred model partly disagrees with observed ERK activities for later time points. While most of the trajectories switch to their steady state between 60 and 120 min in our model, experiments in [167] and [181] show that ERK activity is sustained for at least 2-3h. Since our dataset only contained measurements for up to 60 min, this fact was not taken into account in the model calibration procedure. However, the response duration of the signaling cascade upon stimulation with NGF also seems to show a large variation and to depend in particular on the experimental protocol and distinct clonal PC-12 cell lines, as also stated in [167]. In [163] or [75], for example, MEK (MAP-2 kinase) and ERK activities are almost completely down to the basal level already after 2h (see [163], Fig 3 and [75], Fig 9B, curve without treatment with TPA). Hence it is not completely clear how to describe the activity of the module quantitatively in order to enrich the model with knowledge on the long-term behavior. However, we started to investigate the effect of assuming different minimal switching times for ERK activity, which is illustrated in Additional file A.9. As expected, filtering for trajectories that still have a substantial remaining activity after two and three hours, respectively, increases the ratio of bistable versus quasi-bistable trajectories, since only quasi-bistable trajectories are filtered out. Thus, we think that our model is generally able also to

match the long-term behavior of the cascade.

There is an ongoing debate about relations between ultrasensitivity and/or bistability in responses of single cells on the one hand and the occurrence of bimodality in heterogeneous cell populations on the other hand (see e.g. [24, 129]). It is clear that ultrasensitive and bistable systems can lead to bimodal responses, for example caused by variations in protein contents or stochastic fluctuations. An example that considers the role of mutual inhibition in a gene regulatory network for metastatic transitions and the appearance of stable subpopulations of genetically identical prometastatic cells is described in [143]. On the other hand, for the MAPK pathway it has been shown that bimodality can also emerge from graded single cell responses caused by a broad distribution of ERK pathway activation thresholds [24]. This example reveals that the relation between bistability and bimodality is actually more subtle than a simple one-to-one relation.

We had decided in this study to use a data-driven approach and to adapt model granularity to the data available for model calibration. This results of course in a very simplified model, and the situation *in vivo* is much more complex. Specific aspects regarding the MAPK signaling pathways are discussed in literature and have also partly been implemented in models. One of the most recent interesting studies investigates the role of feedbacks and their time scales by using pulse experiments on a single cell level (see also [29] for a commentary on this). Kocieniewski et al. [128], for example, focus on the role of the two different MEK isoforms and their contribution in the regulation of the ERK response. According to this study, response duration and amplitude are regulated by the ratio and the total amount of both isoforms, respectively. Moreover, localization of proteins and their regulation via scaffolding proteins, together with nucleoplasmic shuttling, is known to play a major role in the regulation of the pathway [2, 129, 180]. Cross-talk and interactions with other cellular pathways is another important aspect [129, 172], which is difficult to take into account in any modeling approach. However, it is an important and interesting question how single modules such as the MAPK signaling cascade behave embedded in a

larger and more complex network. Several studies in recent years hint to the fact that network complexity is intimately linked to functional robustness, meaning that the network structure, and in particular interlinked feedback loops, contribute to a reliable performing of tasks in the presence of perturbations and noise [17, 42, 44, 218, 236].

Furthermore, we have not explicitly taken into account fluctuations in protein content, although we are aware that this is a major source of variability in cell populations. The total amounts of Raf, MEK and ERK do not explicitly appear anymore in the rescaled and normalized model version that we used for our study, hence it is not possible to take absolute fluctuations into account. However, we investigated the effect of varying absolute concentrations by varying the coefficients $s_i,\ i = 1, 2, 3$ in a narrow range about its nominal values $s_i = 1$ and considered the sensitivity of bistability and quasi-bistability to these parameters. Exemplary results are shown for variations in ERK (s_3) in Additional file A.10. Figures for variations in s_1 and s_2 look very similar. Interestingly, the classification of trajectories seems to be very sensitive to these parameters. As can be seen, a moderate reduction in s_i is sufficient to destroy bistability and quasi-bistability almost completely, while bistability is strongly enhanced upon a slight increase in s_i. This is a surprising result, since it is known that stochastic gene expression events can, for example, result in coefficients of variation of about 20-30% in the content of individual proteins [62, 180]. This raises the general question about reliability and robustness of decision processes under such variations. To our knowledge, minimal models for bistability, as used here, are often not robust with respect to such fluctuations and parameter variations, which might trigger further investigations in this direction.

A further critical point in our modeling study is the normalization of model outputs to a particular time point. This normalization was necessary since the dataset used for model calibration only provides relative information. Signals are given in arbitrary units, and the scaling factors are different for each antibody and can also vary across membranes. Thus normalization to a reference experiment is required to make measurements from different experiments comparable and is standard in representing biological data and for modeling [128].

This normalization, however, affects variances of observables, and precludes comparison with experiments where total protein levels matter, such as absolute heights of ppERK peaks under different conditions. Thus, including some information about total protein levels could highly enrich the modeling process in the future.

Finally, recently a new modeling approach, called ODE constraint mixture modeling, was introduced [88]. This approach combines advantages of mechanistic modeling approaches with statistical mixture models to describe heterogeneous cell populations. This framework allows to infer subpopulation structures and dynamics from single cell snapshot data. Since the data used here for model calibration represent only population averages, we did not explicitly take subpopulation structures into account. However, at least the dose response profiles of ERK after stimulation with NGF seem to consist of two or more subpopulations, which was also exploited to mimic the respective dose response curve. Thus, exploiting this framework is another interesting task for future investigations.

2.1.6. Contribution

A large part of this project was constituted by defining a model by iterating different model possibilities. I contributed substantially in this process of refining our final model by discussing, selecting and testing various model setups. Furthermore, I implemented the sampling procedure for the parameter estimation and the model predictions with help from Caterina Thomaseth. I also performed the steady state analysis from the viewpoint of the simulations and gave input to the study design. I was actively involved in the discussion of the results and implications during all stages of the project. Finally, I contributed to writing the manuscript and the visualizing the results.

2.2. Results in the overall context

In this project we investigated the feedback mechanism of the MAPK signaling pathway, which can produce transient and sustained activa-

tion of ERK dependent on the stimuli. The sustained ERK response induced by stimulation with NGF was herein the major interest.

We used data provided in [199], who investigated the network structure of the MAPK module. Thus, we worked with a limited amount of experimental data with few replicates. Working with the data collected in [199], the data preparation step was simplified to extracting the relevant experimental conditions and corresponding measurements from the publication without a need for extensive further processing. Only normalization to a reference experiment was required in order to allow comparison of the data across different experimental conditions.

Following the inferred network topology in [199], we developed our mathematical model. The challenge in this step constituted defining a model as simple as possible, but with enough detail to properly capture the distinct mechanisms of the positive and negative feedback loops. We used mass action kinetics to model most of the pathway. Only the input function $u(t)$ and the positive feedback connection were modeled using a sigmoidal curve to support the sustained ERK response under NGF stimulation. Nevertheless, by fixing one parameter in each case, the number of parameters remained modest, which also allowed for faster computations. In this step we also included further knowledge of system behavior, as we simplified the double phosphorylation of MEK to a single step. Thus, we avoided additional parameters and a state variable for which no data was available.

For model calibration, I applied the Markov-Chain Monte Carlo (MCMC) method. This choice allowed to infer uncertainties of underlying distributions, which in this setting refers to the distribution of steady states.

This approach allowed us to analyze the estimated posterior distribution of simulated trajectories, originating from the extensive MCMC sampling procedure. In turn, this enabled us to identify quasi-bistable trajectories, whose parameter combinations might not have been selected by point estimators such as maximum likelihood estimation. A multi-start local optimization indicated promising parameter sets which served as adequate initialization points for MCMC

sampling. Through this combination, I exploited the fast ML method to find not just good starting points, but also appropriate parameter bounds through repeated runs for the extensive and time-consuming MCMC method to sample from the posterior distribution of model parameters.

For the model validation, there was no additional test data, as all data was necessary for model training. In order to show the validity of the model, we instead selected additional experiments from [199], that were not used in the model calibration step, and mimicked the experimental settings with model simulations. As not all experimental conditions were included in our model, we carefully studied the biological connections to find a way to influence our model correspondingly. One example for this approach is the feedback breaking experiment via inhibition of PKC. PKC is described to be inhibited by Gö7874, which results in the loss of sustained ERK activation under stimulation with NGF in [199]. As neither Gö7874 nor PKC are included in our model, I instead considered the positive feedback from ERK to Raf to be completely eliminated and implemented this by switching off the feedback connection. Proceeding similarly, I could show a qualitatively good agreement between reported additional experimental data in [199] and model simulations given the previously calibrated model for six different settings (Figure 2.5).

In the final step of model analysis, we investigated our model with regard to the positive feedback. As mentioned in the Introduction, analysis tools depend on the purpose of the investigation and are in general individually selected. Nevertheless, we also employed 2D scatterplots (Figure A.4) to investigate correlations among parameters. As the analysis of the positive feedback constituted the major research interest, further analysis was centered around the feedback mechanism. In this context, we were interested in the specific way the trained model would achieve a prolonged ERK activation. To analyze the sustained ERK activation, we employed two approaches. The first approach performed by Caterina Thomaseth used the circuit-breaking algorithm for steady state analysis. Calculation of the steady states of the system for the MCMC parameter samples revealed that only 10% of all systems were bistable, which contradicted the observed

sustained response of our simulated trajectories. As a second approach, I complemented the steady state analysis by additionally inspecting the simulated model trajectories. I classified all simulated trajectories by their long-time behavior in three categories, consisting of monostable, quasi-bistable and bistable systems. I chose 600 min as the critical time point for the decision of bistability versus quasi-bistability, assuming that trajectories that remained in the active state until this point, had a high probability of remaining active. As the threshold of the activation level was chosen such that there are no monostable systems, 10% of the systems were classified as bistable, hereby confirming the results of Caterina Thomaseth. The large majority of 90% of trajectories were quasi-bistable and thus able to sustain ERK activation for a prolonged time before returning to the inactive stable steady state.

We further investigated the mechanism generating quasi-bistability by analyzing the behavior of monostable systems that display long-term ERK activation. The CBA was employed to infer about steady states and the circuit-characteristic for ERK activity at different time points (Figure 2.8). In addition, the absolute value of the vector field in the same time frame was evaluated (Figure 2.9b). Results were used to suggest that quasi-bistability is related to regions with slow dynamics of the state space after the input gets close to zero.

For additional analysis, we visualized different aspects of our model. We investigated the range of parameter values in which quasi-bistability exists by individually varying each parameter about the maximum-a-posteriori estimator and classifying the corresponding system in terms of stability (Figure A.5). This approach was used to infer the sensibility of quasi-bistability to changes in the parameters and revealed that the classification scheme is very sensitive to variations in all but two parameters. Another analysis aspect considered the ability of our model to reflect long-term behavior of the system. Long-term activation of ERK could not be included in the calibration process as our dataset only consisted of measurements for up to one hour. Inspired by results on ERK activity in [167] and [181], we evaluated the distribution of bistable and quasi-bistable trajectories that retain significant activity for up to two or three hours.

Results suggest that our model is fundamentally able to describe the long-term behavior of the system. In a final step, we also employed the CBA and simulations to assess the influence of variations in the total protein amount on the classification of sample trajectories with respect to their stability. As Figure A.7 shows exemplarily for variations in ERK, for a moderate reduction bistability and quasi-bistability disappear. On the other hand, already small increases result in bistable trajectories.

In summary, we efficiently extracted data from literature and developed a small yet sufficiently accurate model of the MAPK signaling pathway. In order to keep computation time and model complexity to a minimum, we carefully adjusted the model to be just complex enough to fit the data with special attention to the positive feedback as subject of our research interest. In accordance with our study purpose, we applied MCMC for model calibration as it enables an estimation of the posterior predictive distribution of any quantity of interest. As the experimental data did not suffice for separation in training and test data, we made use of additional experiments for model validation by thoughtfully mimicking new settings with our model and qualitatively comparing the results. Our simple model was able to sufficiently fit the experimental data from experiments used for training as well as the data used in the validation step. Thus, we finally analyzed our model in terms of the sustained ERK activation in the positive feedback structure. In this step we combined an analytical approach to steady state analysis with a simulation-driven approach to explain sustained ERK response in our model. The validated model was finally used to investigate different aspects of the sustained ERK activation.

Chapter 3.

The tumor suppressor protein DLC1 maintains protein kinase D activity and Golgi secretory function

3.1. Published manuscript and contributions

This chapter corresponds to the following contribution:

> *A. Jensch, Y. Frey, K. Bitschar, P. Weber, S. Schmid, A. Hausser, M. A. Olayioye, and N. E. Radde. "The tumor suppressor protein DLC1 maintains protein kinase D activity and Golgi secretory function". In:* J Biol Chem *293.37 (2018), pp. 14407–14416*

3.1.1. Abstract

Many newly synthesized cellular proteins pass through the Golgi complex from where secretory transport carriers sort them to the plasma membrane and the extracellular environment. The formation of these secretory carriers at the trans-Golgi network is promoted by the Protein Kinase D (PKD) family of serine-threonine kinases. Here, using mathematical modeling and experimental validation of the PKD activation and substrate phosphorylation kinetics, we reveal that the expression level of the PKD substrate Deleted in Liver Cancer 1 (DLC1), a Rho GTPase activating protein that is inhibited by PKD-mediated phosphorylation, determines PKD activity at the

Golgi membranes. RNAi-mediated depletion of DLC1 reduced PKD activity in a Rho-Rho-associated protein kinase (ROCK)-dependent manner, impaired the exocytosis of the cargo protein horse radish peroxidase and was associated with the accumulation of the small GTPase RAB6 on Golgi membranes, indicating a protein-trafficking defect. In summary, our findings reveal that DLC1 maintains basal activation of PKD at the Golgi and Golgi secretory activity, in part by down-regulating Rho–ROCK signaling. We propose that PKD senses cytoskeletal changes downstream of DLC1 to coordinate Rho signaling with Golgi secretory function.

3.1.2. Introduction

Protein Kinase D (PKD), comprising PKD1, PKD2 and PKD3, is a family of serine/threonine protein kinases that localizes to trans-Golgi network (TGN) membranes where it controls protein secretion [63, 152]. PKD further plays an important role in the regulation of actin cytoskeleton remodeling and cell motility [170]. Membrane recruitment of PKD depends on diacylglycerol (DAG), a lipid second messenger that also activates novel PKCs, which in turn phosphorylate and activate PKD [18]. At the Golgi membranes, PKD phosphorylates and regulates the lipid kinase PI4KIIIβ [92], the BAR domain protein Arfaptin-1 [71] and the lipid transfer proteins CERT and OSBP [65, 168], which together coordinate the formation and budding of secretory vesicles. Although the biochemical mechanisms underlying PKD activation and its downstream substrates have been studied extensively, little is known about the negative regulation of PKD and whether PKD activity is subject to any feedback regulation. While it is intuitively clear that Golgi secretory activity must adapt to changes in the cellular environment, how extracellular cues and signals emanating from the plasma membrane are relayed to the Golgi complex to coordinate secretion is still elusive.

Rho proteins are ubiquitously expressed small GTPases that coordinate actin and microtubule cytoskeleton rearrangements, thereby regulating diverse cellular processes such as cell adhesion and migration, cell division and membrane trafficking [107]. When bound

to GTP, Rho GTPases can associate with different effector proteins triggering the activation of downstream signaling. The cycle between the active GTP-bound and inactive GDP-bound state is regulated by the GEF proteins that promote the exchange of GDP for GTP, whereas GAP proteins accelerate the intrinsic GTPase activity leading to the inactivation of the Rho protein [32]. The RhoA isoform is associated mainly with the plasma membrane where it controls actin stress fiber formation and acto-myosin contraction. Overexpression of constitutively active RhoA or the GEF protein Lbc was shown to increase basal PKD activity [248]. Later, RhoA activation induced by oxidative stress or the loss of cell-cell adhesions was reported to increase PKD activity by a mechanism involving novel PKCs and the cytoplasmic kinases ROCK and Src [46, 214]. Vice versa, PKD has also been implicated in controlling RhoA activity, for example, by the direct phosphorylation and stabilization of the RhoA effector protein rhotekin [179] and by the functional inactivation of the RhoGAP DLC1 [204]. Upon PKD-mediated phosphorylation, DLC1 is bound and sequestered by 14-3-3 proteins, thereby preventing it from inactivating Rho-GTP. These observations raise the question whether a positive feedback involving Rho GTPase signaling exists, maintaining cellular PKD activity and thus Golgi secretory function.

Here we test this hypothesis by a data-driven modeling approach that captures PKD activation in dependence of DLC1-mediated Rho regulation. Intriguingly, our model anticipates a DLC1-dependent negative effect of Rho signaling on PKD activity. These predictions were confirmed in subsequent cellular experiments, which further uncovered a novel role for DLC1 in the regulation of protein secretion from TGN membranes.

3.1.3. Results

3.1.3.1. Reciprocal activation of PKD and RhoA

To address if PKD activity might be subject to Rho-dependent feedback regulation we first examined whether the expression of active RhoA leads to PKD activation and concomitant inhibitory phosphorylation of the RhoGAP DLC1, a direct PKD substrate. Owing to

its GAP activity and tumor suppressor function, the overexpression of DLC1 is associated with strong morphological changes that can eventually lead to cell death. The endogenous protein, however, is expressed at very low levels that preclude quantitative analysis of DLC1 phosphorylation. We thus used as a model system the previously established Flp-In GFP-DLC1 cell line in which GFP-DLC1 expression can be induced by doxycycline addition [204]. These cells were transiently transfected to express constitutively active (ca) or dominant negative (dn) RhoA and treated with doxycycline for 16 h. Compared to the GFP vector control cells, expression of RhoA-ca increased whereas RhoA-dn suppressed PKD activity, as measured by an antibody that recognizes the autophosphorylated kinase (pPKD) (Fig. 3.1A,B). In cells expressing RhoA-ca, this was associated with increased DLC1 phosphorylation detected by a PKD substrate antibody reactive with the phosphorylated PKD consensus motif [52] (Fig. 3.1A,B). Note that HEK293 cells mainly express PKD2 and PKD3 and little PKD1. To test if active PKD, in turn, stimulates RhoA activation, we expressed in HEK293T cells a genetically encoded RhoA FRET biosensor together with either constitutively active (ca) or kinase-dead (kd) PKD1. In line with our hypothesis, expression of the active PKD1 significantly increased the FRET ratio of the RhoA biosensor measured in the cell lysates, whereas inactive PKD1 failed to do so (Fig. 3.1C). Immunoblotting of the cell lysates confirmed expression of the PKD variants (Fig. 3.1D). These results suggest a molecular pathway in which RhoA activates PKD, which phophorylates and functionally inactivates DLC1, potentially resulting in an overall positive feedback (Fig. 3.1E).

To confirm that the activation of endogenous PKD also leads to DLC1 substrate phosphorylation, we stimulated the cells with the microtubule-depolymerizing agent nocodazole which activates PKD at Golgi membranes [64]. Nocodazole treatment of Flp-In GFP-DLC1 cells expressing GFP-DLC1 elevated the PKD phosphorylation levels, which was blocked by the selective PKD inhibitor kb NB 142-70 (kb-NB) and more efficiently by the more potent but less specific PKC/PKD inhibitor Gö-6976 (Fig. 3.2A). This was accompanied by increased DLC1 phosphorylation, which was also suppressed by phar-

macological PKD inhibition (Fig. 3.2A). Significance of the effect of the inhibitors was investigated with an F-test. For this, we compared two parametrized model variants. In the null hypothesis H_0, the inhibitor does not act significantly, the alternative hypothesis assumes a significant influence of the inhibitor under investigation on the pPKD and pDLC1 time courses. Results are shown in Fig. 3.2B. According to our test statistics, the pPKD and pDLC1 time courses are significantly below those of the control experiments for the inhibitor Gö-6976, while the effect is not significant for the inhibitor kb-NB (see supporting Section B.1). Similar results were obtained upon stimulation of cells with the phorbol ester phorbol 12,13-dibutyrate (PDBu), an analogue of DAG. Phorbol ester treatment enhanced PKD and DLC1 phosphorylation, both of which could be blocked by kb-NB (Fig. B.1).

3.1.3.2. Computational modeling suggests negative feedback regulation of PKD activity

Based on the hypothesized model structure (Fig. 3.1E) and the measured outputs in the time series experiments, we formulated a model with two state variables that represent phosphorylated PKD and DLC1, respectively. We used mass action kinetics for phosphorylation and dephosphorylation kinetics,

$$\dot{\text{pPKD}} = k(\text{DLC1},\theta)\text{PKD} - \theta_1\text{pPKD} \tag{3.1a}$$

$$\dot{\text{pDLC1}} = \theta_2\text{pPKD} \cdot \text{DLC1} - \theta_3\text{pDLC1}. \tag{3.1b}$$

The PKD phosphorylation rate $k(\text{DLC1},\theta)$ depends on DLC1 via Rho and on the experimental treatment of the cells and is specified in Table B.4. We eliminated unphosphorylated PKD and DLC1 by assuming mass conservation of respective total amounts and normalized both state variables accordingly (supporting Sections B.1 and B.2). For model calibration we exploited maximum-likelihood estimation (MLE), which requires the choice of an appropriate error model for observed outputs. Selection of an error model was done as a data-driven pre-processing step, which is computationally more efficient

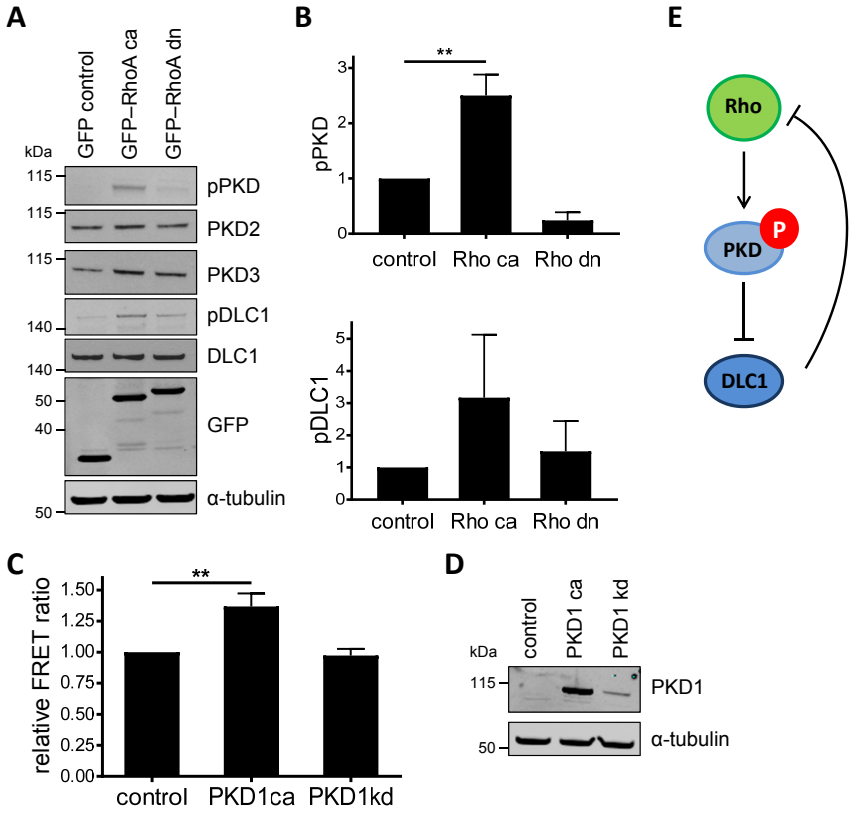

A
GFP control
GFP–RhoA ca
GFP–RhoA dn
kDa
115
115
115
140
140
50
40
50
pPKD
PKD2
PKD3
pDLC1
DLC1
GFP
α-tubulin
B
**
pPKD
3
2
1
0
control
Rho ca
Rho dn
pDLC1
5
4
3
2
1
0
control
Rho ca
Rho dn
E
Rho
P
PKD
DLC1
C
**
relative FRET ratio
1.50
1.25
1.00
0.75
0.50
0.25
0.00
control
PKD1ca
PKD1kd
D
control
PKD1 ca
PKD1 kd
kDa
115
50
PKD1
α-tubulin

Figure 3.1.: **Reciprocal activation of PKD and RhoA.** (A) Flp-In GFP-DLC1 cells were transiently transfected with vectors encoding GFP, constitutively active (ca) or dominant negative (dn) GFP-RhoA. Six hours post transfection, GFP-DLC1 expression was induced with doxycycline. The next day cells were lysed and lysates analyzed by immunoblotting. (B) Band intensities from three independent experiments were quantified and normalized to the loading control and control sample (mean ± SEM). (C) HEK293T cells were transiently cotransfected with vectors encoding PKD1 ca or kinase-dead (kd) and a RhoA FRET biosensor [176]. The next day cells were lysed and FRET ratio was analyzed. Shown are the mean FRET ratios from three independent experiments normalized to the control (± SEM). (D) Cell lysates from (C) were analyzed by immunoblotting. (E) Positive feedback hypothesis in which Rho activates PKD, which phosphorylates and inactivates the RhoGAP DLC1, to support further Rho activation. Data in (B) and (C) were analyzed by one-way ANOVA followed by Dunnett's multiple comparisons test. Only statistically significant changes are indicated. **$p<0.01$

than a combination of error model selection and model calibration. We set up different error models by combining additive normal and multiplicative log-normal error models with biased and unbiased variance estimators. We additionally compared different variants of pooling standard deviations. Comparison was done in terms of suitable information criteria, as shown in Table B.2 in supporting Section B.1. Based on these results we decided to use an additive normal error model and six standard deviations for the modeling study. For each experiment we employed a point control normalization to the highest signal value to avoid normalization to low signals with a low signal-to-noise ratio [50]. The resulting optimization problem consisted of 16 parameters in total (see supporting Section B.2).

Fig. 3.2C shows the calibrated courses of pPKD and pDLC1 after addition of nocodazole in the control case versus treatment with the PKD inhibitors kb-NB and Gö-6976 (compare Fig. 3.2A). Experimental data are indicated by dots. Trajectories were obtained by taking all parameters from the optimization runs into account that gave reasonable model fits. The response to nocodazole treatment is well captured, although the steady state value prior to nocodazole addition is slightly underestimated for both variables in the control case. Both PKD inhibitors reduce the phosphorylation rate of PKD and thus slow down the dynamics of the system, which is more pronounced for Gö-6976 than for kb-NB, as suggested by the data. Respective model fits to measurements of the system response after treatment with PDBu (Fig. B.1) and overexpression of Rho (Fig. 3.1A,B) show that the model also captures these experiments (Fig. B.5). Our approach further provided estimates for standard deviations of measured outputs, which are in good agreement with the empirical ones (Fig. B.4).

Plausibility of the model was tested by a parametric bootstrapping approach, in which we used the inferred stochastic model to generate many datasets with the same size and structure as the experimental data used for model calibration. Then we calculated the likelihood of these datasets by using simulations with the maximum likelihood estimator θ_{MLE} and used these values to estimate a probability density $p(J_{\mathrm{opt}})$ via kernel density estimation. This was compared to the like-

lihood value of the real experimental data (Fig. 3.2D and supporting Section B.2). The objective function value of the MLE falls into the center of this distribution, indicating that we are neither in a poor fitting regime nor that we encounter an overfitting problem. Overall, the model is able to describe all experiments and model granularity constitutes a good trade-off between complexity and flexibility to adapt to different experimental conditions.

We also analyzed values and correlations of inferred parameters (Fig. B.3). Strikingly, the optimizer consistently assigned very small values to the parameter θ_6, a measure for the influence of DLC1 on PKD. Given this, PKD dynamics do not seem to be affected by DLC1 in the inferred model. Such a qualitative statement about the network structure, however, cannot only be based on dimensionless parameter values that have been rescaled in the normalization procedure and cannot directly be compared to another. Thus, we used the model to quantify the influence of DLC1 directly onto pPKD by simulating the expected fold change in pPKD in response to altered DLC1 total amounts, which is reflected by the scaling parameter s_1 in Fig. 3.3. Fig. 3.3A shows that pPKD is only minimally affected, even when DLC1 amounts are very low, suggesting that DLC1 does not inhibit PKD activity.

We then calculated the profile likelihood of the feedback parameter θ_6 around the maximum-likelihood value (Fig. 3.3B). Since $\widehat{\theta}_6$ is very small, we also allowed negative θ_6 values, which corresponds to a sign change of the influence of DLC1 onto pPKD in the network and hence implies a new network structure referred to as model 2. Surprisingly, the overall fit quality improved with negative θ_6 values, with an optimum at $\theta_6^* = -240$. The expected fold change in pPKD using θ_6^* as a new estimate is shown in Fig. 3.3C. Here, pPKD levels are tightly regulated by DLC1 and increase with increasing DLC1 amounts in an almost perfect linear way. Our model simulations and the profile likelihood analysis are thus supportive of an overall negative rather than a positive feedback.

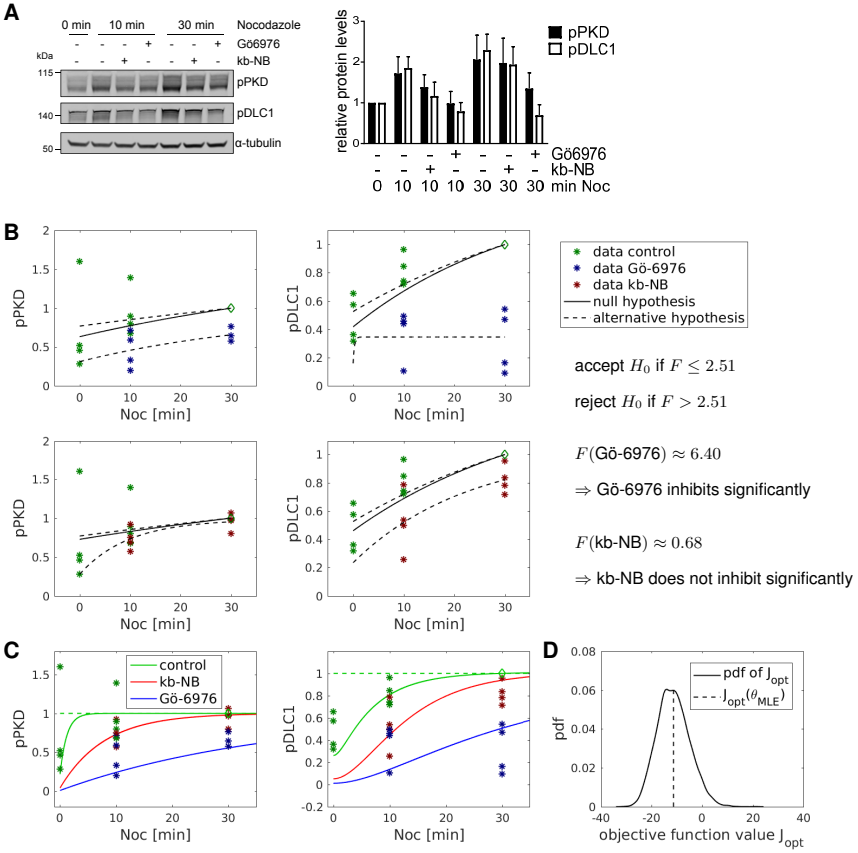

A
0 min
10 min
30 min
Nocodazole
Gö6976
kb-NB
kDa
115
140
50
pPKD
pDLC1
α-tubulin
relative protein levels
pPKD
pDLC1
0 10 10 10 30 30 30 min Noc
B
pPKD
pDLC1
Noc [min]
data control
data Gö-6976
data kb-NB
null hypothesis
alternative hypothesis
accept H_0 if $F \leq 2.51$
reject H_0 if $F > 2.51$
F(Gö-6976) ≈ 6.40
⇒ Gö-6976 inhibits significantly
F(kb-NB) ≈ 0.68
⇒ kb-NB does not inhibit significantly
C
control
kb-NB
Gö-6976
D
pdf of J_{opt}
$J_{opt}(\theta_{MLE})$
pdf
objective function value J_{opt}

Figure 3.2.: **Modeling of nocodazole-induced PKD dynamics.** (A) Flp-In GFP-DLC1 cells were treated with doxycycline. The next day, the cells were treated with kb-NB or Gö-6976 for 2 h, followed by nocodazole treatment for the times indicated. Cells lysates were analyzed by immunoblotting. Band intensities from four independent experiments were quantified and normalized to the loading control and control sample (mean ± SEM). (B) Significance of the effect of the inhibitors Gö-6976 and kb-NB was investigated via an F-test, according to which Gö-6976 suppresses phosphorylation time courses of PKD and DLC1 significantly with a 5% level of significance, while the null hypothesis H_0 cannot be rejected and hence the effect of kb-NB is not significant. (C) Dots indicate re-normalized experimental data from (A), with normalization points denoted by diamonds, together with trajectories of the calibrated model that all lie on top of each other. (D) Model validation via a bootstrapping approach, in which the inferred stochastic model was used to resample new experimental data to estimate the distribution of the maximum-likelihood objective function value.

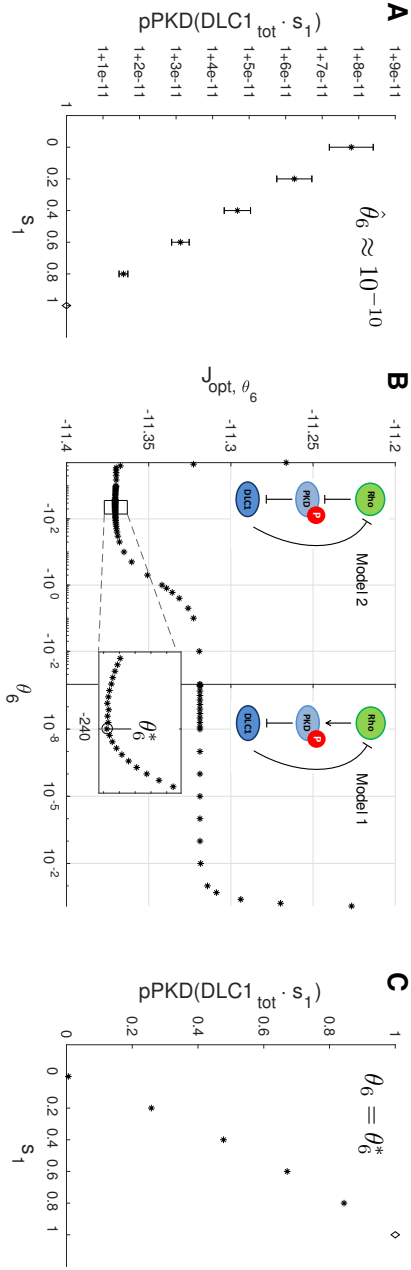

A
pPKD(DLC1 tot · s1)
1+9e-11
1+8e-11
1+7e-11
1+6e-11
1+5e-11
1+4e-11
1+3e-11
1+2e-11
1+1e-11
$\hat{\theta_6} \approx 10^{-10}$
s1
B
Jopt, θ6
-11.2
-11.25
-11.3
-11.35
-11.4
Rho
PKD
P
DLC1
Model 2
Model 1
θ6
θ_6^*
-240
C
pPKD(DLC1 tot · s1)
$\theta_6 = \theta_6^*$
s1

Figure 3.3.: **Mathematical modeling suggests a DLC1-dependent negative feedback of Rho signaling on PKD activity.** (A) Fold changes in pPKD upon changes in the DLC1 total amount as predicted by the model. (B) Profile likelihood for the feedback parameter θ_6. (C) Using model 2, pPKD clearly decreases with decreasing DLC1 amounts.

3.1.3.3. Experimental manipulation of DLC1 levels confirm Rho-Rock dependent negative regulation of PKD activity

To experimentally test the model prediction that PKD activity is sensitive to DLC1 expression levels, we first increased cellular DLC1 expression by doxycycline addition in Flp-In GFP-DLC1 cells and, secondly, reduced the endogenous DLC1 expression by RNAi-mediated DLC1 downregulation, followed by the measurement of PKD autophosphorylation. Intriguingly, comparison of the PKD phosphorylation in doxycycline-treated Flp-In GFP-DLC1 cells with the PKD phosphorylation in untreated cells revealed an increase by 1.5-fold (Fig. 3.4A). Next, we downregulated DLC1 by siRNA transfection in uninduced Flp-In GFP-DLC1 cells (Fig. 3.4C), resulting in the reduction of basal PKD phosphorylation (Fig. 3.4B, 0 min). Of note, PKD phosphorylation was still increased upon nocodazole treatment of cells (Fig. 3.4B, 10 and 30 min), ruling out a general inactivation mechanism. The suppression of basal PKD phosphorylation in cells depleted of DLC1 was partially rescued by pharmacological ROCK inhibition (H1152), indicating that Rho-ROCK signaling suppresses PKD in DLC1-depleted cells, whereas ROCK inhibition had no effect on the pPKD levels in the control (siNT) cells (Fig. 3.4D). Importantly, reduced PKD phosphorylation by DLC1 depletion was confirmed using independent siRNAs (Fig. B.11A). These data strongly support model 2, wherein DLC1, by downregulating Rho-ROCK signaling, contributes to the activation of PKD.

We extended model 2 to include these new experiments. The resulting model fit is shown in Fig. 3.5. Details of the revised model and the estimation procedure are given in supporting Sections B.3 and B.4. Larger variability of trajectories result from the additional constraints in the optimization problem due to the new experiments. Comparing trajectories of models 1 and 2 for the nocodazole experiments (Fig. 3.2C), the negative feedback slows down the system dynamics and compensates for the differences in the effectiveness of the two inhibitors. All other experiments are also well captured (Fig. B.7). Plausibility of the revised model was again confirmed by a bootstrapping approach (Fig. 3.5E). Taken together, the revised

model is superior in fitting the experimental data, thus strengthening the hypothesis that, downstream of DLC1, Rho negatively controls PKD activity, resulting in an overall negative feedback.

3.1.3.4. DLC1 regulates PKD activity at Golgi membranes and protein secretion

The primary localization of PKD is at Golgi membranes, but PKD was also found to associate with the plasma membrane, mitochondria and translocate to the nucleus. To determine whether the Golgi-localized PKD pool is sensitive to DLC1 regulation, we employed a previously described PKD reporter that allows the specific determination of PKD activity at the Golgi membranes [64]. Indeed, in HEK293T depleted of DLC1, phosphorylation of the Golgi-localized PKD reporter as measured in cell lysates was reduced by 50% (Figs. 3.6A and B.11B). We confirmed these results by ratiometric imaging of PKD reporter phosphorylation, measuring specifically the signal intensity at the Golgi membranes (Fig. 3.6B). Compared to the control cells, in cells lacking DLC1 Golgi-localized PKD activity was significantly reduced.

PKD activity at the TGN is essential for secretory vesicle formation. Considering the novel molecular link between DLC1 and PKD, we reasoned that DLC1 depletion should affect Golgi function. To quantitatively measure protein secretion, we used Flp-In T-REx 293 cells inducibly expressing FLAG-labeled horse radish peroxidase (HRP) fused to a signal peptide that directs HRP to the secretory pathway. ssHRP is a well-characterized model cargo that is known to be secreted in a PKD-dependent manner [16, 65]. Compared to siRNA-transfected control cells, the HRP activity measured in the supernatant of cells lacking DLC1 was significantly reduced and was similar to the HRP activity contained in the supernatants of cells treated with a PKD inhibitor (Fig. 3.6C).

Finally, we sought to validate our results in an independent cell line. In agreement with the observations in HEK293T cells, silencing of DLC1 in U2OS cells reduced basal and nocodazole-induced PKD phosphorylation (Fig. 3.6D). The post-Golgi carriers produced by PKD are known to be positive for the small GTPase RAB6 [237].

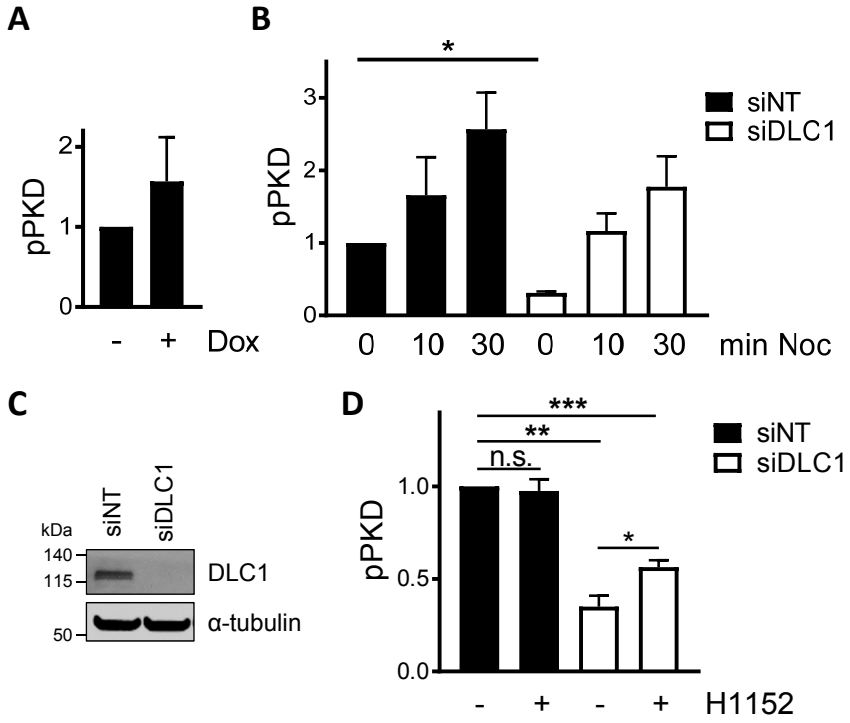

A
pPKD
2
1
0
- + Dox
B
*
pPKD
3
2
1
0
siNT
siDLC1
0 10 30 0 10 30 min Noc
C
siNT
siDLC1
kDa
140
115
DLC1
50
α-tubulin
D
***
**
n.s.
*
pPKD
1.0
0.5
0.0
siNT
siDLC1
- + - + H1152

Figure 3.4.: **PKD activity positively correlates with DLC1 expression.** (A) Flp-In GFP-DLC1 cells were induced with doxycycline and PKD activation was analyzed by immunoblotting. Band intensities from three independent experiments were quantified and normalized to the loading control and control sample ± SEM. (B) Uninduced Flp-In GFP-DLC1 cells were transfected with control siRNA (siNT) or DLC1-specific siRNA (siDLC1). After 72 h, the cells were treated with nocodazole as indicated, lysed and PKD activation was analyzed by immunoblotting. Band intensities from three independent experiments were quantified and normalized to the loading control and control sample ± SEM. *$p<0.05$ (one-sample t-test) (C) Validation of DLC1 knockdown by immunoblotting. (D) Uninduced Flp-In GFP-DLC1 cells were transfected with the indicated siRNAs. After two days, cells were treated with H1152 where indicated. PKD activation was analyzed by immunoblotting. Band intensities from three independent experiments were quantified and normalized to the loading control and control sample ± SEM. *$p<0.05$ (paired two-sample t-test), **$p<0.01$, n.s. = not significant (one sample t-test).

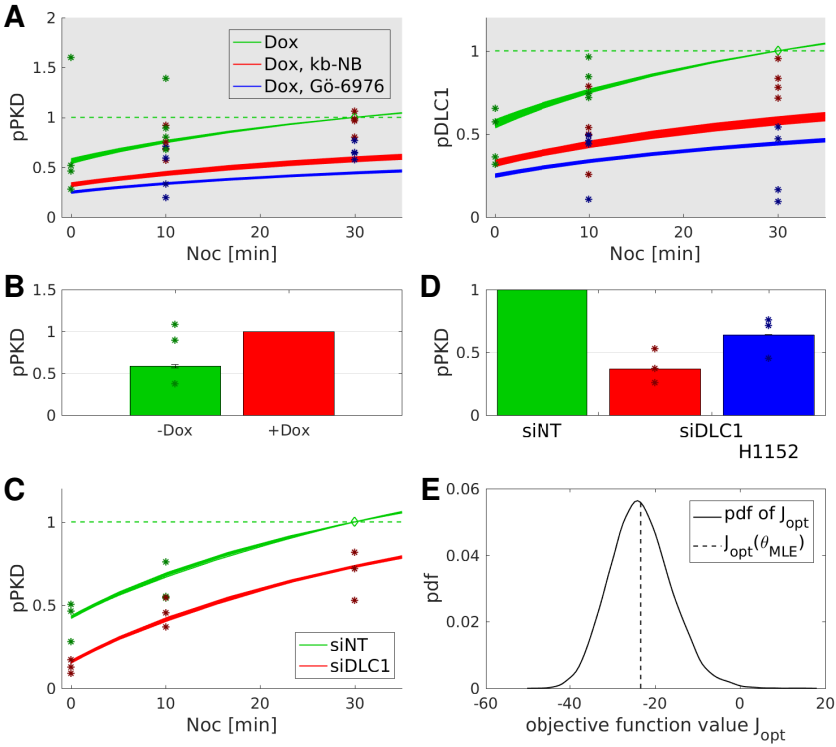


Figure 3.5.: **The negative feedback model 2 is superior** in fitting all experimental data. Dots indicate re-normalized experimental data, with normalization points denoted by diamonds. (A) Nocodazole induced PKD activation (cf. Fig. 3.2A) (B) pPKD fold change (cf. Fig. 3.4A) (C) PKD activation dynamics (cf. Fig. 3.4B) (D) pPKD fold change (cf. Fig. 3.4D) (E) Model plausibility was tested by a bootstrapping approach.

Whereas RAB6 was dispersed in the control cells, in cells lacking DLC1, RAB6 accumulated at the Golgi complex (visualized by GM130 staining), indicative of a trafficking defect (Fig. 3.6E). Note that the Golgi complex also appeared more compact in DLC1-depleted cells. Based on these findings we conclude that PKD activity at the TGN membranes and Golgi secretory function are positively regulated by DLC1.

3.1.4. Discussion

Here we used a combination of experiments on the PKD regulation network and mathematical modeling of phosphorylation dynamics to describe the molecular interactions between PKD, Rho and DLC1. Our study reveals a previously unknown molecular connection between DLC1 and PKD that controls the basal activation state of PKD at the TGN membranes and Golgi secretory function. TGN-derived vesicles formed by PKD deliver cargo to the plasma membrane [152]. Cargo is also specifically delivered to focal adhesions [216], cell adhesion sites that connect the intracellular actin cytoskeleton via integrins with the extracellular matrix. In many cell types including the U2OS cells used in this study DLC1 localizes to FAs (Fig. B.11C), to which it is recruited via the binding of talin and tensin adaptor proteins [34]. Knockdown of DLC1 causes an increase in actin stress fibers and FAs, consistent with the elevated Rho signaling in cells lacking DLC1 [99]. The downregulation of PKD activity in the absence of DLC1 could thus be the result of a homeostatic feedback, requiring no further cargo delivery to FAs. It was recently reported that PKD also localizes to FAs [53], raising the possibility that DLC1 directly affects this particular PKD pool. However, using a Golgi-localized PKD reporter we clearly show that it is the basal PKD activity at the TGN membranes that is sensitive to the DLC1 expression levels.

Negative feedback is generally known to stabilize systems, e.g. by making signaling pathways robust against variations in total protein concentrations, as demonstrated for ERK activity in the MAP kinase cascade [62]. Here, we observe that the DLC1 expression level has a strong influence on basal PKD activity (Figs. 3.3 and B.10).

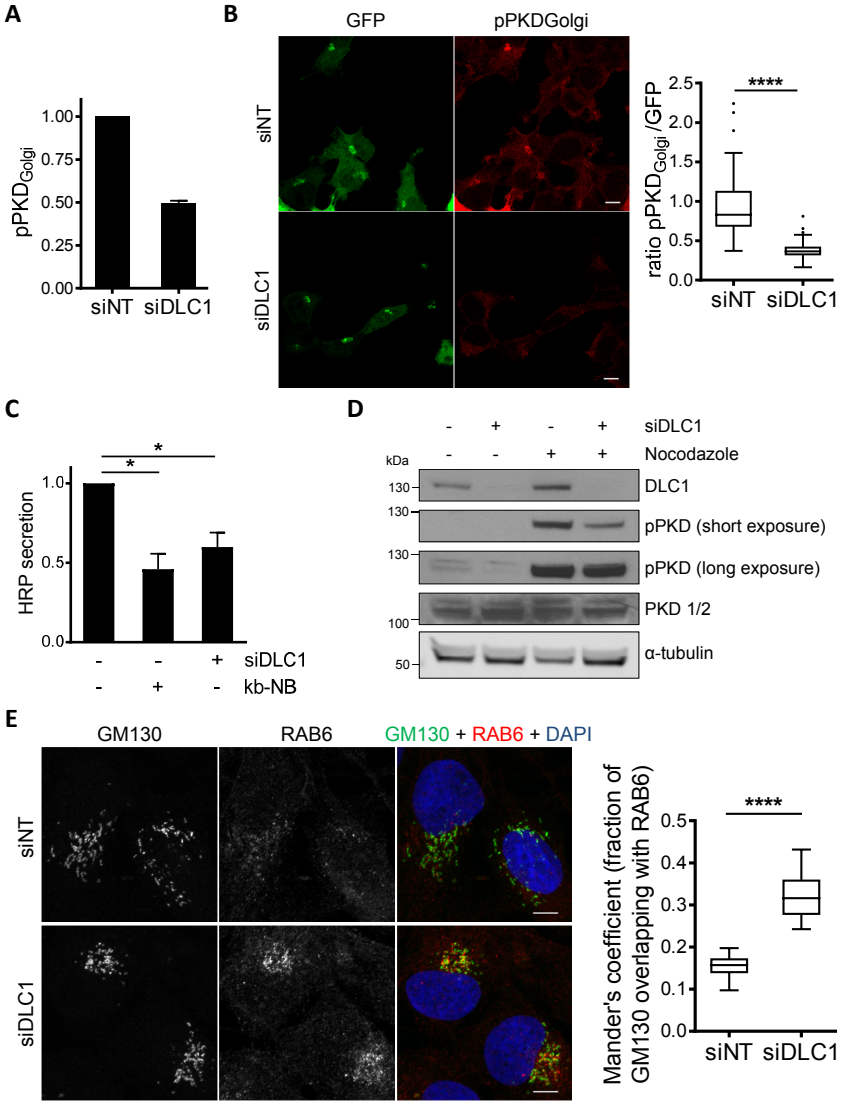

A
pPKD$_{Golgi}$
1.00
0.75
0.50
0.25
0.00
siNT siDLC1
B
GFP
pPKDGolgi
siNT
siDLC1
ratio pPKD$_{Golgi}$/GFP
2.5
2.0
1.5
1.0
0.5
0.0
****
siNT siDLC1
C
HRP secretion
1.0
0.5
0.0
*
*
- - + siDLC1
- + - kb-NB
D
- + - + siDLC1
kDa - - + + Nocodazole
130 DLC1
130 pPKD (short exposure)
130 pPKD (long exposure)
PKD 1/2
100
50 α-tubulin
E
GM130
RAB6
GM130 + RAB6 + DAPI
siNT
siDLC1
Mander's coefficient (fraction of GM130 overlapping with RAB6)
0.5
0.4
0.3
0.2
0.1
0.0
****
siNT siDLC1

Figure 3.6.: **DLC1 depletion impairs protein secretion.** (A,B) Two days post siRNA transfection, HEK293T cells were transfected with the vector encoding the PKD reporter. (A) The next day, cells were lysed and lysates analyzed by immunoblotting. The $pPKD_{Golgi}$-Reporter signal was normalized to the GFP signal. Shown are the mean values of two independent experiments ± SEM (B) The next day, cells were fixed and stained with an antibody reactive with the phosphorylated reporter. PKD activity at the Golgi was determined by ratiometric imaging of 92 cells from two independent experiments. Scale bar=10 μM. ****$p<0.0001$ (two-sample t-test). (C) Flp-In T-REx 293 Flag-HRP cells were transfected with the indicated siRNAs. Expression of FLAG-HRP was induced with doxycycline. The next day, the medium was replaced with serum-free medium containing either kb-NB or DMSO. The supernatant was collected after 5h for HRP measurements. *$p<0.05$ (one-sample t-test). (D) U2OS cells were transfected with the indicated siRNAs. After 3 days, the cells were treated with nocodazole or DMSO for 10 minutes and lysed. The lysates were analyzed by immunoblotting with the indicated antibodies. (E) U2OS cells were transfected with the indicated siRNAs. After three days, the cells were fixed and stained with antibodies specific for GM130 and RAB6. The images shown are maximum intensity projections of several confocal sections. Co-localization of GM130 and RAB6 (n=11; N=3) was analyzed with ImageJ. Scale bar=10 μM. ****$p<0.0001$ (two-sample t-test)

Similarly, phosphorylated DLC1 is very sensitive to changes in the amount of PKD. Surprisingly, this does not result in strong negative feedback regulation and robustness of pPKD to variations in total PKD concentrations. Very small fractions of phosphorylated PKD and DLC1 relative to the total protein amounts might provide an explanation, reminiscent of our previous finding that PKD activity is relatively insensitive to changes in CERT protein levels (c.f. [240], Fig. 5). Although alterations of the PKD protein level result in considerable fold changes in DLC1 phosphorylation, the abundance of unphosphorylated DLC1, which regulates PKD activity as part of the feedback, appears to buffer the system. If this is the case, we anticipate that regulation via the feedback is hardly visible in the dynamics of PKD activity upon nocodazole stimulation, since the fold change of unphosphorylated DLC1 is not much affected in this scenario. Consequently, when comparing the time courses of pPKD in control and siDLC1 experiments after stimulation with nocodazole, we expect that pPKD increases in both scenarios, and that PKD activity in the siDLC1 experiment remains constantly below those of the control experiment, which is in agreement with Fig. 3.4. Although Rho can activate PKD in response to stress signals [46, 214], these Rho pools must be spatially distinct from the one regulated by DLC1, which inactivates PKD. We have not observed any colocalization of DLC1 with Golgi markers, but we cannot rule out that DLC1 directly regulates PKD at the TGN, as both RhoA and ROCK have been found at Golgi membranes [176, 182] and actin remodeling has been associated with cargo sorting at the TGN [28, 197]. Regardless of where DLC1 exerts its function, PKD at the Golgi membranes appears to sense the cellular F-actin state whereby secretion is coordinated. A challenge for the future is the identification of the downstream signaling molecules that link Rho-ROCK with Golgi-localized PKD. 14-3-3 protein and chaperon p32 binding were previously reported to decrease PKD activity [93, 221], but the psysiological conditions that engage these regulatory mechanisms are unknown. 14-3-3 binding is mediated by phosphorylation of serines 205/208 (RRLSNVSLT) and serines 219/223 (IRTSSAELST) within the C1 domain of activated PKD by a yet unknown upstream kinase. Intriguingly, serines 205

and 219 match the consensus sequence for ROCK kinases [10], raising the possibility that ROCK directly phosphorylates and negatively regulates PKD. In DLC1 knockdown cells, pharmacological ROCK inhibition only partially rescued the PKD phosphorylation, thus other Rho effectors or Rho-independent mechanisms could also be involved in the regulation of PKD and secretory trafficking. For example, p122RhoGAP, the rat homolog of DLC1, was initially identified as a phospholipase C δ1-binding protein that activates its PIP2-hydrolyzing activity [101]. Thus, DAG produced by PIP2 hydrolysis may contribute to PKD activation, although a later study failed to demonstrate stimulation of PLCdδ1 enzyme activity by human DLC1 [95]. DLC1 further comprises a START domain, which is found in a number of lipid transfer proteins [9]. The lipid ligand for the DLC1 START domain still remains to be identified. START domain-mediated lipid transfer could be involved in the modification of the local membrane lipid composition or it could provide specific lipid metabolizing enzymes with their substrate, thereby contributing to PKD recruitment and/or activation at the TGN membranes.

In different types of human cancers, the expression of DLC1 is frequently downregulated due to gene deletion or promoter hypermethylation [177]. Work in cancer cell lines and mouse models of cancer have revealed a tumor and metastasis suppressor function for DLC1. In cells lacking DLC1, the aberrant Rho signaling and actin remodeling could lead to the missorting of cargo at the level of the Golgi, altering the cellular secretome and the communication of the cancer cells with the microenvironment. Considering that the different PKD isoforms have been associated with both oncogenic and tumor suppressive functions in dependence of the tumor context [194], it is tempting to speculate that DLC1 loss could support neoplastic transformation in part by dampening PKD activity. Alternatively, elevated PKD activity could lead to the functional inactivation of DLC1 by phosphorylation and sequestration through 14-3-3 protein binding. In future studies it will be of particular interest to clarify the reciprocal regulation of DLC1 and PKD in cancer cells of different tissue origin.

3.1.5. Experimental Procedures

3.1.5.1. Antibodies

Primary antibodies used were: mouse mAb anti-DLC1 (#612020) and mouse mAb anti-GM130(#610823, both BD Biosciences), rabbit pAb anti-PKCμ (D20) to detect PKD 1+2, rabbit pAb anti-GFP (FL) (sc-8334) and rabbit pAb anti-paxillin (H-114, all Santa Cruz Biotechnology, Dallas, TX, USA), rabbit pAb anti-Phospho-(Ser/Thr) PKD Substrate Antibody (#4381) to detect pDLC1, rabbit mAb anti-PKD2 (D1A7), rabbit mAb anti-PKD3 (D57E6) and rabbit pAb anti-RAB6 (D37C7, all Cell Signaling Technology, Danvers, MA, USA), mouse Ab anti-GFP (#11814460001, Roche Applied Sciences, Basel, Switzerland). The polyclonal rabbit antibodies against PI4KIIIβ pS294 and autophosphorylated PKD (pS910 in human PKD1) were described previously [91, 92]. For immunoblotting the following secondary antibodies were used: donkey anti-rabbit IgG and goat anti-mouse IgG coupled to IRDye 680RD or IRDye 800 LW (all LI-COR Biosciences, Lincoln, NE, USA) and HRP-conjugated sheep anti-mouse IgG and anti-rabbit IgG (both GE Healthcare, Piscataway, NJ, USA). Secondary antibodies used for immunofluorescence were goat anti-mouse IgG and anti-rabbit IgG coupled to AlexaFluor488 or AlexaFluor 546 (both Thermo Fisher Scientific, Waltham, MA, USA), respectively.

3.1.5.2. Plasmids and siRNAs

Vectors encoding constitutively active (ca, Q63L) and dominant negative (dn, T19N) RhoA were generated from a wild-type RhoA construct (kindly provided by John Collard from The Netherlands Cancer Institute, Amsterdam, The Netherlands) by a PCR approach using the QuickChange site-directed mutagenesis system (Stratagene California, La Jolla, CA, USA) according to the manufacturer's instructions. Expression vectors encoding kinase-dead (kd, K612W) and ca (S738E/S742E) PKD1 (PKCμ) were described previously [93]. pTriEx-RhoA FLARE.sc Biosensor WT was a gift from Klaus Hahn (Addgene plasmid #12150). The Golgi PKD activity reporter was

described previously [64]. ON-TARGETplus® non-targeting control SMARTpool siRNA (D-001810-10, Dharmacon, Lafayette, CO, USA) was used as negative control (siNT). siMAX DLC1 siRNA (siDLC1) was custom synthesized by MWG Biotech, Ebersberg, Germany (5'-UUAAGAACCUGGAGGACUATT-3'). Custom designed Silencer®Select human DLC1 siRNAs (siDLC1#2, s530697 and siDLC1#3, s530699) were from Thermo Fisher.

3.1.5.3. Cell culture

HEK293T cells and Flp-In T-REx 293 Flag-HRP cells were cultured in RPMI 1640 (Thermo Fisher) supplemented with 10% FCS at 37°C in a humidified atmosphere of 5% CO_2. Flp-In T-REx 293 GFP-DLC1 cells and U2OS cells were grown in DMEM (Thermo Fisher) supplemented with 10% FCS. TurboFect (Thermo Fisher) was used for plasmid transfections and Lipofectamine RNAiMAX (Thermo Fisher) for siRNA transfections. The following reagents were used: Doxycycline (10 ng/ml), Gö-6976 (5 μM), H1152 (10 μM) and nocodazole (10 μM, all from Merck, Darmstadt, Germany), kB NB-172-40 (5 μM) and PDBu (100 nM, Tocris Bioscience, Bristol, United Kingdom).

3.1.5.4. Western blotting

Cells were lysed in NEB extraction buffer [50 mM Tris (pH 7.5), 150 mM NaCl, 1% NP40, 1 mM sodium orthovanadate, 10 mM sodium fluoride, and 20 mM β-glycerophosphate, Complete protease inhibitors (Roche)] and lysates were clarified by centrifugation at 16,000xg for 10 min. Equal amounts of protein were loaded onto NuPAGE® 4 - 12% Bis-Tris precast gels (Thermo Fisher) and then blotted onto PVDF membrane using the iBlot device (Thermo Fisher). Membranes were blocked in 0.5% blocking reagent (Roche) in PBS containing 0.1% Tween-20 in PBS-T for 1 hour at room temperature, incubated with primary antibodies, followed by HRP-conjugated or IRDye-conjugated secondary antibodies and visualization was carried out using the ECL detection system (Thermo Fisher) or the Odyssey

device (LI-COR), respectively. Quantification of signals was carried out with the Odyssey imaging software.

3.1.5.5. Rho biosensor measurements

Cells were transiently transfected with a plasmid encoding the RhoA FLARE.sc biosensor along with PKD1 expression vectors. 24 h after transfection cells were lysed in FRET buffer (50 mM Tris (pH 7.5), 5 mM β-glycerophosphate, 5 mM sodium fluoride, 0.5% Triton X-100) and FRET ratio was measured in a multiwell plate reader (Tecan, Männedorf, Switzerland). CFP was excited at 433 nm and CFP emission was detected at 475 nm. The FRET signal was measured by exciting CFP and detecting Citrine emission at 527 nm. The FRET ratio was calculated by dividing the FRET signal by the CFP signal.

3.1.5.6. Ratiometric imaging to determine Golgi PKD activity

Ratiometric imaging of the Golgi PKD activity reporter was carried out as previously described [64]. Briefly, HEK293T cells were seeded on collagen-coated glass cover slips and transfected with siRNAs. After two days, cells were transfected with the vector encoding the Golgi PKD activity reporter. The next day, cells were fixed with 4% paraformaldehyde (PFA), permeabilised with 0.2% Triton X-100 and stained with an antibody reactive with the phosphorylated reporter and Alexa546-labelled secondary antibody. Samples were mounted with Fluoromount G (Southern Biotech, Birmingham, AL, USA) and analyzed with a confocal laser scanning microscope (LSM710, Zeiss, Jena, Germany). EGFP was excited with the 488-nm line of the argon laser and emission was detected in the spectral window 496–553 nm. Alexa546 was excited with the 561-nm line of a DPSS laser and emission was detected from 563-621 nm. Laser powers were adjusted to prevent fluorophore saturation and identical laser settings were maintained throughout the experiment. Maximum intensity projections of confocal stacks were analyzed with the ZEN software (Zeiss). In reporter expressing cells, the Golgi region of interest was

defined in the EGFP channel. With the output mean intensity values, the ratio of Alexa546 to EGFP signal was calculated.

3.1.5.7. Immunofluorescence

U2OS cells were seeded on collagen-coated glass cover slips and transfected with 10 nM siRNAs. After three days, cells were fixed with 4% PFA. Staining and imaging was carried out as described above. Quantification of co-localization was carried out with the JaCoP plugin in ImageJ [30].

3.1.5.8. HRP secretion assay

Three days post siRNA transfection, Flp-In HEK293 Flag-HRP cells were replated into 24 well plates. 6 hours after seeding Flag-HRP secretion was induced by doxycycline addition (10 ng/ml). The next day, the medium was replaced by phenol red and serum free medium containing doxycycline and the PKD inhibitor kb-NB 142-70 or DMSO, respectively. The supernatant was collected after 5 h for HRP activity detection by addition of ECL reagent and measurement of the chemiluminescent signal using a multiplate reader (Tecan).

3.1.5.9. Statistical analysis

Data are shown as mean ± S.E.M.; 'n' refers to the number of analyzed cells or images and 'N' to the number of independent experiments. Statistical significance was analyzed by the indicated statistical tests (GraphPad Prism version 7.03; GraphPad Software, La Jolla, CA, USA). p-values below 0.05 were considered as significant (*$p<0.05$; **$p<0.01$; ***$p<0.001$; ****$p<0.0001$).

3.1.6. Contributions

In this project, much effort went into determining a mathematical model to describe the first and second biological hypothesis of the network structure. I was actively involved in the discussions of possible modeling options in this important step. Likewise, I played an

active role in researching, selecting and defining the applied methods and analysis tools. Additionally, I implemented the model and all analysis tools (up to the usage of existing software). Consequently, I performed all model analysis and discussed results with the other authors. Finally, I also assisted in writing the manuscript, mainly contributing to the supplementary information, which contains details on the methods and tools applied in the study.

3.2. Results in the overall context

The goal of this project was the refinement of knowledge on the DLC1-PKD regulation complex. In this collaborative work, the network topology involving DLC1, PKD and Rho was investigated by a combination of experimental and mathematical modeling approaches.

In the data processing step, I supported the testing of effect significance of the experimental inhibitors kb-NB and Gö-6076 by using a parametrized dynamic model to describe two hypotheses and applying an F-test as described in Appendix B.1.3 and shown in Figure 3.2. This modeling approach allowed to include the dynamic behavior in the hypothesis test. To that end, we calculated the test statistic from the residual sum of squares after fitting the model for each hypothesis to the experimental data.

Our choice of MLE for model calibration resulted in the need to choose an appropriate error model for the data and correspondingly the model output. In order to identify the error model that could optimally describe the experimental data, I selected and compared a number of different options as detailed in Appendix B.1.4. These options consisted of the three levels: distribution, variance estimator and pooling of standard deviations. The 12 correspondingly defined model variants were compared using several information criteria. The error model was selected according to the criteria corrected for small sample sizes. As shown, this approach allows for an in depth comparison of different error model options of levels as specific as the modeler deems necessary. In comparison to applying this step as part of the model calibration step, application as a pre-processing step

reduces the required computational time as the calibration tends to be repeated until the optimal mathematical model and corresponding suitable parameter bounds are identified.

The number of experimental data points was relatively small compared to the number of experimental conditions. Additionally, experimental data only comprised measurements for phosphorylated PKD and DLC1. In this sparse data setting, we chose a simplified two state variable model to describe the system dynamics. This especially means that Rho, which is a main contributor to the DLC1-dependent feedback regulating PKD activity in our model hypothesis (Figure 3.1E), is not directly represented in the mathematical model. The model utilizes mass action kinetics as the fundamental modeling approach in order to consider the unknown system behavior and keep it as simple as possible. However, the PKD phosphorylation rate $k(\text{DLC1}, \theta)$ required particular attention. First, our feedback hypothesis is incorporated in the PKD phosphorylation rate. In the hypothesized model structure, DLC1 inhibits Rho activation which results in an inhibition of PKD phosphorylation. Thus, we modeled the influence from DLC1 on PKD by decreasing the basal rate of PKD phosphorylation dependent on the amount of unphosphorylated DLC1. Moreover, all experimental treatments are modeled as direct influences on the PKD phophosphorylation rate. As Rho is not a model variable, we described constitutively active Rho by increasing PKD phosphorylation rate by a fixed value θ_9. All factors incorporated in the PKD phosphorylation rate are specified in Table B.4.

In this study, we investigated the hypothesis of an overall positive feedback involving Rho GTPase signaling on PKD activity. Accordingly, we chose maximum likelihood estimation as method for model calibration. After comparing different approaches regarding the reliability, I employed the Pattern Search algorithm to solve the optimization problem. The use of a multi-start approach enabled me to find a range of optimal and near-optimal parameter sets. Thus, the model hypotheses could be evaluated while keeping the computational effort relatively low.

The calibrated models of the first and second hypothesis both

showed a good accordance with the experimental data. However, as the amount of available experimental data was small, a separation in training and test data was not feasible. Also, as Western Blot data requires normalization to a reference point, the number of data points was even more restricted. In order to nevertheless test the model validity in an objective manner, I implemented a novel bootstrapping approach that sampled new data from the calibrated model (Figure 3.2D). If the calibrated model followed the data too closely, thus overfitting, the calibrated model would fit the majority of the sampled data worse and thus lead to worse objective function values than the experimental data set. Conversely, if the calibrated model fits the data too loosely, thus underfitting, we assume that the calibrated model fits the sampled datasets better than the experimental data. The resampled data was used to estimate the distribution of the maximum-likelihood objective function value. Plotting this estimated distribution against the objective function value of the calibrated model can give an easy visual indication on the plausibility of the model when the number of experimental data points is not sufficient for other methods.

Analyzing the model, I also considered the values and correlations of the inferred parameter sets via scatterplots (Figure B.3). This approach allows the assessment of possible 2D correlations among parameters and how well parameters could be defined given the model complexity and available experimental data. Inspection of parameter values also revealed that the parameter representing the influence of DLC1 on PKD is constantly optimized to very small values. Thus, in order to investigate if the small values of this parameter actually implicate the influence of DLC1 on PKD activity, I simulated PKD activity in response to varying DLC1 total amounts by adjusting the scaling parameter s_1 (Figure 3.3A). As the simulation results suggested a very small influence of DLC1 on PKD activity and thus a weak feedback connection, I performed a profile likelihood analysis on the feedback parameter θ_6 (Figure 3.3B). The results showed that the experimental data could be better described using negative values for the feedback parameter. Using the optimal value of θ_6 found via the profile likelihood analysis, I repeated the simulation of

PKD activity in response to varying total amounts of DLC1. As can be seen in Figure 3.3C, the change in θ_6 increased the influence of DLC1 amounts on pPKD. Taking these findings into account finally lead to the second model hypothesis comprising a negative feedback mechanism from DLC1 on PKD.

This new model hypothesis was tested via additional experiments. These comprised experimental conditions of increased and decreased DLC1 expression as well as inhibition of ROCK. The experiments support the hypothesis that DLC1 increases PKD phosphorylation. Furthermore, the experiments indicate an important role of Rho-ROCK signaling for the feedback mechanism. Hereby, Rho activates ROCK which in turn inhibits PKD activation. Combining new and old experiments, we returned to the data pre-processing step and repeated all steps of the modeling procedure. For the second model hypothesis I followed the same five-step procedure as for the first model hypothesis. Relevant changes only appear in the modeling and model calibration steps.

The complexity of the second model increased compared to the first model as a result of the additional experimental treatments. The first set of experiments showed an increase in PKD activity for constitutively active Rho, we thus kept the corresponding term from model 1. However, the new experiments revealed that Rho-ROCK signaling negatively regulates PKD activity. This negative regulation is suppressed by DLC1 down-regulating Rho activity. Consequently, we assumed that constitutively active Rho inhibits PKD activation via ROCK signaling. To include this aspect in our model, we introduced a parameter θ_{10} to reduce the strength of the positive feedback from DLC1 on PKD for this experimental setup. All adjusted or new factors incorporated in the PKD phosphorylation rate are specified in Table **??**. For model calibration, I employed a gradient-based optimizer, which is faster than Pattern Search, to compensate for the increased model complexity.

Finally, the second model was used for a simulation-based analysis of the feedback by investigating the interaction strengths between the model variables PKD and DLC1. Varying the total amounts of PKD and DLC1, I simulated pDLC1 and pPKD fold changes, respectively

(top row in Figure B.10). Hereby, pDLC1 increases linearly with the PKD amount and similarly pPKD is a linear function of the total DLC1 amount, suggesting a strong connection of the activation of one component with the amount of the other. In a second part, I considered the influence of changes in total PKD on pPKD as well as from total DLC1 on pDLC1 (bottom row in Figure B.10). Results show, that changes in total DLC1 amounts strongly influence DLC1 phosphorylation, while changes in total PKD amounts are not propagated through the feedback.

Additional experiments were performed to test whether the PKD pool, located at the Golgi membranes, is sensitive to DLC1 regulation. The experimental results showed a reduction of Golgi-localized PKD in cells depleted of DLC1. Therefore, we deduced that Golgi secretory function should be affected by a lack of DLC1. This hypothesis was again confirmed experimentally.

In summary, we combined experimental with mathematical modeling approaches to study the molecular interactions between DLC1, PKD, and Rho. Our results reveal that DLC1 controls the basal PKD activity at the Golgi and Golgi secretory function. This mechanism involves Rho-ROCK signaling, which is down-regulated by DLC1.

From a methodological point of view, the number of available experimental data was limited by constraints such as the time and personal needed to perform additional experiments as well as the cost.

Thus, for the selection of all methods, starting from the modeling up to the data analysis step, I had to take the sparsity of the data into consideration. The sparsity forces strong restraints on the applicability of existing methods. Consequently, I developed new methods and introduced novel applications of existing methods such as the pre-processing for error model selection and the use of the F-test for inhibitor testing. Whenever possible, standard procedures, such as MLE for model calibration, were used.

Chapter 4.

ROSIE: RObust Sparse Ensemble for outlIEr detection and gene selection in cancer omics data

4.1. Published manuscript and contributions

This chapter corresponds to the following contribution:

> *A. Jensch, M. B. Lopes, S. Vinga, and N. Radde. "ROSIE: RObust Sparse ensemble for outlIEr detection and gene selection in cancer omics data". In:* Stat Methods Med Res *31.5 (2022), pp. 947–958*

4.1.1. Abstract

The extraction of novel information from omics data is a challenging task, in particular, since the number of features (e.g. genes) often far exceeds the number of samples. In such a setting, conventional parameter estimation approaches lead to ill-posed optimization problems, and regularization may be required. In addition, outliers can have a large impact on classification accuracy.

Here we introduce ROSIE, a sparse and robust ensemble classification approach, which combines sparse and robust classification methods for outlier detection and feature selection and further performs a validity check by using a bootstrap approach. Therefore, we selected three sparse and robust methods: Sparse robust discriminant analysis with sparse partial robust M regression, Robust and sparse K-means clustering and robust, and sparse logistic regression

with elastic net penalty. Outliers of ROSIE are determined by the rank product test using outlier rankings of all three methods, and important features are selected as features commonly selected by all methods.

We apply our methodology to RNA-Seq data from The Cancer Genome Atlas (TCGA) to classify observations into Triple-Negative Breast Cancer (TNBC) and non-TNBC tissue samples. The pre-processed dataset consists of $16,600$ genes and more than $1,000$ samples. We demonstrate that ROSIE selects important features and outliers in a robust way. Identified outliers are concordant with the distribution of the commonly selected genes by the three methods, and results are in line with other independent studies. Furthermore, we discuss the association of some of the selected genes with the TNBC subtype in other investigations. In summary, ROSIE constitutes a robust and sparse procedure to identify outliers and important genes through binary classification. Our approach is ad hoc applicable to other datasets, fulfilling the overall goal of simultaneously identifying outliers and candidate disease biomarkers to the targeted in therapy research and personalized medicine frameworks.

4.1.2. Introduction

Genomics, proteomics, metabolomics, transcriptomics - omics data exist in a wide variety and enable research in just as many medical fields. For example, omics data have been applied in the fields of toxicology (e.g., Thomas et al. [227], Sutherland et al. [224]), nutritional science (e.g., Zhang et al. [250], Kato et al. [120]) and disease research (e.g., Kan et al. [118], Reid et al. [190], Anda-Jáuregui and Hernández-Lemus [11], Paczkowska et al. [173]). The extraction of novel information from omics data is challenging. In particular, classification based on transcriptomics data is hampered by a large feature space and a comparably low number of individuals ($n \ll p$), leading to ill-posed optimization problems. The large p, small n setting is one important problem of the curse of dimensionality and requires a special treatment. A variety of sparse methods that reduce the dimensionality of the feature space have been proposed in this con-

text. Examples include data-based statistical methods such as Linear Discriminant Analysis [57, 156], penalized likelihood functions [45], variable selection methods or shrinkage approaches [202], Support Vector Machines [81] and many more. These methods usually require efficient algorithms.

In addition, transcriptomics data frequently contain erroneous or noisy values. Independent of whether these values are caused by measurement errors or inherent outlying behavior, they can influence the classification process of all the remaining patients [68]. Robustness to outliers can be achieved by robust methods which identify outliers (also denoted as influential samples) during the classification process. A novel approach for outlier detection by Lopes et al. [149], for example, applies a consensus approach that combines the inherent residual measures of several classification methods to obtain a consensus ranking of samples in terms of their outlierness. Since feature selection and also outlier detection methods are based on different assumptions, their performance also varies depending on the specific characteristics of the dataset which they are applied to. Likewise, a comparison of different methods in an in silico study also depends to a considerable extend on the model which has been used for data generation, since every method has its strengths and weaknesses and there is not a single best solution. The idea of Ensemble approaches is to combine several methods which return the same kind of output in order to increase accuracy and reduce the number of false positively selected features. It has already been shown that the sparse Ensemble approach of Lopes et al. [149] achieves high accuracy in feature selection compared to other sparse and robust classifiers in settings where the number of outliers is low [223]. However, in datasets with a larger proportion of outliers, these might have an impact on the classification, and thus on the results of outlier detection and feature selection. Therefore, important features can be missed in the selection.

Combining the idea of an Ensemble approach with the need for robustness against outliers, we propose to use an Ensemble of robust sparse methods, which we name RObust Sparse ensemble for outlIEr detection and feature selection (ROSIE). The general workflow of

ROSIE (i) combines sparse and robust classification methods for outlier detection and feature selection and (ii) performs a validity check in terms of altered data.

To build our Ensemble, we selected three sparse and robust methods with freely available implementations in R packages, to perform supervised (classification) and unsupervised (clustering) learning tasks: Sparse robust discriminant analysis with sparse partial robust M regression [98, 207] (SPRM-DA or SPRM), Robust and sparse K-means clustering [131, 132] (RSK-means), and Robust and sparse logistic regression with elastic net penalty [136, 137] (enetLTS). For each method, a ranking of outlierness for all features is obtained and combined to a single consensus ranking by calculating the Rank Product (RP). Outlierness is subsequently assessed using the RP test. Bootstrap samples drawn from the original dataset are used to verify results.

This pipeline is evaluated on simulated data. Results show that the procedure identifies outliers reliably in different settings. Subsequently, ROSIE is applied to a transcriptomic breast cancer dataset to differentiate triple-negative breast cancer (TNBC) from other breast cancer types (non-TNBC). TNBC is an aggressive breast cancer subtype, with a marked heterogeneity and a poor survival, for which the selection of new biomarkers for the development of new targeted therapies is of clinical relevance [174]. ROSIE is indeed able to select features in a robust way. Moreover, several of the selected genes have been associated with TNBC in other experimental and machine learning contexts, which corroborates the biological significance of the genes selected by ROSIE.

4.1.3. Methods

4.1.3.1. Ensemble procedure

The Ensemble procedure is illustrated in Figure 4.1A. It can be divided into two parts. In the first and main part (Figure 4.1A (left)), three classification methods are applied independently from each other to the dataset. Hyperparameters for each method are optimized during this step. Since all methods are sparse and robust,

each of them returns a list of selected features and a measure for the outlier ranking of the samples. Commonly selected features are marked as important. Moreover, using the RP test to achieve a consensus ranking, we finally obtain a list of outliers by evaluation of the corresponding q-values.

The second part (Figure 4.1A (right)) consists of a validity check which verifies the results of the main part with resampled data. For this purpose, several bootstrap sets are taken from the original dataset while preserving the dataset size and proportion of samples labeled with 0 and 1, respectively. The classification methods are applied to these bootstrap samples using the optimal hyperparameters identified in the main part. The resulting lists of outliers and selected features are used to evaluate the results of the main part.

4.1.3.2. Ensemble methods

We selected three inherently different methods for classification in order to obtain independent ranking results. A schematic depiction of the approaches with arbitrary data points of two classes is given in Figure 4.1B. A formal description of each method, the choice of hyperparameters, as well as the ranking of outliers and the selection of features are detailed in Supporting Information Section C.1.

4.1.3.3. Outlier identification

The identification of outliers by combining the results of different classifiers can in the simplest way be achieved by finding the intersection of samples tagged as outliers by each method. But not only do not all methods provide such tags, this procedure also does not have any statistical background. We therefore apply an Ensemble method based on the RP technique [35]. This non-parametric statistical technique is based on the RP from different methods and permits the calculation of significance rankings for all samples. Therefore, as depicted in the Ensemble workflow (Figure 4.1A), we require the outlier rankings for each classification approach. As the classifiers differ in their procedure of classification and outlier detection, rank-

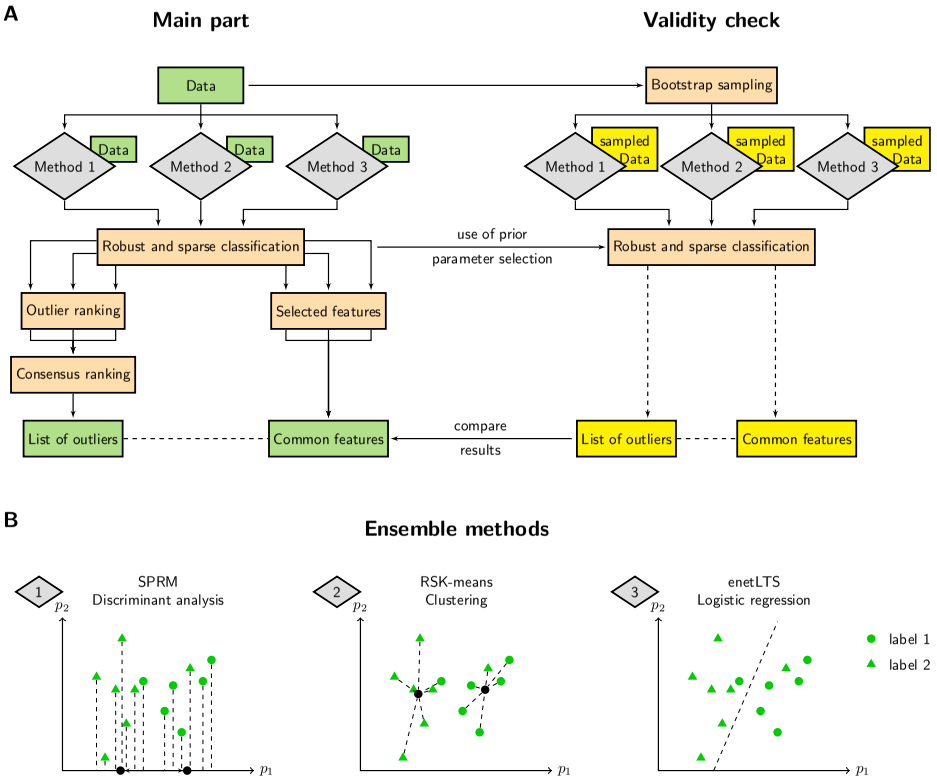

A
Main part
Validity check
Data
Bootstrap sampling
Method 1
Method 2
Method 3
Data
sampled Data
Robust and sparse classification
use of prior parameter selection
Outlier ranking
Selected features
Consensus ranking
List of outliers
Common features
compare results
B
Ensemble methods
1
SPRM
Discriminant analysis
2
RSK-means
Clustering
3
enetLTS
Logistic regression
label 1
label 2
p_1
p_2

Figure 4.1.: **ROSIE workflow and robust and sparse classification methods.** A) Three robust and sparse methods perform classification on the dataset. Each method provides an outlier ranking and selected features. Rankings are combined to acquire an outlier list. Important features are taken as the intersection of all three selected feature sets. Validity of the method is assessed by repeatedly classifying bootstrap sampled datasets and comparing the results with the main part. B) Simplified representation of the underlying classification methods, i.e., sparse robust discriminant analysis with sparse partial robust M regression (SPRM), robust and sparse K-means clustering (RSK-means) and robust and sparse logistic regression with elastic net penalty (enetLTS) for exemplary data comprising two classes and two features (p_1, p_2).

ings are obtained in an individually adjusted fashion, as described for each of the methods. Independent of the ranking rule, an average approach (software settings `ties.method = "average"`) is applied for tied values. Thus, for each sample $i \in \{1, \ldots, n\}$ we obtain three rank values $R_l(i)$, $l \in \{1, 2, 3\}$.

In order to combine these rankings to one consensus ranking, we calculate the RP for each individual as $RP(i) = \prod_{l=1}^{3} R_l(i)$. Subsequently, samples are ranked according to their RP values. Corresponding p-values are then determined using the approach of Heskes et al. [96]. Statistical testing of all p-values increases the risk of type I errors (false positives), since for each test a type I error can occur. In order to control the type I error in multiple testing, the expected proportion of type I errors among all significant test results, i.e., the False Discovery Rate (FDR) [219, 220], can be considered. While a False Positive Rate of 5% implies that on average 5% of true null hypotheses are rejected, an FDR of 5% means that on average 5% of all rejected null hypotheses are actually true. As a measure of the FDR, so-called q-values are calculated based on the p-values. q-values as measures of the FDR are the analogue of the p-values as measures of the False Positive Rate and provide a mechanism to control the rate of false discoveries in multiple testing problems.

4.1.3.4. Validity check

In order to assess the robustness of ROSIE towards variations in the data, we repeat the classification and evaluation steps for different alterations of the original data created by bootstrap sampling (see Figure 4.1A, right side). For m data variations, the samples are separated in m blocks of approximately equal size while keeping the proportion of the classes. Each block is subsequently filled to original size with data points that are sampled with replacement from the complete dataset. Again, we ensure preservation of the case proportion. This sampling strategy ensures that each sample is contained in at least one bootstrap block. In the next step, classification is performed for each block given the parameters that were selected in the main Ensemble run. Finally, the entirety of influential samples

found in the bootstrap runs are compared with the influential samples of the main run. Likewise, we examine the match of selected features found in the main run and the bootstrap runs. In addition to validating our procedure, this approach can be used to reduce the number of features to be evaluated by considering only those that have been repeatedly selected also in the bootstrap runs.

4.1.3.5. Simulation Study

In order to evaluate our ensemble compared to each individual method in a controlled setting in which the ground truth is known, we performed a simulation study on artificial data comprising 3200 features and 200 samples (as detailed in Supporting Information Section C.3). Outliers were created in two different ways. First, a subgroup of samples was randomly selected and their labels switched. This reflects errors in the a priori classification. This was performed for 5% and 15% of the samples, respectively, leading to two datasets. Second, in order to mirror outliers in gene expression in the third dataset, 15% of the features were randomly selected and their standard deviation computed. Then, 5% of the samples were randomly selected and the values corresponding to the selected features increased by three times the respective standard deviation.

4.1.3.6. Triple-negative breast cancer data/ Data preparation

We considered a dataset consisting of RNA sequencing (RNA-Seq) data of breast cancer patients from The Cancer Genome Atlas [191] Breast Invasive Carcinoma data collection. The Cancer Genome Atlas [231] comprises one of the largest collections of omics datasets for more than 33 different cancer types and $20,000$ individual tumor samples. The dataset used to evaluate ROSIE was the one used in Lopes et al. (2018) [149], corresponding to the Breast Invasive Carcinoma RNA-Seq Fragments Per Kilobase per Million (FPKM), excluding the clinical variables subset. The dataset was obtained using the brca.data R package [235], as described by the authors [149].

The dataset consists of $1,019$ patients (samples) in total, of which 160 are TNBC (class membership $y_j = 1$) and 859 non-TNBC ($y_j = 0$). The expression of three receptors was used to assign class labels to the samples. Patients are labeled as TNBC when the genes for the estrogen receptor and progesterone receptor are not expressed while the human epidermal growth factor receptor 2 (HER2) is not overexpressed.

HER2 measurements based on three different readouts were available for a classification of samples, HER2 (via immunohistochemical testing (IHC)) level, HER2 (via IHC) status and the HER2 level measured by fluorescence in-situ hybridization testing (FISH). Altogether, 28 patients showed non-concordance between two of the resulting HER2 labels, of which 4 were assigned to the TNBC and 24 to the non-TNBC group. We refer to these patients as *suspect* samples. For 8 out of these 28 suspect samples, HER2 decision also decides label. We note here that although the non-TNBC group consists of several subgroups, they are all assumed to be similar enough, such that binary classification is not hampered.

The huge amount of raw data was reduced by considering for the analysis only protein coding genes reported by the Ensembl genome browser [247] and the Consensus Coding Sequence project [178]. By additionally removing genes whose expression level remained constant across all patients, a subset of $19,688$ genes (features) was extracted. We further reduced the number of genes to $16,600$ in a final step of data preparation, as SPRM is restricted in the data size it can process. Reduction was performed by employing the function filterVarImp from the R package caret [109], which performs class prediction for a series of feature subsets. For each subset sensitivity, specificity and subsequently the receiver operating characteristic (ROC) curve are computed. The area under the ROC curve is then used as the measure of variable importance. By sorting the features according to their variable importance, we discarded those with the lowest variable importance, such that $16,600$ features remained. Data was log transformed for further analysis.

4.1.4. Results and discussion

4.1.4.1. Simulation Study: ROSIE reliably detects outliers in different settings

In order to investigate the performance of our procedure in detecting outliers, we applied ROSIE to the three simulated datasets. Details about the classification settings and the choice of the hyperparameters are given in Supporting Information C.2 and C.4. Results were compared with those of the individual approaches by ROC analysis (Figure 4.2). For the first dataset (Figure 4.2 left), SPRM performs best, tightly followed by enetLTS. Since RSKC is an unsupervised learning approach, which does not use the a priori labels, its performance is comparable to a random classification for the first and the second dataset, as expected. enetLTS outperforms the other approaches on the second dataset. Since ROSIE takes the outlier rankings of all three methods into account, it naturally cannot be the best method for a single dataset. However, its ROC curve is still acceptable even though RSKC completely fails in these particular scenarios. However, RSKC by far outperforms the other two methods on the third dataset, and enetLTS is not better than random. Also for this scenario, ROSIE still gives reasonable results. Moreover, ROSIE has the best overall AUC value when averaging over all three scenarios.

In summary, this analysis shows that the performance of the individual methods vary significantly and strongly depend on the particular dataset at hand and the kind of outliers, while ROSIE is able to compensate for the failure of one of the methods. Moreover, on average, ROSIE detects outliers more reliably in terms of averaged AUC values. Since for real datasets the outlier percentage and noise levels are usually unknown a priori, ROSIE can indeed provide robust results in a situation of lack of detailed information.

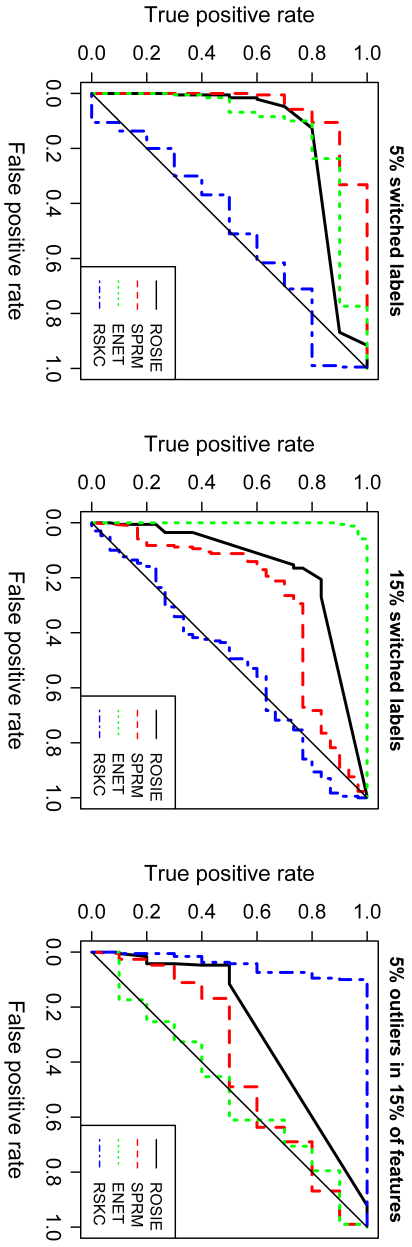

True positive rate
0.0 0.2 0.4 0.6 0.8 1.0
False positive rate
0.0 0.2 0.4 0.6 0.8 1.0
5% switched labels
ROSIE
SPRM
ENET
RSKC
True positive rate
0.0 0.2 0.4 0.6 0.8 1.0
False positive rate
0.0 0.2 0.4 0.6 0.8 1.0
15% switched labels
ROSIE
SPRM
ENET
RSKC
True positive rate
0.0 0.2 0.4 0.6 0.8 1.0
False positive rate
0.0 0.2 0.4 0.6 0.8 1.0
5% outliers in 15% of features
ROSIE
SPRM
ENET
RSKC

Figure 4.2.: **ROC curves for simulation study results.** Results comparing ROSIE with single methods for three outlier settings. Average AUC values: ROSIE (0.81), ENET (0.79), SPRM (0.76), RSKC (0.65)

Table 4.1.: **Summary of classification results.** Number of selected features and number of misclassifications for SPRM, RSK-means and enetLTS.

	SPRM	RSK-means	enetLTS
# of selected genes	2,982	511	70
Misclassifications	68	63	63

4.1.4.2. Breast cancer dataset: ROSIE selects features and outliers in a robust way

We examined the TNBC dataset with the three previously described methods SPRM, RSK-means and enetLTS. Details about the classification settings and parameter selections are given in the Supporting Information Section C.2. Final parameter combinations for each method are listed in Table C.3. Table 4.1 includes the number of selected genes and misclassifications for each of the methods. The methods result in similar numbers of misclassifications. A majority of 56 samples were commonly misclassified by all three methods (see Figure C.1 for a Venn diagram of misclassified samples). The number of selected genes highly differs between the three methods, with differences up to two orders of magnitude. SPRM selects the largest number of genes, 2,982, in the classification process. Interestingly, the 511 genes picked by RSK-means are a subset of this selection. Furthermore, only two of the 70 genes picked by enetLTS are not part of it. Taken together, a set of 54 genes was selected by all three methods (see Figure C.2 for a Venn diagram of selected genes and Table C.4 for a list of gene names). In summary, we have a remarkable agreement between the three methods regarding the set of misclassified samples as well as the set of selected genes.

After aggregating outlier rankings for all three methods and calculating the q-value for each sample, 11 samples with $q < 0.05$ were identified as influential (Table 4.2). All influential samples are of type non-TNBC, while all but one of these samples are classified as TNBC by each of the three classification methods. Also, that one is still

misclassified by two of the methods. This list shows that ROSIE has the potential to detect potential misclassifications also in cases where labels are initially missing. Furthermore, the list of misclassifications is enriched by suspect cases, which is further reassuring.

Five bootstrap samples were used to validate results. The three classification methods were applied to each of these samples, and commonly selected features and a list of influential samples were identified. A summary of the individual bootstrap optimization runs is given in Table C.5 in the supplementary material. All influential samples were repeatedly selected as influential in all bootstrap runs they are part of. Influential samples appear in one up to all five of the bootstrap blocks with a mean appearance of 2.9 times. Moreover, 22 ($\approx 41\%$) of the commonly selected features of the main run are also commonly selected in all five bootstrap runs. Another 14 ($\approx 26\%$) are commonly selected in four bootstrap runs, while only three ($\approx 6\%$) are not commonly selected in any bootstrap run (see Table C.4). Taken together, this analysis shows that ROSIE is able to select features and outliers in a robust way regarding variability in the data.

4.1.4.3. Breast cancer dataset: Influential samples identified by ROSIE match well with the commonly selected genes

In a first analysis step, we considered the correlation coefficients between the commonly selected genes. Figure 4.3 shows the corresponding heatmap of correlation coefficients. The genes show a clear separation into two blocks of predominantly moderate positive correlations while correlations between genes of different blocks are predominately moderate negative. The smaller block consists of 13 genes that are known to be downregulated in TNBC, for example *AGR2*, *TBC1D9* and *TGFB3*. The larger block comprises 41 genes that show upregulated behavior in TNBC samples, for example *FOXC1*, *UGT8* and *HORMAD1* (block membership of all 54 genes is noted in Table C.4). Among commonly selected genes, absolute values of correlation coefficients range from 0.22 to 0.95. In comparison, Figure C.3 presents a corresponding heatmap of 54 randomly selected genes from the full dataset. Here, no such interrelated groups can be

	SPRM	RSK-m	enetLTS	RP	p-values	q-values	miscl. rate
TCGA-E9-A22G	1	94	1	94	0	0.0014	100
TCGA-A2-A0YJ	2	90	8	1440	0	0.0230	100
TCGA-A2-A4S1	61	1	43	2623	$1 \cdot 10^{-4}$	0.0243	67
TCGA-A7-A13E	9	154	2	2772	$1 \cdot 10^{-4}$	0.0243	100
TCGA-A2-A04U *	5	168	4	3360	$1 \cdot 10^{-4}$	0.0243	100
TCGA-LL-A6FR	13	79	5	5135	$2 \cdot 10^{-4}$	0.0296	100
TCGA-AR-A0TP	10	91	6	5460	$2 \cdot 10^{-4}$	0.0296	100
TCGA-AR-A251	3	296	7	6216	$2 \cdot 10^{-4}$	0.0299	100
TCGA-AN-A0FJ *	6	78	22	10296	$4 \cdot 10^{-4}$	0.0410	100
TCGA-OL-A5S0	8	402	3	9648	$4 \cdot 10^{-4}$	0.0410	100
TCGA-AN-A0FL *	39	13	24	12168	$5 \cdot 10^{-4}$	0.0444	100

Table 4.2.: **Summary for influential samples found by Ensemble procedure.** Shown are acquired ranks per method, Rank Product (RP), statistical p- and q-values, misclassification percentage and percentage of significant q-values in bootstrap runs. Suspect cases are marked with an asterisk (*). All influential samples were repeatedly selected as influential in all bootstrap runs they were included in.

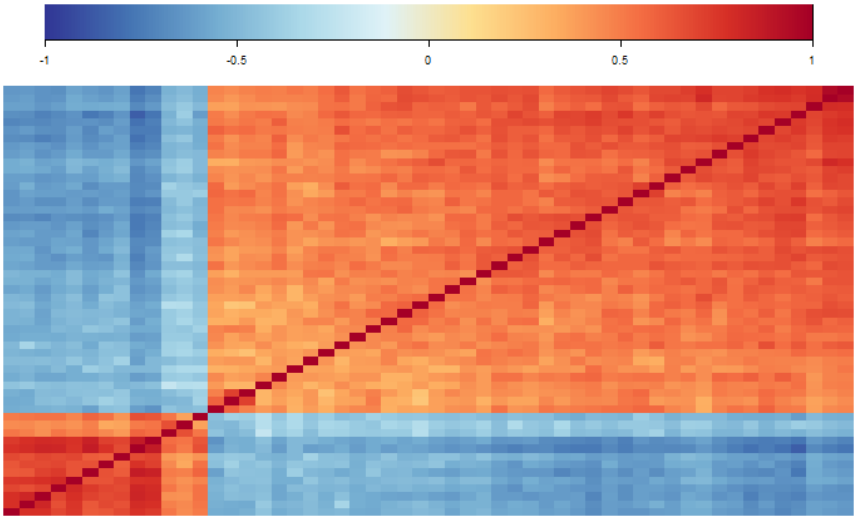


Figure 4.3.: **Correlation analysis of selected features.** Heatmap of correlation values of the 54 commonly selected features.

identified, and weak correlations dominate.

As the correlation values hint to a strong connection among the selected genes, we examined possible distinctive behavior of TNBC, non-TNBC and influential samples via density estimates.

For this purpose, we estimated two 1D marginal densities of commonly selected genes using all but the influential samples that were labeled as TNBC and non-TNBC, respectively, according to the TNBC markers, as described before. Figure 4.4A shows such density estimates exemplarily for six commonly selected genes. Density estimates of the non-TNBC group are represented by the solid red lines, respective estimates of the TNBC group are represented by the green dashed lines. Vertical lines illustrate the medians of both groups. In general, the densities of the non-TNBC group, which comprises around 84% of all samples, are for many genes close to normal distributions, while shapes of densities of the TNBC group vary substantially.

Densities partially show good separation between TNBC and non-TNBC groups, such as *FOXC1*, *AGR2* and *TBC1D9*. Here, the TNBC groups can roughly be summarized as right skewed or left skewed curves, respectively, with a median far from the non-TNBC median. In contrast, *HORMAD1* shows two distinct peaks for the TNBC group, one of which is in good agreement with the non-TNBC peak.

Along with this, values of markers for samples returned as influential by ROSIE are depicted in blue. They all were assigned to the non-TNBC group according to their markers prior to the classification approach. It can be seen that they strongly match the density curves of the TNBC group in all six plots and, related to that, they are distributed closely around the TNBC median. Particularly for *HORMAD1*, the influential sample values tend to have larger gene expression values, and thus fit particularly to the higher TNBC mode. Also, the median for influential and TNBC samples is very similar regarding *HORMAD1*.

The density plots for *UGT8* show another possible behavior of TNBC samples. Here, TNBC samples are rather uniformly distributed over a wide range of values that overlaps with the non-TNBC curve. Influential individuals are also widely spread, but the median still aligns with the TNBC samples.

Finally, *TGFB3* shows two overlapping curves with a seemingly bad separation of TNBC and non-TNBC. Still, the influential samples tend away from the non-TNBC peak and spread around the TNBC median instead.

Since density curves may not properly reflect the fact that around 84% of the samples are non-TNBC and the sample size for TNBC is comparably small, we present histograms of the TNBC and non-TNBC groups in Figure C.4. Overall, Figure 4.4A shows that the influential samples selected via the RP test match well with the commonly selected genes, which in turn supports the potential of our ROSIE approach to identify important features and influential samples.

Based on these results, we asked the question whether the selected genes are primarily those which are differentially expressed between

the two groups. Therefore, we applied edgeR ([193], version 3.26.8) to the dataset in order to identify differentially expressed genes. In total, 7529 genes were found to be differentially expressed by this analysis. There is a very good agreement between the two methods. In particular, all genes found by ROSIE are among the differentially expressed genes identified by edgeR, thus reassuring that these are indeed correlated with the classification. Moreover, all those genes have a quite low false discovery rate, as can be seen by a ROC analysis with the genes found by ROSIE as ground truth (Figure C.5). This analysis shows that ROSIE is able to identify DEGs as important features.

Analysis on influential samples and potential biomarkers on the TCGA dataset was also conducted by Lopes et al. [149] and Segaert et al. [206]. The concordance of the three approaches are illustrated in the Venn diagrams in Figure 4.5. Lopes et al. used an Ensemble approach of sparse classification methods to identify outliers in the TCGA dataset using the RP statistics. 24 influential samples were identified, four of which coincide with our findings. The large difference in the number of outliers found by Lopes et al. and ROSIE is probably due to the fact that the three methods that were used in their ensemble approach are much more similar than in our approach.

Conversely, Segaert et al. used a single robust and sparse method, enetLTS, for outlier detection. Their results comprise 43 influential samples which include all of our findings. Both publications also present a set of genes as potential biomarkers for TNBC. Five genes which we identified as potential biomarkers were also found by the sparse Ensemble [149], while 26 are in common with Segaert et al. [206]. Overall, this shows that our results are in line with other independent studies on the same dataset and additionally provide novel genes as potential putative biomarkers.

4.1.4.4. Breast cancer dataset: Genes selected by ROSIE are associated with TNBC types in other studies

As the goal of this study is to show the capability of ROSIE for identifying biomarkers and influential samples in oncology data, we

exemplarily investigate the biological background of three of the 54 selected genes. In the following, we will thus illustrate the significance of our findings by discussing the biological importance of the genes *HORMAD1*, *AGR2* and *TBC1D9* for TNBC, which presented especially strong indications of importance in literature.
HORMA domain containing 1 (*HORMAD1*) is one of the genes repeatedly selected also in the bootstrap runs of the Ensemble procedure. As HORMA domains play a role in chromatin binding, the protein encoded by *HORMAD1* has been suggested to be involved in meiosis and its expression as a potential marker for cancer [41]. In previous studies analyzing differentially expressed genes between TNBC and non-TNBC, *HORMAD1* has already been highlighted as one of the key upregulated genes differentiating TNBC and non-TNBC [39, 249]. Additionally, *HORMAD1* overexpression, referring to the higher *HORMAD1* levels of the second mode of TNBC samples, has also been reported to contribute to Homologous Recombination Deficiency and to be a potential composite predictive biomarker for sensitivity to platinum-based chemotherapy in TNBC patients [239].

Similarly, *AGR2*, which has also been repeatedly selected in the Ensemble procedure, was listed among the top downregulated genes differentially expressed between TNBC and non-TNBC [39, 249]. It has been shown that *AGR2* is coexpressed with the estrogen receptor in breast cancer cell lines [229]. In addition, it has been associated with cell migration and metastasis [58].

Finally, *TBC1D9* is a gene whose function has only recently been revealed to be involved in the regulation of selective autophagy via regulating *TBK1* activation, which in turn is often associated with cancer [169]. Another recent study employed machine learning algorithms and survival outcome of breast cancer patients to identify three potential genes for the discrimination between TNBC and non-TNBC [134]. Thereby, *TBC1D9* was selected, and overexpression of *TBC1D9* was furthermore shown to be connected to a better prognosis [134].

These aspects reinforce our findings of genes important for TNBC classification and the importance of identifying outlying individuals whose unique gene markup might influence their prognosis and drug

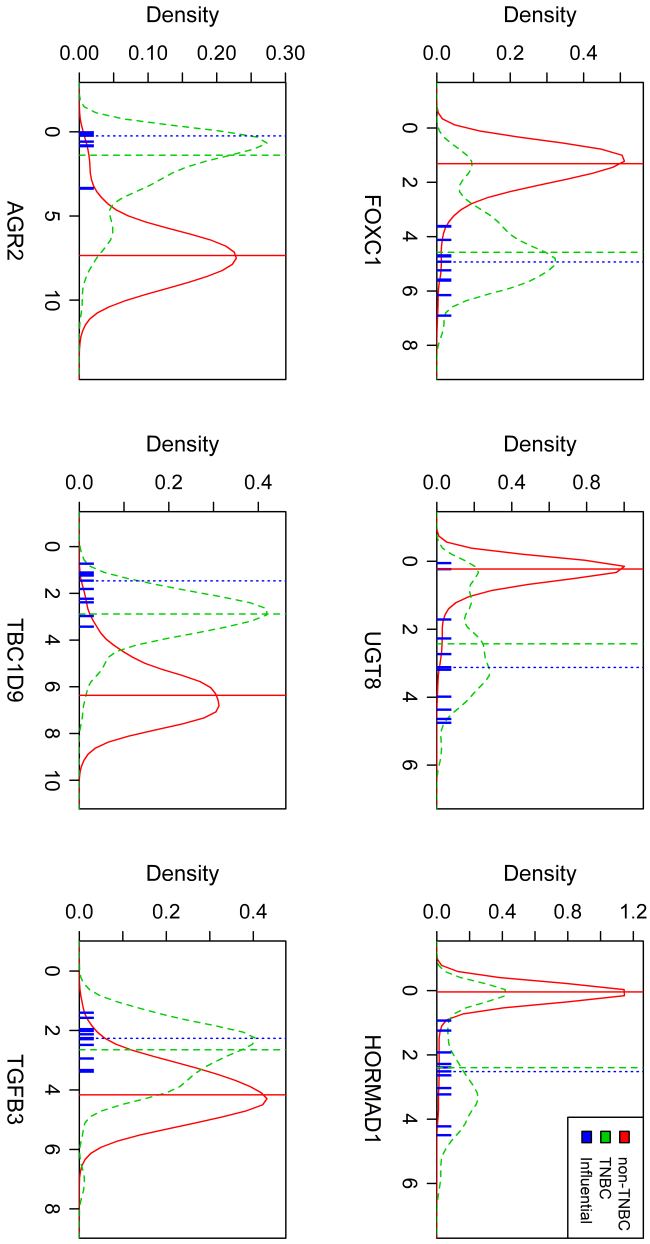

Density
FOXC1
AGR2
UGT8
TBC1D9
HORMAD1
TGFB3
non-TNBC
TNBC
Influential

Figure 4.4.: **Relation between influential samples and commonly selected genes.** Estimated densities of gene expression of selected features grouped by TNBC (green dashed line) and non-TNBC (red line). Vertical lines represent respective group medians. Blue markers depict influential samples.

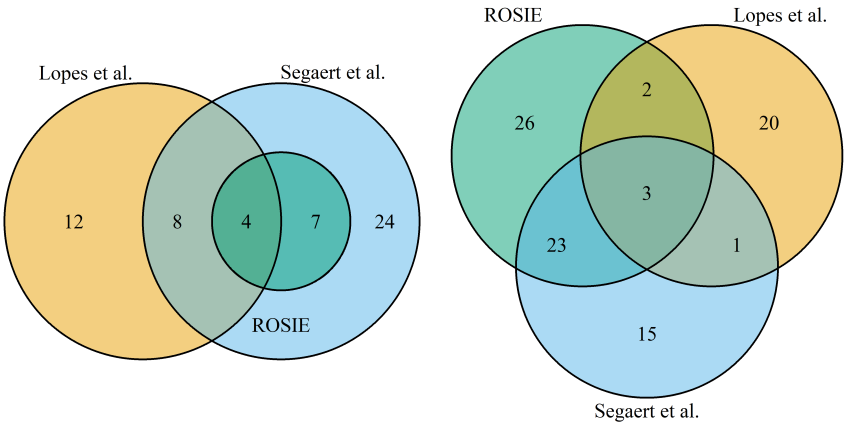


Figure 4.5.: **Venn diagrams comparing different classification approaches.** Comparison of identified outliers (left) and selected genes (right) from ROSIE, the sparse Ensemble approach by Lopes et al. [149] and the robust approach enetLTS by Segaert et al. [206].

sensitivity.

4.1.5. Conclusions

In this study, we have presented ROSIE, a robust and sparse Ensemble approach for outlier detection and feature selection from high dimensional datasets. ROSIE combines different robust and sparse methods which are individually applied to the dataset. Thereby, hyperparameters are adjusted individually for each method, and a ranking of outliers as well as a set of selected features are defined. ROSIE combines these results into a consensus ranking by evaluating the q-values via the RP test, and by defining the set of selected features as features commonly selected by all methods. A validity check is done via a bootstrap approach. ROSIE was validated on simulated datasets and subsequently applied to RNA-Seq data from the TCGA for classification into TNBC and non-TNBC tissue samples.

Applying our Ensemble approach, we managed to reduce a set of 16,600 genes to 54 possible biomarkers for TNBC. ROSIE was able to identify features and outliers in a robust way. Furthermore, the identified set of potential biomarkers seems promising, since several of those genes also appear in other studies on differently expressed genes between TNBC and non-TNBC.

A survival analysis that compared TNBC cases, non-TNBC cases and outliers shows that outliers are all censored at early time points (Figure C.6). In our opinion, this does not allow for any conclusions regarding similarity or dissimilarity between non-TNBC and outliers. If outliers were similar to the other class (here TNBC) in this analysis, one could argue that this probably hints to just a wrong labeling of those samples. However, this is not the case and needs further investigation in the future.

The workflow which we have presented can also be applied to other datasets with a large feature space and a low number of samples. In particular, it can handle outliers in the dataset. Overall, compared to the application of a single robust and sparse method, Ensemble approaches that combine inherently different methods might be superior in distinguishing spurious from true findings.

In future work, it remains to be seen how much the results of our Ensemble approach depend on the individual methods which are combined. In our application study, for example, we have observed a large similarity between SPRM and enetLTS, which overshadows the ranking results of RSK-means. As K-means can be seen as a special case of tclust [68], a more flexible robust clustering approach which is particularly designed to fit clusters with different scatters and weights, it would for instance be interesting to replace trimmed k-means by this more general approach or even more advanced versions [67, 69] in future applications.

Furthermore, ROSIE suffers from long run times, especially for RSK-means, which has the longest run time despite the smallest number of parameter combinations in the parameter selection step. This needs to be addressed to make ROSIE applicable to larger datasets in future work.

4.1.6. Contribution

This project focused on outlier detection and feature selection in the setting of classification. I implemented the ensemble approach including the validity check and further validation approaches. Additionally, I performed the analyses of the identified outliers and selected genes. This included the suggestion, implementation and visualization of techniques. I was actively involved in the discussion of the results and implications during all stages of the project. Finally, I significantly contributed to writing the manuscript.

4.2. Results in the overall context

This project considered the problem of extracting outlying (or influential) sample points and features of interest when performing classification. We approached these tasks with a novel ensemble scheme that combines robust and sparse classification methods. The robust and sparse ensemble scheme that we named ROSIE was applied to a breast cancer data set to differentiate triple negative from non triple negative breast cancer.

The RNA Seq data of breast cancer patients were extracted from the online database The Cancer Genome Atlas. We pre-processed the data by creating class labels according to the expression level of three receptors. Additionally, dimension reduction was performed in order to decrease the amount of raw data. As a first step, data for proteins not reported by the Ensembl genome browser or the Consensus Coding Sequence project was excluded. In the next step, genes with constant entries across all samples were omitted as well. Finally, because one of the classification methods imposed an upper bound on the number of features, we employed another method to further reduce the data set. In order to ensure that genes important for the distinction between the two classes remain in the reduced data set, a measure of variable importance with respect to the class separation was calculated. This filter method allowed us to define the number of genes to keep while prudently reducing the feature space. In preparation for the subsequent model analysis, we additionally

extracted survival data which includes the survival time and censoring status of all patients.

A key aspect of this project was the identification of outliers and important features which may serve as targets for biomarker research. Previous research by [223] showed the strength of a sparse classifier ensemble for outlier detection and feature selection. As higher numbers of outliers were shown to be detrimental for feature selection in the same context, we decided to use an ensemble of sparse and robust classification methods. The basic classification models were selected to have different working concepts. This approach is based on the idea that through the variety of classifiers, the ensemble would be able to provide reliable results for any data set, while similar methods would have similar strengths and weaknesses. For the calibration step, method implementations that included robustness and sparsity were chosen.

After learning the classification models, the results were used to identify outliers and important features. The outlier detection was based on the RP technique in order to provide a statistical background. Hereby, for each method an outlier ranking of the samples was determined. The acquired ranks across the three classifiers were multiplied for each sample and ranked according to this RP. This ranking was then evaluated in terms of q-values to allow a control of the false discovery rate. Important features were selected as intersection of the feature sets provided by the three classification models.

The validation of the results of our ensemble approach was hindered by the fact that for the TNBC data set the ground truth is known neither for outliers nor for important genes. As a consequence, standard validation approaches as introduced in Chapter 1 are not feasible as the classifier performance could not be quantified. Therefore we considered different approaches to investigate the reliability of ROSIE.

Our ensemble scheme includes a validity check employing a bootstrap sampling approach. Hereby, outliers and important genes, determined by the consensus approach, are compared for varied data sets sampled from the original breast cancer data set. I implemented

a type of block bootstrap sampling that guaranteed each sample to be present in at least one validation data set. Special attention was also given to ensure that the ratio of TNBC and non-TNBC samples remained the same as in the original data set. Computation time was significantly reduced by reusing the parameter combinations selected for the original dataset. Evaluating the results showed that the identified influential samples were repeatedly selected as influential in all bootstrap runs they were contained in. Furthermore, we examined the selected genes by comparing them with the results of edgeR, a method for identifying differentially expressed genes. The set of genes identified by edgeR as differentially expressed contains all genes selected by ROSIE. Visualizing the comparison of genes found by edgeR and our selection, by computing a ROC taking our selected genes as ground truth, revealed that the the important genes selected by ROSIE have relatively low false discovery rates.

In order to examine the abilities of ROSIE in comparison to the single classifiers with respect to outlier detection in diverse data settings, we created three artificial data sets with differently conditioned outliers and analyzed the results using ROC curves. The ROC curves show that none of the three classification methods performs best on all data settings or they even completely fail in outlier detection. On the other hand, while ROSIE cannot provide better results than the best method for any data setting, it also remains reliable even when single methods fail. This impression was confirmed by evaluating the average AUC values of the three single methods and ROSIE across the three outlier settings.

As a step that partially validates our findings while also analyzing the results, we scanned the existing literature. We compared our findings with other machine learning approaches that examined the same data set. Additionally, we investigated the biological background of a subset of our selected genes, displaying that our approach is able to identify potentially interesting genes for further research.

For further analysis of the commonly selected genes, we considered the correlations among those genes. The visualization of the correlation values with a heatmap provided further inside into structural relations among the genes (Figure 4.3). Identified outliers were eval-

uated under a clinical context, by employing Kaplan-Meier curves (Figure C.6). Using the survival data of the patients in our data set, we looked at survival probabilities for TNBC, non-TNBC and outlying samples. Outliers and commonly selected features were also analyzed in a combined fashion. For a subset of the commonly selected genes we created density plots of the expression values of the TNBC and non-TNBC groups, while our outliers were marked separately (Figure 4.4). This way, I evaluated whether outliers match with their predefined class or exhibit different behavior. In addition, the separation of expression values between the two classes could also be considered. As there are many more non-TNBC than TNBC samples, the normalized density plots may be misleading with respect to actual sample sizes. To clarify this, I also provided histograms corresponding to the density plots (Figure C.4).

In summary, we proposed a novel approach for outlier detection and gene selection based on an ensemble classification scheme. Data dimensions were reduced in a pre-processing step to increase usability of the results and reduce computational runtime. ROSIE combines the results of three classifiers to determine outliers and important features. The power of ROSIE was evaluated by creating artificial datasets with different outlier settings to compare the accuracy of ROSIE in comparison to each single classifier. In addition, the validity of the findings with respect to variations in the data composition was tested via a bootstrap approach. Several new datasets are hereby sampled from the original data while ensuring that each sample is present in at least one such dataset. We also compared our set of important features with the results of edgeR, a tool for differential expression analysis. Further comparisons with previous studies were conducted with respect to our set of outliers and selected genes. Outliers and important genes were used for further analysis.

Chapter 5.

Conclusion

This chapter concludes the thesis by summarizing and discussing the findings of the preceding chapters. In addition, an outlook on a few potential future research directions is presented.

5.1. Summary

In the previous chapters I presented three modeling studies. While all studies aimed at gaining deeper understanding of biological systems and followed the five modeling steps described in Chapter 1, each study highlights different purposes, challenges and methods. Here, I will shortly summarize the content of each publication.

Chapter 2 features a well studied signaling pathway, the MAPK signaling pathway. The presented work focused on the investigation of sustained ERK activation when the system is stimulated with NGF. Methods for all modeling steps were chosen to allow extensive exploration of this mechanism. In particular, we employed MCMC-sampling to generate samples from the posterior distribution of the parameters. The model was validated by predicting the outcomes of a set of perturbation experiments which were not used for model calibration. By combining CBA, an analytical approach to steady state analysis, with simulation-based analyses we studied the sustained ERK activity in our model. We investigated different aspects of sustained activation of ERK, with particular focus on the mechanism behind ERK response. Results revealed that quasi-bistability can contribute to the observed ERK response.

The second project (Chapter 3) features the tumor suppressor protein DLC1 and its role in maintaining PKD activity at the Golgi membranes. We defined a mathematical model for the first model hypothesis describing an overall positive feedback that maintains PKD activity and involves DLC1-dependent Rho signaling. This model hypothesis was subsequently questioned as model analysis suggested an overall negative rather than a positive feedback. As a result, additional experiments were conducted, which strengthened the hypothesis of negative feedback. Thus, we defined a second mathematical model including these new experiments. The analysis of the model fit and parameter sets of the calibrated second model supported the second model hypothesis. Overall, our results reveal that DLC1 contributes to the activation of PKD at the Golgi and Golgi secretory activity by downregulating Rho-ROCK signaling.

The third project (Chapter 4) introduces ROSIE, a new ensemble classification approach, which combines three sparse and robust classifiers for outlier detection and feature selection. This proposed method further performs a bootstrap-based validity check to assess the robustness of ROSIE towards variations in the data. In addition, we conducted a simulation study to evaluate ROSIE in comparison to the three individual methods. ROSIE was applied to RNA-Seq data of TNBC and non-TNBC samples to identify patients with a conspicuous gene expression profile compared to their class and to extract potential biomarkers. We validated our approach using four different methods, which consist of the validity check, the simulation study, a comparison of results with other studies and methods, and the investigation of the biological relevance of our findings.

5.2. Discussion

Each study of biological systems requires individual consideration to choose appropriate methods for all modeling steps. The presented studies demonstrate the selection process for three different study purposes. Consequently, various methods for data pre-processing, modeling, model calibration, model validation, and analysis were

introduced and applied. This includes many established methods that were adapted to the specific requirements. In the following paragraphs, I will highlight some of the methodological choices and discuss some novel approaches.

Data pre-processing

Starting with the data pre-processing step, the second study presented in Chapter 3 included new experimental data. As experimental application of inhibitors requires the evaluation of the effect, performing significance tests is a standard procedure. However, standard approaches do not take time courses into account. In Chapter 3 significance of the inhibitors was therefore instead tested by representing the null and alternative hypotheses with time dependent curves that were fitted to the data.

As exclaimed in the Chapter 1, in classification problems dimension reduction represents a common data preparation task. In Chapter 4 different approaches to tackle this problem were showcased. Selecting only genes that are listed in specific databases was a useful tool to reduce the magnitude of the dataset. In different scenarios using other data sources, this approach may however not be feasible. On the other hand, deleting uninformative variables with constant entries across all samples is always applicable. Nevertheless, this may not substantially reduce the data dimension. The final approach applied in Chapter 4 is a filter method that performs independently from the model choice. This allows to reduce the data dimension to an arbitrarily large subspace.

System modeling

Proceeding from the data pre-processing to the modeling step, the studies presented in the previous chapters used standard approaches. In Chapter 2 and 3, mass action kinetics constituted the basic approach for modeling the systems. Nevertheless, as also shown in those studies, careful consideration of biological knowledge can be used to either simplify the model or enhance model precision for different

aspects. This is important in order to keep the number of parameters as low as possible while the model granularity still suffices to represent the system components under study, especially in the case of sparse datasets. The modeling step in Chapter 3 highlights the importance of meticulously evaluating the relations among system components. Here, only PKD and DLC1 are included as model variables, while Rho and ROCK are indirectly represented in the PKD phosphorylation rate. Consequently, the influence of experiments altering Rho-ROCK signaling had to be interpreted and modeled in terms of the overall effect of DLC1 on PKD activity.

On the contrary, for classification tasks it is possible to use one of the magnitude of existing classification models. In Chapter 4 the model selection was influenced by the planned analysis. As the project applied an ensemble approach to identify outliers and potential biomarkers, three classifiers had to be selected. The methods were selected to use different classification mechanisms in order to allow the proposed procedure to work reliably on different datasets. However, other methods may also contribute to improving the ensemble scheme. Adding additional classifiers, for example an SVM, could potentially improve accuracy with respect to outlier identification and feature selection. On the other hand, additional classifiers would also add to the computation time, which may be undesired.

Model calibration

For the model calibration step, all three studies relied on existing methods. Still, the choice of a method is strongly influenced by the study purpose. While the study presented in Chapter 2 aimed at investigating the feedback mechanism and employed several simulation studies in the process, the project described in Chapter 3 aimed at understanding the system structure and evaluated different model hypotheses. Consequently, the first study employed MCMC sampling which allowed excessive simulation studies using the sampled parameter sets. While MCMC sampling enables the investigation of the posterior predictive distribution of model predictions, the calibration process is computationally expensive. Therefor, applying a

point estimator such as the ML estimator may be a superior choice for hypothesis testing as conducted in Chapter 3. For classification problems, different calibration methods exist as well and also need to be selected with the study purpose in mind. The project in Chapter 4 thus applied classifiers that had been enhanced to be robust and sparse. In addition, further steps were included to identify outliers and important features. For outlier detection, samples were ranked in terms of outlierness for each classifier. These rankings were combined via the rank product and evaluated according to q-values to control the rate of false discoveries. Important features are chosen as the intersection of the feature sets selected by the three classification methods, thus increasing reliability of the results while reducing the number of features that could be targeted for further investigation.

Model validation

Model validation is a key step to ensure that any conclusions drawn from model analysis can be relied upon. Although standard techniques exist, as they rely on the presence of sufficient data amounts, they may not be applicable depending on the data availability and composition. In Chapter 2 validation was thus performed using data from additional experiments. However, as apparent in this project, careful consideration towards the implementation of new conditions in the model is necessary. As the project presented in Chapter 4 used a classifier ensemble to identify outliers and potential biomarkers, the validation step was aimed at these entities instead of the classification results. The ground truth for outliers and biomarkers in the dataset is however unknown and different approaches were combined to investigate the validity of ROSIE. This included the comparison of results with other studies and methods. Moreover, we investigated the biological background of the genes selected by ROSIE in order to highlight the relevance of our findings for TNBC research. Additionally, in order to illustrate the abilities of ROSIE in comparison to each single classifier, a simulation study was performed. The simulated data was corrupted with outliers, which were subsequently assessed by ROSIE. Comparing the outlier findings of ROSIE and the three

classifiers via ROC curves for different data settings demonstrated the versatility of ROSIE in comparison to single methods.

Model analysis

The choice of analysis tools also strongly influences other process steps as seen above. At the same time, the step of model analysis is especially dependent on the study purpose. Nevertheless, certain considerations are common in this step. For systems biology studies, this refers to an analysis of the parameter space. Scatter plots can be used to visually inspect 2D correlations and the variance of the parameters given multiple parameter sets. Correlation analysis in general can be used for model reduction, as for example in [55]. Profile likelihoods are another useful approach to analyze the parameterized model. Even though the main focus of the profile likelihood analysis is the investigation of parameter identifiability, Chapter 3 demonstrate its application to simultaneously invalidate a model hypothesis and suggesting a new one.

In classification studies the parameters are not of inherent interest as they do not represent biological functions. Instead, correlations can be inferred for the feature space. Using heatmaps to visualize correlation coefficients as in Chapter 4 permits the identification of correlated clusters of features. Focusing alternatively on the influential samples, Kaplan-Meier curves can provide additional insight into the relevance of the findings as information on the survival time of each sample is incorporated. However, especially for small a number of outliers, Kaplan-Meier curves may be hard to interpret when the small outlier population does not resemble either class. Finally, visualizing the relation between the identified influential samples and the potential biomarkers can support findings of both categories. In Chapter 4, this was implemented using simple density estimates of the feature values of the two groups. In addition, the values for the influential samples were marked separately, which allows the comparison of outliers with their predefined class.

Returning once again to systems biology, modeling studies introduce the great possibility to predict system behavior under varying

conditions without relying on time-consuming or even infeasible experiments. By varying scaling factors to simulate changes in protein amounts, system behavior can be predicted for different experimental conditions. Additionally, the granularity of changes to be investigated can be almost chosen arbitrarily as model simulation is relatively fast in general. The studies in both Chapter 2 and 3 employ simulation studies of this kind with 23 and 5 different settings, respectively. Alternatively, simulation studies can also be used to predict system behavior for a prolonged time beyond the experimental data. In Chapter 2, this approach was exploited to assess whether the model could in general reflect long term ERK activity. In the broader context of system analysis, as compared to model analysis, further experimental work can provide valuable additional insight into the system. In Chapter 3, the relation of DLC1 to Golgi-localized PKD and Golgi secretory function was experimentally investigated. Combining the results of the modeling study and these additional experiments revealed that DLC1 positively regulates PKD activity at the Golgi and Golgi secretory activity.

In addition to adapting the aforementioned methods, some novel approaches have been introduced as well. The study in chapter 3 employs MLE for model calibration, which requires the selection of an error model. In order to compare different error model hypotheses, a computationally efficient approach was proposed. Thereby, different error models are compared in a data pre-processing step that does not involve the calibration of the model. Consequently, computation time is not restrictive for this method and it allows the comparison of various error model options. Different information criteria can be used to compare the model variants. Hereby, it is important to consider the number of data points in regard to the number of parameters. This becomes apparent in Table B.2 where the top three error models selected by the information criteria AIC and AIC_c, where the latter corrects for small sample sizes, are completely different. In an advanced setting, this pre-processing step could also be used to first select the top contenders for the error model choice. The model can then be calibrated for this subset of error models to determine the optimal choice. This setting may be useful when the selected

information criteria do not yield a clear favorite error model.

Model validation employing training and test data requires sufficiently large data sets to enable splitting. However, in Chapter 3, only few replicates with low time resolution were available and a division of the data was not feasible. A novel bootstrapping approach was instead introduced to assess potential overfitting or underfitting of the model. Hereby, artificial datasets are sampled from the inferred stochastic model. The likelihood value of observing these datasets, given the calibrated model, is then compared with the likelihood of observing the experimental data, given the calibrated model. If the model fits the experimental data better than most of the sampled data, this may suggest that the model overfits the experimental data. In contrast, if the model fits the experimental data worse than most of the sampled data, this may hint towards underfitting. Nevertheless, as this approach only considers over- and underfitting, it may not identify other irregularities. For example, the inspection of the parameter scatterplots in Chapter 3 revealed that the good model fit was achieved by minimizing the influence of the feedback in the first model hypothesis. This also highlights the importance of evaluating model quality from different perspectives.

Bootstrap sampling constitutes a valuable tool for model validation. It was also applied in Chapter 4 to evaluate the robustness of the outlier detection and gene selection to variations in the data. Here, a block bootstrap approach was chosen to ensure that each sample is represented in the validation process. We furthermore adjusted the sampling procedure such that the ratio of the two classes equaled the original data for each bootstrapped dataset. The classification step with subsequent determination of outliers and important genes was repeated for each dataset. In order to reduce the overall computation time, we decided to omit the selection of hyperparameters and instead reuse the parameter combinations that were chosen for the original dataset. Nevertheless, for projects involving smaller datasets and thus faster computation times, selecting the hyperparameters for each bootstrapped dataset may improve classification accuracy and subsequently conformity of the detected outliers and selected features. Further consideration could also be given to the sampling scheme.

Given the presented procedure, identified outliers were present in one up to all five bootstrap blocks, with a mean appearance of 2.9 times. Ensuring that these samples are present in each bootstrap block may increase the significance of the validity check.

The final methodological approach I wish to highlight is the combination of analytical tools with simulation studies as performed in Chapter 2. The study focused on the analysis of the feedback mechanism supporting sustained ERK activity. We approached this problem from two different perspectives. On the one hand, we employed the CBA to perform a steady state analysis of the system for the MCMC parameter sets and classified the outcome in bistable and monostable systems according to the number of steady states. Using CBA for the steady state analysis utilizes the network topology of the signaling pathway to perform efficient steady state calculations. On the other hand, we simulated the model for the MCMC parameter sets and classified the trajectories into bistable, quasi-bistable, and monostable systems depending on the level of ERK activity at 60 min and 600 min. Simulating model trajectories allows to observe the system behavior at any time point of interest. Combining these approaches showed that network modeling studies may produce effects undetectable through standard analytical methods. In this case, 90% of the parameter sets were monostable according to the CBA, yet all simulated trajectories showed sustained ERK activity of at least 60 minutes. We then again used the CBA to investigate the mechanism behind quasi-bistability.

In conclusion, modeling of biological systems is generally based on five steps that consist of data preparation, system modeling, model calibration, model validation, and model analysis. Each step comprises its own challenges and often a vast choice of available methods. The selection of methods for the different steps is strongly influenced by the research question, but also by decisions in other steps. In this work I presented three studies investigating biological systems and the complex decision processes to fulfill the study purposes. I also presented and discussed the methods required under the different project conditions. In addition, each study highlighted not only individual adaptations of existing methods to adjust to the circumstances

but also novel approaches to different aspects of the modeling process. Circumstances hereby refers to for example the availability of data, which greatly influences especially the data pre-processing and model validation steps. In this thesis, all projects were confronted with sparse data and novel approaches to address this problem were presented for the data pre-processing, model validation, and model analysis steps.

5.3. Outlook

Based on the presented work, several future research directions exist.

The feedback loop model of DLC1 for PKD activity at the Golgi membranes enables excellent research possibilities. In future research, an expansion of the model could be interesting with the goal of describing the mechanisms by which DLC1 is involved in focal adhesion modification. Such model would be useful to study and predict changes of focal adhesions in response to mutations downregulating DLC1.

Besides the expansion of the biological scope on PKD regulation, I presented a new ensemble approach that combines feature selection and outlier detection results from three different methods. The presented ROSIE approach encourages the selection of different classifiers in order to gather information from another point of view. Further research could also aim at decreasing the runtime or a further reduction of the number of selected features.

In this thesis I proposed two new bootstrap-based approaches for model validation. Further research could help improve these methods, for example by enhancing the sampling procedure. The sampling procedure for the validity check of ROSIE ensures that each sample, and consequently each identified outlier, is present in at least one bootstrap block. In future research, it could be considered to include the identified outliers in each block to gain more information on the validity of the results. However, depending on the number of outliers, this may also introduce bias in the sampled dataset.

Appendix

A. Additional files for Chapter 2

A.1. Additional file 1: A Bayesian framework for ODE model calibration

In a Bayesian parameter estimation framework, every quantity-of-interest is described in terms of a probability distribution. This framework allows to propagate variability in the data to uncertainties in model predictions and is illustrated in Figure A.1. The framework is initialized by encoding prior knowledge about parameters θ in a *prior probability distribution* $p(\theta)$, which is most often simply a uniform distribution within finite boundaries. Data are interpreted in this framework as samples from a parametrized stochastic process, which defines the likelihood function $p(y|\theta)$. In our framework, we employ *stochastically embedded* ODE models, i.e. we assume that the underlying process can be described in a deterministic way (the ODE model) and measurements are disrupted by measurement errors (the error model, also called noise model). The likelihood function is used to update our prior knowledge about model parameters and to transform it into a *posterior distribution* $p(\theta|y)$, which is a distribution of the model parameters conditional on the data. This is obtained via exploiting Bayes' Theorem. This posterior distribution can in principle be transformed into *posterior predictive distributions* $p(\tilde{y}|y)$ for any quantity-of-interest $\tilde{y}$, like e.g. marginals of individual parameters, model states, event times, or discrete features emerging from the model's behavior such as quasi-bistability.

In the particular framework of ODE model parameter estimation, we face the problem that the posterior distribution is not available in closed form. Thus, it is investigated via generating representative samples, which is realized via constructing a Markov chain that converges to the desired target distribution. There are numerous

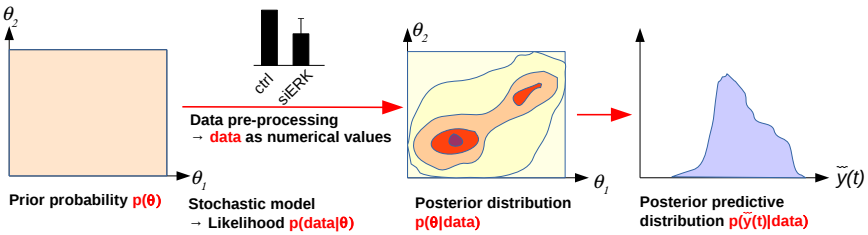


Figure A.1.: Schematic of a Bayesian learning framework.

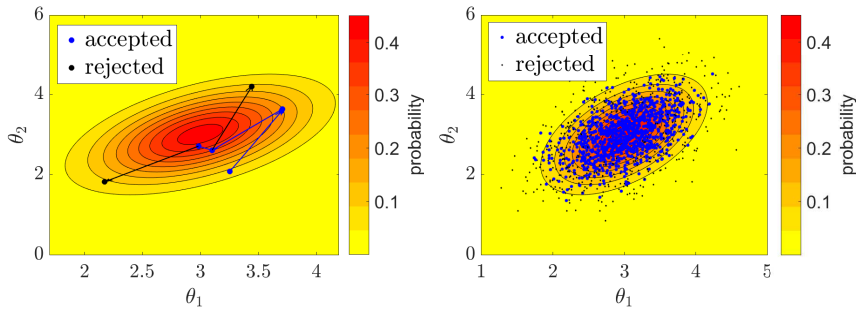


Figure A.2.: Schematic of Markov chain Monte Carlo sampling.

algorithms available for this (for more details and historical work we refer to [72, 90, 159]). The working principle of such an MCMC algorithm is shown in Figure A.2. Situated at θ, the Markov chain proposes a new parameter set θ', which is accepted with a probability that takes the ratio of the values of the target density at θ and θ' into account (Figure A.2 left). If the chain is converged, the set of accepted samples represent the target distribution (Figure A.2 right).

A.2. Additional file 2: Model normalization procedure

We start with the model version shown in Fig 2.3B in the main manuscript

$$
\begin{aligned}
\dot{\text{pRaf}} &= k_1^+(\text{Raf}_{\text{TOT}}s_1 - \text{pRaf})u(t) - k_1^-\text{pRaf} + \\
&\quad + fn\left[-k_{Fn}\text{ppERKpRaf}\right] + \\
&\quad + fp\left[k_{Fp}\frac{\text{ppERK}^5}{\text{ppERK}^5 + g^5}(\text{Raf}_{\text{TOT}}s_1 - \text{pRaf})\right] \\
\dot{\text{ppMEK}} &= k_2^+(\text{MEK}_{\text{TOT}}s_2 - \text{ppMEK})\text{pRaf} - k_2^-\text{ppMEK} \\
\dot{\text{pERK}} &= k_3^+(\text{ERK}_{\text{TOT}}s_3 - \text{pERK} - \text{ppERK})\text{ppMEK} + \\
&\quad + k_4^-\text{ppERK} - (k_3^- + k_4^+\text{ppMEK})\text{pERK} \\
\dot{\text{ppERK}} &= k_4^+\text{pERKppMEK} - k_4^-\text{ppERK} \\
u(t) &= \begin{cases} 0 & t < 0 \\ 1 - \dfrac{t^3}{t^3 + K^3} & t \geq 0. \end{cases}
\end{aligned}
$$

In order to compare this model to the data in [199], variables have to be rescaled and normalized to the same reference experiment as in [199]. The light signals detected in the Western blots were normalized to the signals of the respective total proteins, such that the experimental values represent measures that are proportional to the fractions of phosphorylated proteins. Following this line of argumentation, we rescale the variables of the model accordingly, by defining the dimensionless state variables as

$$
\begin{aligned}
x_1 &= \alpha_1 \cdot \frac{\text{pRaf}}{\text{Raf}_{\text{TOT}} \cdot s_1} \\
x_2 &= \alpha_2 \cdot \frac{\text{ppMEK}}{\text{MEK}_{\text{TOT}} \cdot s_2} \\
x_3 &= \alpha_3 \cdot \frac{\text{pERK}}{\text{ERK}_{\text{TOT}} \cdot s_3} \\
x_4 &= \alpha_4 \cdot \frac{\text{ppERK}}{\text{ERK}_{\text{TOT}} \cdot s_3}.
\end{aligned}
$$

The transformed system in terms of these new variables reads

$$
\begin{aligned}
\dot{x}_1 &= \mathbf{k_1^+}(\alpha_1 - x_1)u - \mathbf{k_1^-}x_1 + \\
&\quad + fn\left[-\tilde{\mathbf{k}}_{\mathbf{Fn}}\frac{1}{\alpha_4}s_3 x_1 x_4\right] + fp\left[\mathbf{k_{Fp}}\frac{x_4^5}{x_4^5 + \left(\frac{\tilde{\mathbf{g}}\alpha_4}{s_3}\right)^5}(\alpha_1 - x_1)\right] \\
\dot{x}_2 &= \tilde{\mathbf{k}}_{\mathbf{2}}^{+}(\alpha_2 - x_2)s_1\frac{1}{\alpha_1}x_1 - \mathbf{k_2^-}x_2 \\
\dot{x}_3 &= \tilde{\mathbf{k}}_{\mathbf{3}}^{+}(1 - x_3 - \frac{1}{\alpha_4}x_4)s_2\frac{1}{\alpha_2}x_2 + \mathbf{k_4^-}\frac{1}{\alpha_4}x_4 - \mathbf{k_3^-}x_3 - \tilde{\mathbf{k}}_{\mathbf{4}}^{+}s_2\frac{1}{\alpha_2}x_3 x_2 \\
\dot{x}_4 &= \tilde{\mathbf{k}}_{\mathbf{4}}^{+}s_2\frac{\alpha_4}{\alpha_2}x_3 x_2 - \mathbf{k_4^-}x_4.
\end{aligned}
$$

Here, bold parameters are unknown and have to be estimated. Gray parameters specify the experimental condition. We have set $\alpha_3 = 1$ w.l.o.g., since pERK was not quantified experimentally. Rescaling of parameters is given by the transformations

$$
\begin{aligned}
\tilde{g} &= \frac{g}{\text{ERK}_{\text{TOT}}} \\
\tilde{k}_{Fn} &= k_{Fn}\text{ERK}_{\text{TOT}} \\
\tilde{k}_2^+ &= k_2^+\text{Raf}_{\text{TOT}} \\
\tilde{k}_3^+ &= k_3^+\text{MEK}_{\text{TOT}} \\
\tilde{k}_4^+ &= k_4^+\text{MEK}_{\text{TOT}}.
\end{aligned}
$$

In the following, to keep the notation as simple as possible, we will neglect the tilde for the rescaled parameters, and therefore consider the obtained ODE model $\dot{x} = f(x, \theta), x \in \mathbb{R}^4_+$, with parameter vector $\theta \in \mathbb{R}^{12}_+$ given by

$$\theta = (k_1^+, k_2^+, k_3^+, k_4^+, k_1^-, k_2^-, k_3^-, k_4^-, k_{Fn}, k_{Fp}, g, K).$$

The coefficients α_i account for the effect of different antibodies and their binding affinities in the Western blot measurements. These are furthermore additionally dependent on the particular experimental conditions and the specialties of the membranes. Thus, in order to enable a comparison across experiments on different membranes, Western blot data are usually additionally normalized to a reference condition. Following the data in Santos et al., we used the states at $t^* = 5$ min as the reference condition for each individual protein for this purpose, and the model outputs were normalized accordingly:

$$\begin{aligned} z_1(t) &= \frac{x_1(t)}{x_1(t^* = 5\text{min})} = \frac{\text{pRaf}(t)}{\text{pRaf}(t^* = 5\text{min})} \\ z_2(t) &= \frac{x_2(t)}{x_2(t^* = 5\text{min})} = \frac{\text{ppMEK}(t)}{\text{ppMEK}(t^* = 5\text{min})} \\ z_3(t) &= \frac{x_4(t)}{x_4(t^* = 5\text{min})} = \frac{\text{ppERK}(t)}{\text{ppERK}(t^* = 5\text{min})}. \end{aligned}$$

These output variables are independent of the scaling factors α_i, yet these are needed to simulate the model output during the optimization. Here we chose the interval $[0, 4]$ to sample the alphas during the MCMC procedure.

A.3. Additional file 3: Formulation of the posterior distribution

For our Bayesian parameter estimation framework we need to formulate the posterior distribution,

$$p(\theta|y) = \frac{l_y(\theta)p(\theta)}{p(y)}. \tag{A.3.1}$$

We will first define the likelihood function $l_y(\theta)$. Therefore, we assume log normally distributed error models for each individual measurement,

$$\tilde{Y}_i(t_k) \sim \log N(\log x_i(t_k), \sigma_{ik}^2),$$

which leads to

$$Y_i(t_k) \sim \log N(\log x_i(t_k) - \log x_i(t^*), \sigma_{ik}^2 + \sigma_{i*}^2)$$

for the normalized data, where σ_{i*}^2 denotes the error of the reference experiment for protein i.

For the global response coefficients R_{ij}, $i, j = 1, 2, 3$ we take the values in [199], which consist of four replicates. As described in the main manuscript, the global response coefficients (GRC) are defined as

$$R_{ij} = 2\frac{\partial \ln(v_i)}{\partial \ln(p_j)} \approx 2\frac{(\bar{v}_i^{(s_j)} - \bar{v}_i^{(c)})}{(\bar{v}_i^{(s_j)} + \bar{v}_i^{(c)})},$$

where $v_1 = \text{pRaf}$, $v_2 = \text{ppMEK}$, $v_3 = \text{ppERK}$. The variables $\bar{v}_i^{(s_j)}$ and $\bar{v}_i^{(c)}$ denote the (quasi) steady state activities of component i in the case of silencing of component j and in the control case, respectively. In order to approximate these steady state values, measurement time points were set to $t_{\text{GRC}}^{\text{EGF}} = 5$ min and $t_{\text{GRC}}^{\text{NGF}} \in \{5, 15\}$ min (for more details we refer to [199] and references therein). Hence

$$R_{ij} \approx R_{ij}(t_k) \approx 2\frac{(v_i^{(s_j)}(t_k) - v_i^{(c)}(t_k))}{(v_i^{(s_j)}(t_k) + v_i^{(c)}(t_k))},$$

where $v_i^{(s_j)}(t_k)$ and $v_i^{(c)}(t_k)$ denote the activities of component i at time point t_k in the case of silencing of component j and in the control case, respectively.

The table in Fig 2.2 in the main manuscript lists estimates $\widehat{\mathbb{E}}(R_{ij}(t_k))$ and $\widehat{\sigma}(R_{ij}(t_k))$ extracted from the data in [199]. To remain consistent

with our hypothesis of log normal distributions for the (normalized) Western blot signals, we decided to use these estimates to obtain respective estimates for the parameters of the quantity

$$z'_{ij}(t_k) \coloneqq \frac{v_i^{(s_j)}(t_k)}{v_i^{(c)}(t_k)},$$

since corresponding measurement values $Y'_{ij}(t_k)$ also follow a log normal distribution. Therefore, we resolved $R_{ij}(t_k)$ for $z'_{ij}(t_k)$ to get

$$z'_{ij}(t_k) = \frac{2 + R_{ij}(t_k)}{2 - R_{ij}(t_k)}.$$

According to this, estimates for the parameters of the log normal distribution of $Y'_i(t_k)$ were set to

$$\widehat{\mathbb{E}}(Y'_{ij}(t_k)) = \frac{2 + \widehat{\mathbb{E}}(R_{ij}(t_k))}{2 - \widehat{\mathbb{E}}(R_{ij}(t_k))}$$

and

$$\begin{aligned}\widehat{\sigma}^2(Y'_{ij}(t_k)) &= \left|\frac{\partial z'_{ij}(t_k)}{\partial R_{ij}(t_k)}\right| \ \widehat{\sigma}(R_{ij}(t_k)) \\ &= \frac{4}{(2 - \widehat{\mathbb{E}}(R_{ij}(t_k)))^2}\widehat{\sigma}(R_{ij}(t_k)).\end{aligned}$$

In summary, the resulting likelihood function reads:

$$\begin{aligned}&l_{\log y}(\theta) = \\ &\prod_m \prod_{i=1}^3 \left(\prod_{t_k} \frac{1}{\sqrt{2\pi\sigma_{imk}^2}} \exp\left[-\tfrac{1}{2} \left(\frac{\log z_i^m(t_k, \theta) - \log y_i^m(t_k)}{\sigma_{imk}} \right)^2 \right] \right) \times \\ &\prod_{j=1}^3 \prod_{t^m_{\mathrm{GRC}}} \left(\frac{1}{\sqrt{2\pi\widehat{\sigma}^2_{ijm}(t^m_{\mathrm{GRC}})}} \exp\left[-\tfrac{1}{2} \left(\frac{\log \widehat{\mathbb{E}}(Y'_{ij}(t^m_{\mathrm{GRC}})) - \log z'_{ij}(t^m_{\mathrm{GRC}}, \theta)}{\widehat{\sigma}^2_{ijm}(t^m_{\mathrm{GRC}})} \right)^2 \right] \right)\end{aligned}$$

Here, $m \in \{\mathrm{EGF}, \mathrm{NGF}\}$ denote experiments with different growth

factors, the indices $i = 1, 2, 3$ and $j = 1, 2, 3$ enumerate the three output variables z_i and the three silencing experiments siRaf, siMEK and siERK, respectively. The time points $t_k \in \{10, 15, 30, 60\}$ min refer to the meausurement time points in the control experiments, and $t_{\mathrm{GRC}}^{\mathrm{EGF}} = 5$ min for the time point that is used to determine the global response coefficients in case of stimulation with EGF and $t_{\mathrm{GRC}}^{\mathrm{NGF}} \in \{5, 15\}$ min for the two time points used in the respective NGF experiments. We note here that $t_k = 5$ min does not appear in the likelihood function, since measurements at this point were used as reference experiments.

Since a priori nothing was known about the values of the parameters

$$\theta = (k_1^+, k_2^+, k_3^+, k_4^+, k_1^-, k_2^-, k_3^-, k_4^-, k_{Fn}, k_{Fp}, g, K) \in \mathbb{R}_+^{12},$$

we decided to use almost non-informative prior distributions. This was done by assuming uniform distributions on the logarithmic scale for all parameters but K in order to allow for covering several orders of magnitude for these parameters. For details on the choice of the prior boundaries and the optimization and subsequent sampling we refer to Additional file A.4.

A.4. Additional file 4: Details on the MCMC sampling procedure

In order to sample from the posterior distribution described in Additional file A.3, all numerical calculations were run on `MATLAB R2014b` (64 bit). The model and data were managed using the toolboxes `SBPD` and `SBTOOLBOX2`. `SBTOOLBOX2` with the CVODE integrator from `SUNDIALS` was employed for the integration of the ODE system. Absolute and relative error tolerances of the integrator were set to `options.abstol=1e-10` and `options.reltol=1e-10`.

In the first step we intended to find good starting values for the Markov chains and appropriate boundaries for the parameter's prior distributions. As described, all parameters except K were sampled in the log space, to cover several orders of magnitudes. Since K

θ	$\log k_1^+$	$\log k_2^+$	$\log k_3^+$	$\log k_4^+$	$\log k_1^-$	$\log k_2^-$
$\hat{\theta}^{\mathrm{MLE}}$	-5.7324	7.3475	7.8110	2.3365	-0.0865	6.2055
θ	$\log k_3^-$	$\log k_4^-$	$\log k_{Fn}$	$\log k_{Fp}$	$\log g$	K
$\hat{\theta}^{\mathrm{MLE}}$	6.8132	-0.4295	17,8312	-5.9037	-5.8563	5.6202

Table A.1.: Estimated MAP parameter values.

describes the decay or switching time in the input, which is expected from the EGF control experiments to lie approximately between 5 and below 10 minutes, we used a uniform distribution with fixed boundaries $[4, 8]$ min for this parameter directly. The boundaries for the other distributions were set heuristically via a trial and error procedure. Therefore, in a first step we optimized the posterior distribution several times with different prior boundaries and adapted the boundaries accordingly to ensure that parameter regions with very high likelihood values are not truncated by the prior distribution. Maximization of $p(\theta|y)$ was done by minimizing $-\log p(\theta|y)$ using the Matlab built-in function `fmincon`. Tolerances on the constraint violation and function value were set to `OPTIONSfmincon.TolFun=1e-6` and `OPTIONSfmincon.TolCon=1e-6`, respectively. To account for possible multiple local minima a multistart algorithm with uniformly distributed initial values was used.

Equipped with a convenient estimate $\widehat{\theta}^{\mathrm{MAP}}$ from this procedure (listed in Table A.1), boundaries were set to $[10^{\widehat{\theta}^{\mathrm{MAP}}-2}, 10^{\widehat{\theta}^{\mathrm{MAP}}+2}]$ for subsequent MCMC sampling [82].

For the implementation the `mcmcstat` toolbox with the method option 'DRAM' was used. To achieve convergence a warm-up period of $5 \cdot 10^5$ samples was carried out prior to the sampling of a parameter chain of length $3 \cdot 10^6$. Four independent chains were initialized using as starting points different parameter estimates with small objective function values. Convergence for the overall chain was assessed with the Gelman–Rubin–Brooks diagnostic using the function `mpsrf`, which returns a potential scale reduction factor R. For testing of the

individual chains the Geweke method was applied. Both diagnostics are implemented in `mcmcstat` [37].

The mean acceptance rate over the four chains was 11% in a first sampling trial. Convergence diagnostics showed $R = 1.0264$ for the Gelman–Rubin–Brooks method, but bad p-values for two chains with the Geweke method. To improve the sample quality a second sampling was carried out with initial parameters chosen from a sub-sample of the first run. The acceptance rate was improved to 20%. All chains passed the convergence test with a p-value of at least 0.8. Overall chain testing resulted in an improved value of $R = 1.0051$.

The estimates of the marginal distributions of the parameters from this second run are shown in Additional file A.5. Highest and lowest indicated values on the abscissa correspond to lower and upper boundaries of the respective prior distributions. Estimates of the MAPs and the means are indicated by dashed gray lines and gray lines, respectively. It can be seen that most of these 1D marginals show a large variance, indicating that these are only vaguely defined, and that the data do not contain much information about these individual parameters. This is indeed not unusual in case of quantitative models and only few data points with high measurement noises. Only the distributions of the parameters k_1^- and k_4^- have significantly lower variances than the respective prior distributions, indicating a high sensitivity of the model output on these parameters.

A.5. Additional file 5: Estimated marginal parameter distributions from the MCMC sample

See Figure A.3.

A.6. Additional file 6: Scatterplot matrix of a subset of the parameters from the MCMC sample

See Figure A.4.

A.7. Additional file 7: Details on the classification scheme with the CBA

In the main text we presented the steps of the circuit-breaking algorithm applied to our network model, represented schematically in Fig 2.6, Subfigs A-C. The final step of the algorithm, needed to obtain all steady-state coordinates for all variables of the system, requires the calculation of the zeros of the circuit-characteristics $c(\kappa, \theta_i)$ for all sample points θ_i. This condition is given by Equation (2.8) in the main manuscript.

For the rescaled and normalized model (see Additional file A.2)

$$\begin{aligned}
\dot{x}_1 =& k_1^+(\alpha_1 - x_1)u - k_1^- x_1 + \\
&+ fn\left[-k_{Fn}\frac{1}{\alpha_4}s_3 x_1 x_4\right] \\
&+ fp\left[k_{Fp}\frac{x_4^m}{x_4^m + \left(\frac{\tilde{g}\alpha_4}{s_3}\right)^m}(\alpha_1 - x_1)\right] && \text{(A.7.1a)} \\
\dot{x}_2 =& \tilde{k}_2^+(\alpha_2 - x_2)s_1\frac{1}{\alpha_1}x_1 - k_2^- x_2 && \text{(A.7.1b)} \\
\dot{x}_3 =& \tilde{k}_3^+(1 - x_3 - \frac{1}{\alpha_4}x_4)s_2\frac{1}{\alpha_2}x_2 \\
&+ k_4^-\frac{1}{\alpha_4}x_4 - k_3^- x_3 - \tilde{k}_4^+ s_2\frac{1}{\alpha_2}x_3 x_2 && \text{(A.7.1c)} \\
\dot{x}_4 =& \tilde{k}_4^+ s_2\frac{\alpha_4}{\alpha_2}x_3 x_2 - k_4^- x_4, && \text{(A.7.1d)}
\end{aligned}$$

this translates into finding the intersection of two one-dimensional

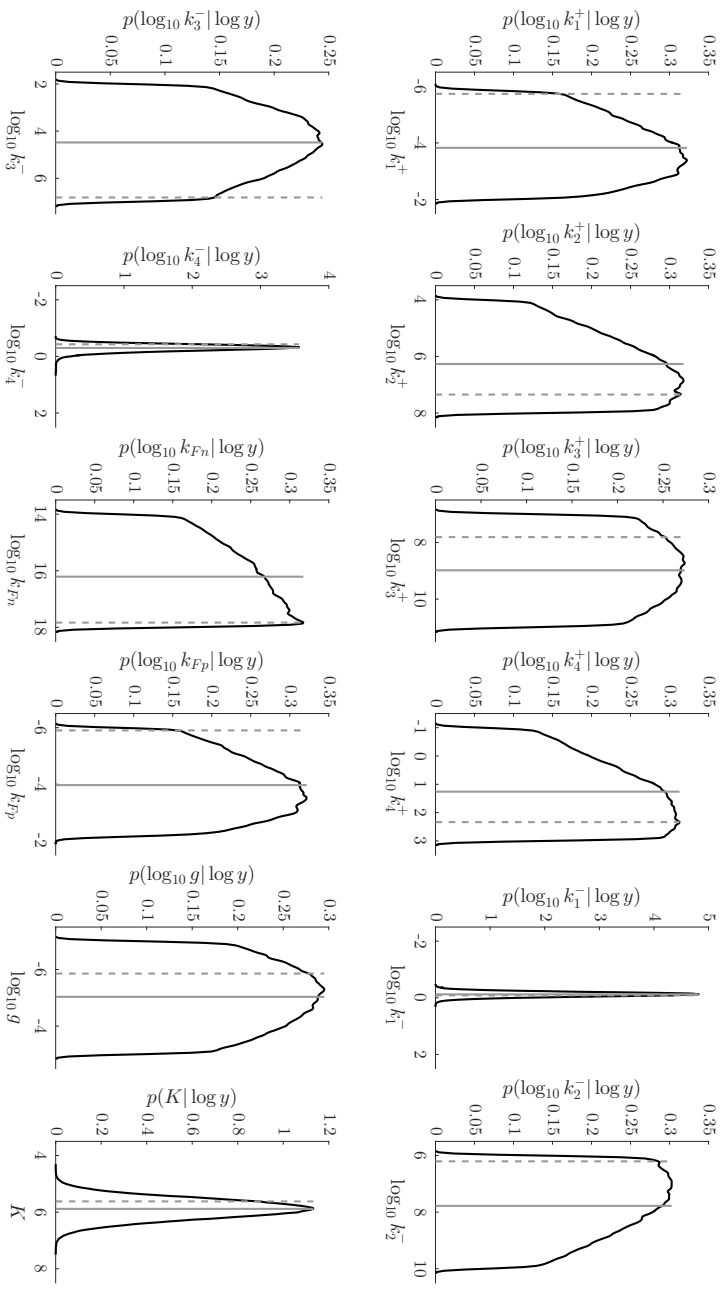

$p(\log_{10} k_1^+ | \log y)$
$\log_{10} k_1^+$
$p(\log_{10} k_2^+ | \log y)$
$\log_{10} k_2^+$
$p(\log_{10} k_3^+ | \log y)$
$\log_{10} k_3^+$
$p(\log_{10} k_4^+ | \log y)$
$\log_{10} k_4^+$
$p(\log_{10} k_1^- | \log y)$
$\log_{10} k_1^-$
$p(\log_{10} k_2^- | \log y)$
$\log_{10} k_2^-$
$p(\log_{10} k_3^- | \log y)$
$\log_{10} k_3^-$
$p(\log_{10} k_4^- | \log y)$
$\log_{10} k_4^-$
$p(\log_{10} k_{Fn} | \log y)$
$\log_{10} k_{Fn}$
$p(\log_{10} k_{Fp} | \log y)$
$\log_{10} k_{Fp}$
$p(\log_{10} g | \log y)$
$\log_{10} g$
$p(K | \log y)$
K

Figure A.3.: Kernel density estimates from the MCMC parameter samples of the marginal parameter distributions. Estimates of the MAPs and the means are indicated by dashed gray lines and gray lines, respectively.

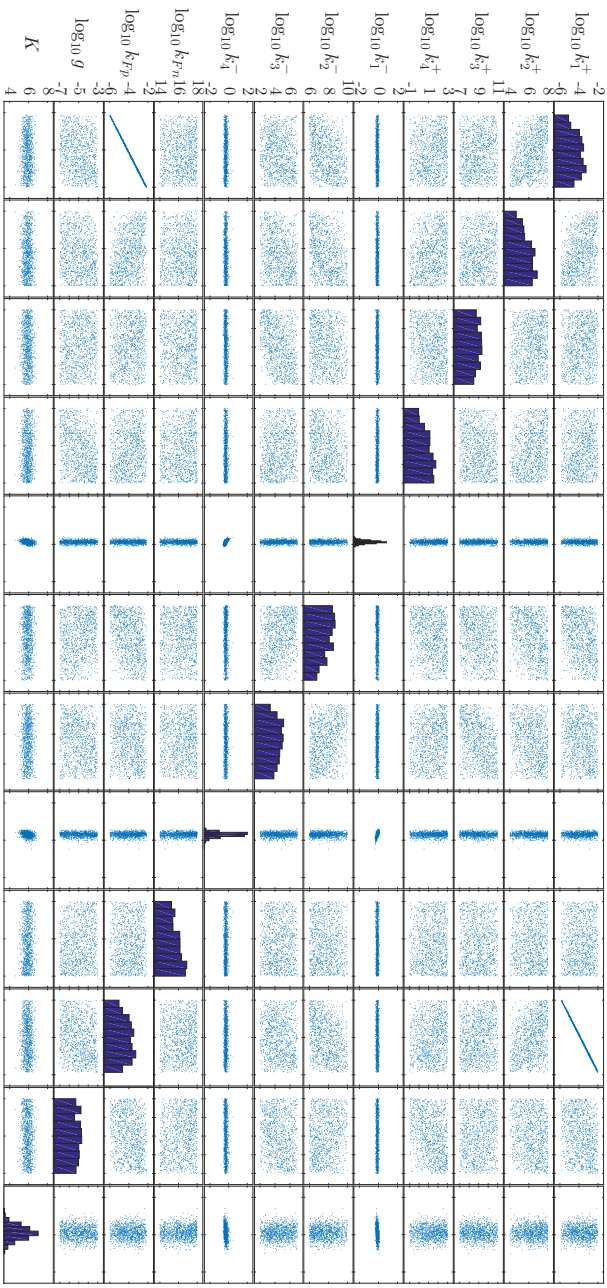

$\log_{10} k_1^+$
$\log_{10} k_2^+$
$\log_{10} k_3^+$
$\log_{10} k_4^+$
$\log_{10} k_1^-$
$\log_{10} k_2^-$
$\log_{10} k_3^-$
$\log_{10} k_4^-$
$\log_{10} k_{Fn}$
$\log_{10} k_{Fp}$
$\log_{10} g$
K

Figure A.4.: 2D scatter plot matrix for the parameters with histograms on the diagonal.

functions of κ in the particular example:

$$k_4^- \kappa = k_4^+ s_2 \frac{\alpha_4}{\alpha_2} \bar{x}_2(\kappa, \theta_i) \bar{x}_3(\kappa, \theta_i), \tag{A.7.2}$$

as can be easily verified by looking at Equation (A.7.1d) of the ODE model (A.7.1). The steady states of the other three state variables as function of κ and of the model parameters are given by the following expressions:

$$\bar{x}_1(\kappa, \theta_i) = \alpha_1 \cdot \frac{k_{Fp} h(\kappa, \theta_i)}{k_1^- + k_{Fp} h(\kappa, \theta_i)} \tag{A.7.3a}$$

$$\bar{x}_2(\kappa, \theta_i) = \alpha_2 \cdot \frac{k_2^+ \frac{s_1}{\alpha_1} \bar{x}_1(\kappa, \theta_i)}{k_2^- + k_2^+ \frac{s_1}{\alpha_1} \bar{x}_1(\kappa, \theta_i)} \tag{A.7.3b}$$

$$\bar{x}_3(\kappa, \theta_i) = \frac{\frac{k_4^-}{\alpha_4} \kappa + k_3^+ \left(1 - \frac{\kappa}{\alpha_4}\right) \frac{s_2}{\alpha_2} \bar{x}_2(\kappa, \theta_i)}{k_3^- + \left(k_3^+ + k_4^+\right) \frac{s_2}{\alpha_2} \bar{x}_2(\kappa, \theta_i)}. \tag{A.7.3c}$$

In Equation (A.7.3a) the function $h(\kappa, \theta_i)$ represents the Hill function

$$h(\kappa, \theta_i) = \frac{\kappa^m}{\kappa^m + (g\alpha_4/s_3)^m}. \tag{A.7.4}$$

Equation (A.7.2) is solved numerically. The set of solutions $\{\bar{\kappa}\}$ corresponds to the steady state coordinates of variable z_3.

A.8. Additional file 8: Sensitivity analysis of the simulation-based classification scheme

Sensitivity analysis of the simulation-based classification scheme, see Figure A.5.

A.9. Additional file 9: Simulation-based classification of sample trajectories with varying minimal switching times

See Figure A.6.

A.10. Additional file 10: Classification of sample trajectories with varying total ERK concentration

See Figure A.7.

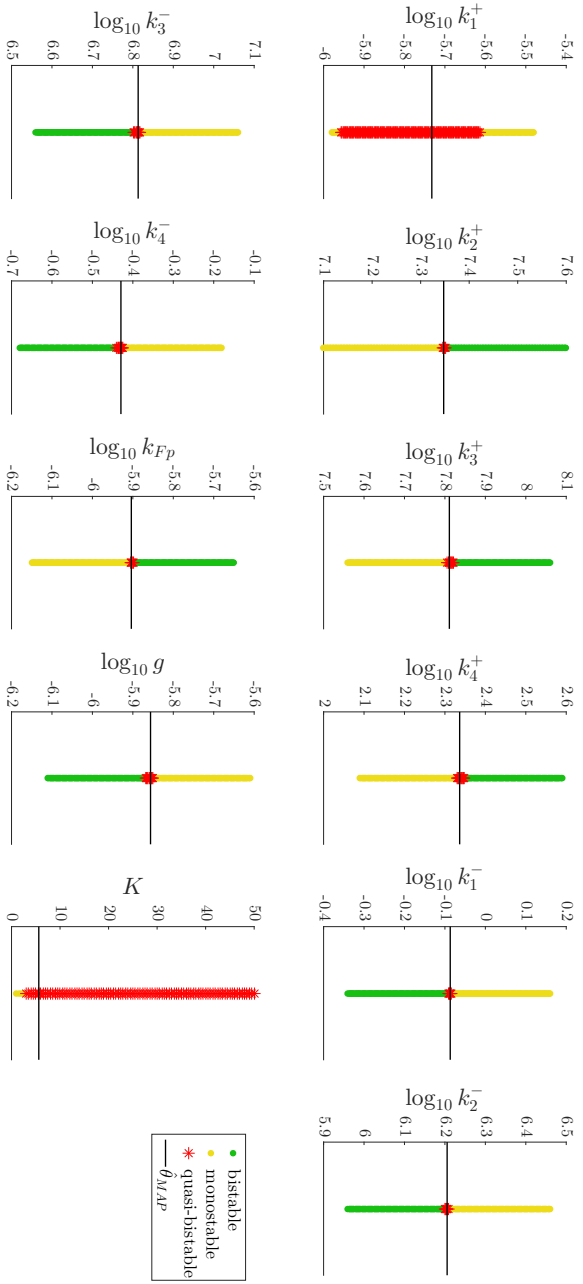

$\log_{10} k_1^+$
$\log_{10} k_3^-$
$\log_{10} k_2^+$
$\log_{10} k_4^-$
$\log_{10} k_3^+$
$\log_{10} k_{Fp}$
$\log_{10} k_4^+$
$\log_{10} g$
$\log_{10} k_1^-$
K
$\log_{10} k_2^-$
bistable
monostable
quasi-bistable
$\hat{\theta}_{MAP}$

Figure A.5.: Model parameters were varied independently each at a time about the maximum-a-posteriori estimator. The simulation-based classification scheme was conducted repeatedly for these variations. Except for the parameters k_1^+ and K, which do not influence the limit sets of the system for $u = 0$, classification is highly sensitive to parameter variations.

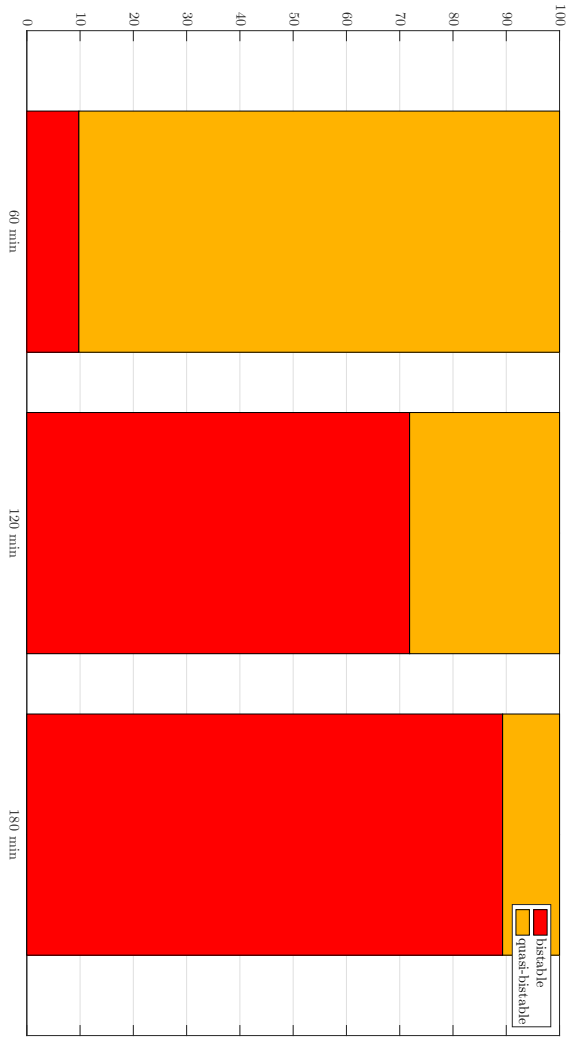

0
10
20
30
40
50
60
70
80
90
100
60 min
120 min
180 min
bistable
quasi-bistable

Figure A.6.: The 60 min case represents the full MCMC sample. For the 120 (180) min case all quasi-bistable trajectories that switch between 60 and 120 (180) min were filtered out.

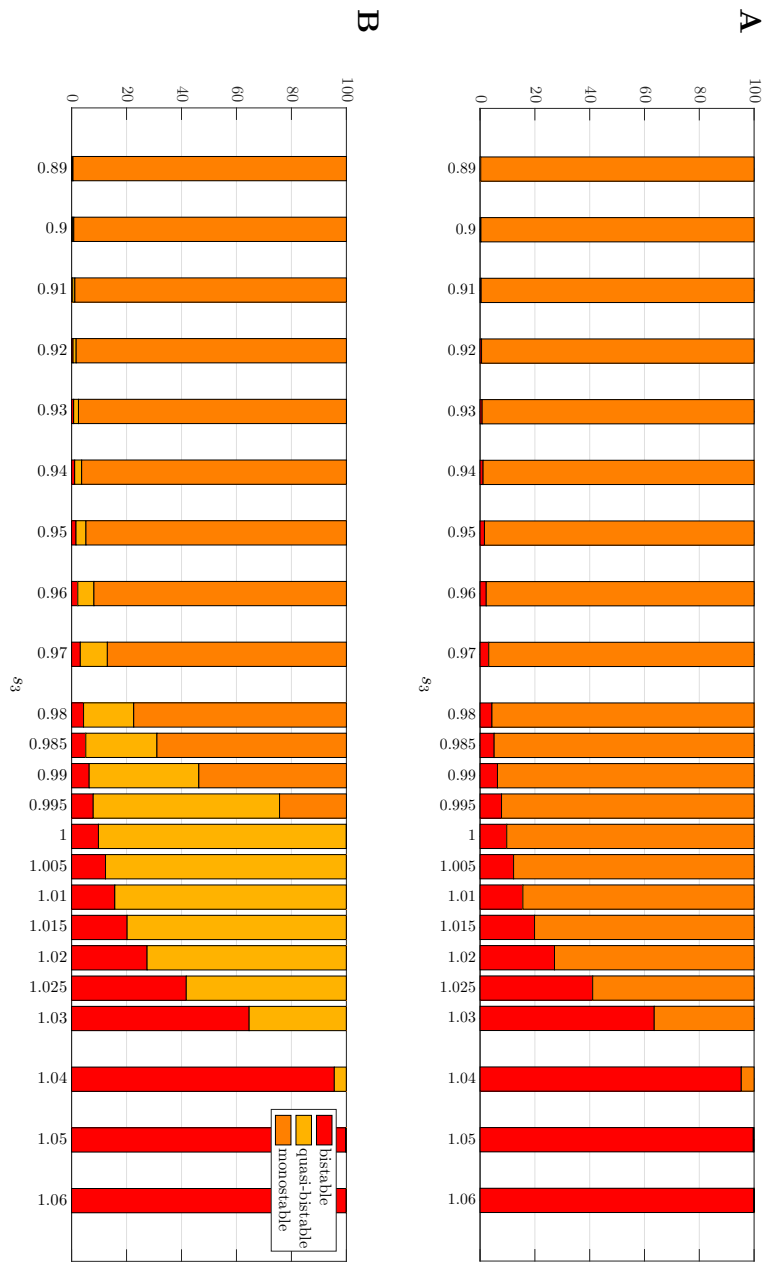

A
B
0
20
40
60
80
100
0.89
0.9
0.91
0.92
0.93
0.94
0.95
0.96
0.97
0.98
0.985
0.99
0.995
1
1.005
1.01
1.015
1.02
1.025
1.03
1.04
1.05
1.06
s3
bistable
quasi-bistable
monostable

Figure A.7.: A. Steady state analysis via the CBA and classification into monostable and bistable trajectories. B. Simulation-based classification into bistable, quasi-bistable and monostable trajectories.

B. Supporting Information for Chapter 3

B.1. Positive feedback model (model 1): Data pre-processing

B.1.1. Additional experimental data

Fig. B.1 shows PKD and DLC1 phosphorylation time courses after stimulation with the phorbol ester PDBu, in the absence or presence of the PKD inhibitor kb-NB.

B.1.2. Normalization of experimental data

Western blot data are usually normalized in a multistep procedure, including background corrections and normalization to a loading control to diminish spurious signals resulting from loading differences. Further normalization is required to enable a comparison across different replicates. For this purpose, normalization to a control experiment is a commonly applied standard procedure, which we also use in this study. Normalization is an important data pre-processing step, which also affects the statistical properties of the normalized data [135, 228] and therefore has an impact on state estimation and hypothesis testing. In accordance with [50], we re-normalized experimental data to the highest signal value to avoid normalization to values with low signal-to-noise ratios. Respective data is shown in Table B.1.

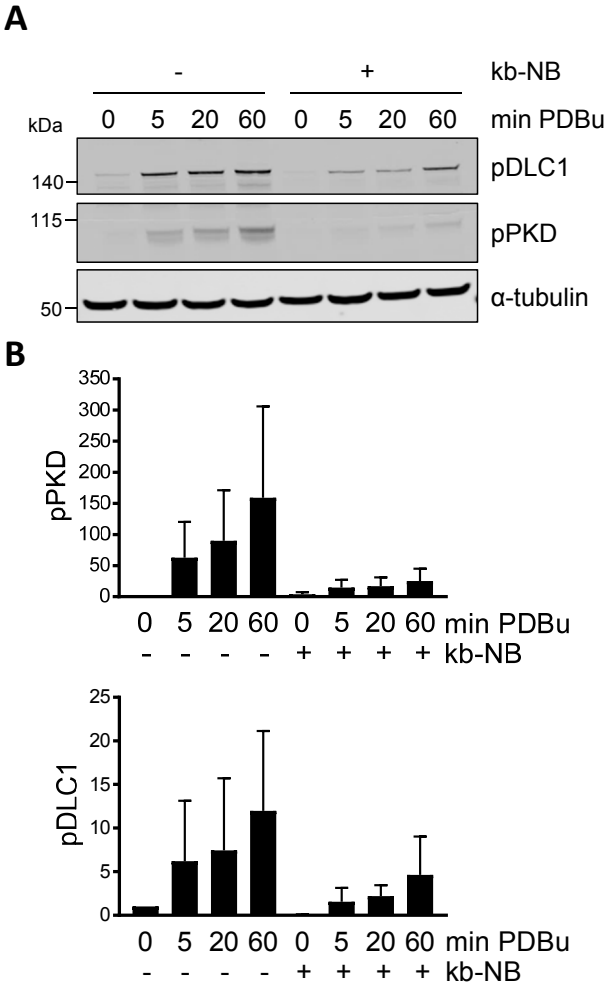


Figure B.1.: (A) Expression of GFP-DLC1 in Flp-In GFP-DLC1 was induced with doxycycline. The next day, cells were treated with the PKD inhibitor kb NB 142-70 for 2 h, followed by PDBu treatment for the times indicated. Cells lysates were analyzed by immunoblotting. (B) Band intensities from four independent experiments were quantified and normalized to the loading control and control sample (mean ± SEM).

Experiment 1: Input PDBu, control (Fig. B.1)								
time	pDLC1				pPKD			
0 min		0.05	0.18			0.00	0.08	
5 min		0.60	0.24			0.39	0.47	
20 min		0.72	0.29			0.56	0.73	
60 min		1	1			1	1	
Experiment 2: Input PDBu, kb-NB (Fig. B.1)								
time	pDLC1				pPKD			
0 min		-	-			0.02	0.04	
5 min		0.14	0.08			0.09	0.18	
20 min		0.17	0.24			0.10	0.27	
60 min		0.42	0.28			0.15	0.43	
Experiment 3: Input nocodazole, control (Fig. 3.2A)								
time	pDLC1				pPKD			
0 min	0.32	0.66	0.57	0.36	0.28	1.60	0.52	0.46
10 min	0.72	0.74	0.96	0.85	0.80	1.39	0.89	0.68
30 min	1	1	1	1	1	1	1	1
Experiment 4: Input nocodazole, kb-NB (Fig. 3.2A)								
time	pDLC1				pPKD			
10 min	0.26	0.50	0.54	0.79	0.57	0.92	0.69	0.75
30 min	0.96	0.72	0.78	0.83	0.98	0.97	0.80	1.06
Experiment 5: Input nocodazole, Gö-6976 (Fig. 3.2A)								
time	pDLC1				pPKD			
10 min	0.11	0.44	0.46	0.49	0.33	0.20	0.71	0.59
30 min	0.09	0.16	0.54	0.47	0.64	0.65	0.77	0.57
Experiment 6: pPKD and pDLC1, control (Fig. 3.1B)								
	pDLC1				pPKD			
	0.14	0.86	0.80		0.53	0.31	0.41	
Experiment 7: pPKD and pDLC1, Rho ca (Fig. 3.1B)								
	pDLC1				pPKD			
	1	1	1		1	1	1	

Table B.1.: Experimental data, normalized to the highest signal value

B.1.3. Significance test for the effect of inhibitors kb-NB and Gö-6976

Significance of the effect of the inhibitors kb-NB and Gö-6076 was investigated by setting up a parametrized dynamic model and using an F-test. For this, we defined and compared two nested parametrized model variants for each inhibitor. In the null hypothesis H_0, the inhibitor does not act significantly and hence the data can be described with a single parametrized hyperbolic curve,

$$H_0: \quad \varphi(t) = \varphi_0 + 1 - \exp(-\lambda t), \tag{B.1.1}$$

with parameters φ_0 and λ. The alternative hypothesis H_1 assumes a significant influence of the inhibitor under investigation on the pPKD and pDLC1 time courses. Therefore, two curves $\varphi^c(t)$ and $\varphi^i(t)$ were defined for the control and inhibition experiment, respectively:

$$H_1: \quad \varphi^c(t) = \varphi_0^c + 1 - \exp(-\lambda^c t) \tag{B.1.2a}$$

$$\varphi^i(t) = \varphi_0^i + b^i(1 - \exp(-\lambda^i t)), \tag{B.1.2b}$$

with parameters φ_0^c, λ^c, φ_0^i, b^i and λ^i, where superscripts c and i symbolize the control and the inhibition experiments, respectively.

For model calibration we chose the method of least squares with normalized data taken from Table B.1. The normalization point at $t = 30$ min was taken into account by a constraint on φ_0,

$$\varphi(30\,\text{min}) = 1 = \varphi_0 + 1 - \exp(-\lambda \cdot 30\,\text{min}), \tag{B.1.3}$$

leading to $\varphi_0 = \exp(-\lambda \cdot 30\,\text{min})$. Analogously, we set $\varphi_0^c = \exp(-\lambda^c \cdot 30\,\text{min})$ for H_1. We define $y_{j,l}^c(t_k)$ and $y_{j,l}^i(t_k)$ as measurement outputs for the control and inhibition experiment, respectively. Thus, the residual sum of squares for H_0 was specified as

$$RSS_{1,j} = \sum_{t_k} \sum_{l=1}^{4} \left[(y_{j,l}^c(t_k) - \varphi(t_k))^2 + (y_j^i(t_k) - \varphi(t_k))^2 \right] \tag{B.1.4}$$

and respectively for H_1

$$RSS_{2,j} = \sum_{t_k} \sum_{l=1}^{4} \left[(y^c_{j,l}(t_k) - \varphi^c(t_k))^2 + (y^i_{j,l}(t_k) - \varphi^i(t_k))^2 \right], \tag{B.1.5}$$

where $j \in \{\text{pPKD, pDLC1}\}$, $k = 1, \ldots, 3$ and $l = 1, \ldots, 4$ denote outputs, the time points with $t_k \in \{0, 10, 30\}$min and the number of replicates per experimental condition. For parameter estimation the residual sum of squares was minimized separately for both measurement outputs using the `fmincon` algorithm in `Matlab 2016b` (64 bit). All optimizer options were set to default with parameter boundaries from 0 to 5 on a linear scale. A multistart optimization using 1000 starting points was performed for all model variants and measurement outputs.

The F-value for the F-test was calculated via

$$F = \frac{\dfrac{RSS_1 - RSS_2}{p_2 - p_1}}{\dfrac{RSS_2}{n_{\text{data}} - p_2}}, \tag{B.1.6}$$

with

$$RSS_1 = RSS_{1,\text{pPKD}} + RSS_{1,\text{pDLC1}} \tag{B.1.7a}$$

$$RSS_2 = RSS_{2,\text{pPKD}} + RSS_{2,\text{pDLC1}}. \tag{B.1.7b}$$

The constants $p_1 = 2$ and $p_2 = 8$ denote the numbers of parameters for H_0 and H_1, respectively. The total number of data points including pPKD and pDLC1 time course data was $n_{\text{data}} = 32$. In order to test H_0, we consider the tails of the $\mathcal{F}$-distribution with $(p_2 - p_1,\ n_{\text{data}} - p_2)$ degrees of freedom. An $\alpha = 5\%$ level of significance corresponds to a critical value $F_\alpha = 2.51$, and H_0 is rejected if the observed F value exceeds F_α. Results of this testing procedure are shown in Fig. 3.2B of the main manuscript.

B.1.4. Selection of an error model

For model calibration we exploit maximum-likelihood estimation, which requires the choice of an appropriate error model for observed outputs. Additive normally distributed error models or multiplicative log-normal error models are most frequently used for this purpose. In Kreutz et al. [135] it was argued that the main source of biological variability and experimental noise is multiplicative and log-normally distributed, which suggests a log-transformation of the data to obtain approximately normally distributed data. The mixed error model from which additive and multiplicative effects are deduced in [135] is, however, not applicable in our setting due to low numbers of replicates per condition. Here we decided to select an error model as a data-driven pre-processing step on the normalized data. Such an a priori analysis is computationally more attractive than integrating the selection of an error model directly into model calibration and allows for a much more comprehensive comparison of different error models. We used additive normal and multiplicative log-normal error models, in combination with maximum-likelihood estimators for the means and the variances. Since the maximum-likelihood variance estimator is biased, we additionally included also unbiased variance estimators. Moreover, we compared independent standard deviation estimation for each condition and each time point with the estimation of partly pooled standard deviations. The first pooling version averages the standard deviation of each experimental condition separately for each output, resulting in 12 standard deviation parameters. For the second pooling the standard deviations were further averaged per Western blot according to Fig. 3.1A, 3.2A and B.1A. We compared all model variants by using different information criteria. Results are shown in Table B.2. Shown are the likelihood values $\mathcal{L}$, the Akaike information criterion (AIC), the corrected Akaike information criterion (AIC_c), the Bayesian information criterion (BIC) and Akaike weights, i.e.

$$\Delta_i = \mathrm{AIC}_{c_i} - \min(\mathrm{AIC}_c)$$
$$\mathrm{AW}_i = \frac{\exp(-\Delta_i/2)}{\sum_{r=1}^{R} \exp(-\Delta_r/2)}.$$

For each case, the three superior ones are color marked. The AIC agrees with the likelihood values in the choice of the best model. Both select the most complex model. This selection is, however, different from that of AIC_c, BIC and AW, all of which also agree in the choice of the best error model. Overall, these results reflect that the later penalize complexity more than the AIC. It should be noted here that both AIC and BIC are approximations that assume a large sample size compared to the number of parameters [4, 205], which is not given here. Thus we judge AIC_c, which corrects for finite sample sizes [38], and Akaike weights the more suitable criteria here, which also completely agree in their ranking. Hence we decided to select the most plausible error model according to these two criteria. According to the evidence Table in [119], AIC_c records positive evidence between the best and the second best model. Based on these results, we decided to use an additive normal error model and six standard deviation for the following modeling study. The six standard deviation pools are represented by six parameters σ_j^i, $i = 1, 2, 3$, $j \in \{\text{pPKD, pDLC1}\}$ depicting the standard deviation for each experiment and output pooled as described above. Table B.3 shows the composition of experiments that were used for pooling.

B.2. Model 1: Modeling and model calibration

In the following we provide details on the positive feedback modeling approach and calibration to experimental data.

B.2.1. Modeling approach and normalization

According to Fig. 3.1E, we built up a simplified two state variable model,

$$\text{pP}\dot{\text{K}}\text{D} = k(\text{DLC1}, \theta)\text{PKD} - \theta_1\text{pPKD} \tag{B.2.1a}$$

$$\text{pD}\dot{\text{L}}\text{C1} = \theta_2\text{pPKD} \cdot \text{DLC1} - \theta_3\text{pDLC1}, \tag{B.2.1b}$$

with model parameters θ. For simplicity and because the ratio of substrate and kinase molecules are unknown we used mass action kinetics

Error model	# parameters	$\mathcal{L}$	AIC	AIC$_c$	BIC	AW
$\mathcal{N}$, unbiased, σ_i	54	1.0e25	-7.19	230.41	121.44	3.5e-41
$\mathcal{N}$, unbiased, $\bar{\sigma}_{12}$	39	1.4e20	-14.73	63.27	78.17	6.9e-05
$\mathcal{N}$, unbiased, $\bar{\sigma}_6$	33	1.7e15	-4.14	44.64	74.46	0.7662
$\mathcal{N}$, biased, σ_i	54	3.9e26	-14.45	223.15	114.18	1.3e-39
$\mathcal{N}$, biased, $\bar{\sigma}_{12}$	39	4.9e20	-17.30	60.70	75.60	0.0002
$\mathcal{N}$, biased, $\bar{\sigma}_6$	33	5.2e14	-1.75	47.03	76.85	0.2317
log $\mathcal{N}$, unbiased, σ_i	54	7.6e25	-11.18	226.42	117.45	2.6e-40
log $\mathcal{N}$, unbiased, $\bar{\sigma}_{12}$	39	1.9e18	-6.21	71.79	86.69	9.7e-07
log $\mathcal{N}$, unbiased, $\bar{\sigma}_6$	33	3.6e12	8.19	56.97	86.79	0.0016
log $\mathcal{N}$, biased, σ_i	54	2.7e27	-18.44	219.16	110.19	9.7e-39
log $\mathcal{N}$, biased, $\bar{\sigma}_{12}$	39	2.1e18	-6.41	71.59	86.49	1.1e-06
log $\mathcal{N}$, biased, $\bar{\sigma}_6$	33	1.9e11	14.08	62.86	92.69	8.4e-05

Table B.2.: Error model selection procedure based on information criteria. In total $N = 80$ measurements were used. Standard deviations were either estimated individually for each condition and time point (σ_i), or by pooling over time series and outputs ($\bar{\sigma}_{12}$) or over experiments and outputs ($\bar{\sigma}_6$). The three superior models are marked in green (top), orange (second top) and yellow (third top).

variance parameter	pooled experiments (state variable)
σ^1_{PKD}	1,2 (pPKD)
σ^2_{PKD}	3,4,5 (pPKD)
σ^3_{PKD}	6 (pPKD)
σ^1_{DLC1}	1,2 (pDLC1)
σ^2_{DLC1}	3,4,5 (pDLC1)
σ^3_{DLC1}	6 (pDLC1)

Table B.3.: Experiments and states used for calculation of the six standard deviation pools depending on the experimental conditions given in Table B.1.

wherever applicable. The PKD phosphorylation rate $k(\mathrm{DLC1}, \theta)$ depends on DLC1 via Rho and on the experimental treatment of the cell culture and is specified in Table B.4 for the different treatments used for model calibration.

We eliminated PKD and DLC1 by assuming mass conservation of respective total amounts,

$$\mathrm{PKD_{tot}} = \mathrm{PKD} + \mathrm{pPKD} \tag{B.2.2a}$$

$$\mathrm{DLC1_{tot}} = \mathrm{DLC1} + \mathrm{pDLC1}. \tag{B.2.2b}$$

Normalization of both state variables to total concentrations,

$$x_1 = \frac{\mathrm{pDLC1}}{\mathrm{DLC1_{tot}}} \tag{B.2.3a}$$

$$x_2 = \frac{\mathrm{pPKD}}{\mathrm{PKD_{tot}}} \tag{B.2.3b}$$

leads to

$$\begin{aligned}\dot{x}_1 = &(\theta_0(1 - \tilde{\theta}_6(1 - x_2)) + \theta_4 u_1 + \theta_5 u_2 + \alpha_3\theta_9) \\ &\times (1 - \alpha_1\theta_7 - \alpha_2\theta_8)(1 - x_1) - \theta_1 x_1\end{aligned} \tag{B.2.4a}$$

$$\dot{x}_2 = \tilde{\theta}_2(1 - x_2)x_1 - \theta_3 x_2 \tag{B.2.4b}$$

where $\tilde{\theta}_6 = \theta_6 \mathrm{DLC1}_{\mathrm{tot}}$, $\tilde{\theta}_2 = \theta_2 \mathrm{DLC1}_{\mathrm{tot}}$ and u_i, α_i are Boolean variables that are used to indicate the treatment. For the sake of simplicity the parameters $\tilde{\theta}_6$ and $\tilde{\theta}_2$ will be called θ_6 and θ_2 in the following. Together with the standard deviations of the error model, the vector of unknown parameters is given by

$$\begin{aligned}\boldsymbol{\theta} = (&\theta_0,\, \theta_1,\, \theta_2,\, \theta_3,\, \theta_4,\, \theta_5,\, \theta_6,\, \theta_7,\, \theta_8,\, \theta_9,\\ &\sigma^1_{\mathrm{PKD}},\, \sigma^2_{\mathrm{PKD}},\, \sigma^3_{\mathrm{PKD}},\, \sigma^1_{\mathrm{DLC1}},\, \sigma^2_{\mathrm{DLC1}},\, \sigma^3_{\mathrm{DLC1}}) \in \mathbb{R}^{16}_{+}.\end{aligned}$$

B.2.2. Likelihood function

For model calibration we used maximum-likelihood estimation with the previously chosen error model. We also included the pooled standard deviations as optimization parameters in order to be more flexible with data points that cannot properly be fitted. We use $y_{ijl}(t_k)$ to denote measurement outputs. The indices $i = 1, \ldots, 7$, $j = 1, 2$, $k = 1, \ldots, 6$ and $l = 1, \ldots, 4$ denote different experimental conditions, enumeration of the outputs, the time points with $t_k \in \{0, 5, 10, 20, 30, 60\}$min and the number of replicates per experimental condition, respectively. According to the error model, we assume

$$Y_{ij}(t_k) \sim \mathcal{N}(z_{ij}(\boldsymbol{\theta}, t_k), \sigma^2_{ij}), \tag{B.2.5}$$

where z_{ij} denotes the normalized output according to Table B.1,

$$z_{ij}(\boldsymbol{\theta},\, t_k) = \frac{x_{ij}(\boldsymbol{\theta},\, t_k)}{x_{\mathrm{ctrl},j}(\boldsymbol{\theta},\, t_{\mathrm{ctrl}})}, \tag{B.2.6}$$

where $x_{\mathrm{ctrl},\,j}$ is the value of the simulated substrate j under the conditions of the experimental data used for normalizing y_{ij} and t_{ctrl} is the normalization time point. The likelihood function then reads

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= p(y|\boldsymbol{\theta})\\ &= \prod_{i=1}^{7} \prod_{j=1}^{2} \prod_{t_k} \prod_{l=1}^{4} \frac{1}{\sigma_{ij}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_{ijl}(t_k) - z_{ij}(\boldsymbol{\theta}, t_k)}{\sigma_{ij}}\right)^2\right].\end{aligned} \tag{B.2.7}$$

Treatment (u, α)	$k(\text{DLC1}, \theta)$	Remark
–	$\begin{cases} \theta_0(1-\theta_6\text{DLC1}) & \theta_6\text{DLC1} \in [0,1] \\ 0 & \text{otherwise} \end{cases}$	unphosphorylated DLC1 decreases the basal PKD phosphorylation rate
PDBu	$k + \theta_4$	PDBu triggers PKD phosphorylation
Noc	$k + \theta_5$	nocodazole triggers PKD phosphorylation
kb-NB	$\begin{cases} k(1-\theta_8) & \theta_8 \in [0,1] \\ 0 & \text{otherwise} \end{cases}$	kb-NB decreases the overall PKD phosphorylation rate
Gö-6976	$\begin{cases} k(1-\theta_7) & \theta_7 \in [0,1] \\ 0 & \text{otherwise} \end{cases}$	Gö-6976 acts similar to kb-NB
Rho ca	$k + \theta_9$	constitutively active Rho increases PKD phosphorylation

Table B.4.: PKD phosphorylation rate $k(\text{DLC1}, \theta)$ depending on the experimental setup and on DLC1

We used the negative log-likelihood function for optimization,

$$\max_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}\in\Theta} -\log \mathcal{L}(\boldsymbol{\theta}) = J_{\text{opt}}, \tag{B.2.8}$$

where Θ is the set of acceptable parameters and J_{opt} is called the objective function value.

B.2.3. Optimization details

In order to evaluate the likelihood function (B.2.7), all simulations of the model (B.2.4) were performed via `Matlab 2016b` (64 bit). For model handling we used the `SBTOOLBOX2` and `SBPD` toolboxes, which make use of the CVODE solver from `SUNDIALS` for integration. Integrator options were set to `options.abstol = 1e-10` and `options.reltol = 1e-10`.

The optimization problem (B.2.8) was solved with the Pattern Search algorithm, which gave most reliable results in several tests for our setting. Pattern Search was introduced by [102] and uses a mesh in the parameter space in order to move step-wise to the minimum of the objective function. During a parameter poll a decrease in the objective function value is called a success and leads to an increase of the mesh size, whereas in case that the objective function value cannot be decreased the mesh size is reduced.

For implementation in `Matlab` we used the internal algorithm `patternsearch` with the following options:

- `OptionsPatternsearch.Cache = 'off'`,
- `OptionsPatternsearch.CompletePoll = 'off'`,
- `OptionsPatternsearch.MeshAccelerator = 'on'`,
- `OptionsPatternsearch.ScaleMesh = 'on'`,
- `OptionsPatternsearch.MaxFunEvals = 9000p`,
- `OptionsPatternsearch.MaxIter = 300p`,

where $p = 16$ is the number of unknown parameters.

As further options we set boundaries for these parameters. We used a logarithmic scale for the reaction rate parameters $\theta_0, \ldots, \theta_9$,

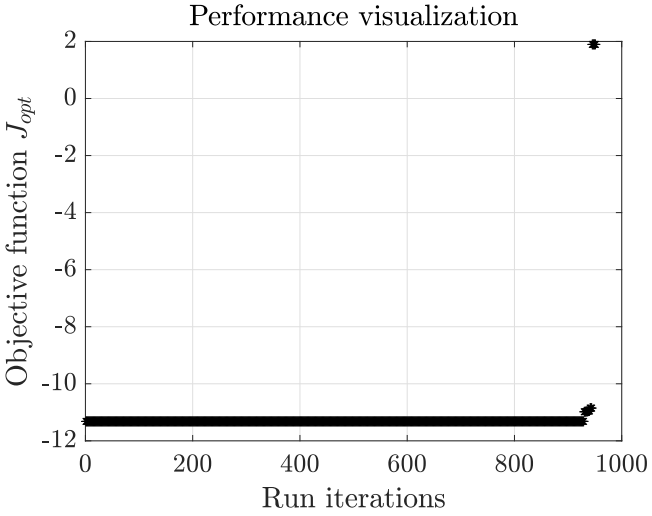


Figure B.2.: Sorted objective function values of converged parameter sets for model 1.

which allows to cover many orders of magnitude. Our boundaries comprise four orders of magnitude and were adjusted through several optimization runs, such that the optima that were found do not lie on one of the boundaries. Boundaries for the parameters θ_7 and θ_8, the efficiencies of the two PKD inhibitors, were set to [0, 1] according to (B.2.4). For the standard deviations σ_j^i we used the empirical estimates σ_{emp} to set reasonable boundaries, which were finally set to $\sigma_j^i \in \left[\frac{2}{3}\sigma_{\text{emp}}, \frac{3}{2}\sigma_{\text{emp}}\right]$.

Optimization was performed with 1000 latin hypercube samples as starting values. The initial conditions $x_1(0)$ and $x_2(0)$ were under steady state assumption calculated according to the conditions specified for each experiment. Fig. B.2 shows the resulting objective function values for 951 converged parameter sets[1], 929 of which form a nice plateau.

[1]Ensuring that the magnitude of the mesh size is less than the specified tolerance and constraint violation is less than those specified in options.ConstraintTolerance.

Figs. B.3 and B.4 show parameter scatterplots. The parameters $\theta_2, \theta_3, \theta_4, \theta_5, \theta_7$ and θ_8 are clearly identifiable from the experiments. It is also plausible that $\theta_5 > \theta_4$, since kb-NB is able to abolish the signal almost completely in case of stimulation with PDBu, while PKD is still considerably activated in case of stimulation with nocodazole. Similarly, θ_7 is close to the maximal value 1 and larger than the influence of the inhibitor kb-NB, which also reflects experimental observations. The basal PKD phosphorylation and dephosphorylation rates θ_0 and θ_1, respectively, have a much larger uncertainty. This is also true for the parameter θ_9, which characterizes the effect of RhoA on PKD that also showed a high variance in the experiments. Surprisingly at first glance, these three parameters are almost perfectly correlated. Having a closer look at the parameters, this can easily be explained in the following way: The optimizer assigns quite small values to the parameter θ_6, which is a measure for the influence of DLC1 on PKD. Given this, PKD dynamics is hardly affected by DLC1. In this case, the strong correlations directly follow from a steady state analysis of decoupled PKD. The correlation between θ_0 and θ_1 follows from the steady state conditions in the control case, while the correlation of θ_9 with both parameters results from the observed steady state ratio conditions in the Rho ca experiments.

We used the 929 parameter sets from the plateau in Fig. B.2 to evaluate the model fit and to have an estimate of the resulting uncertainty due to non-identifiable parameters. In addition to the dynamic response of the system to nocodazole treatment with and without PKD inhibitors, which are shown in Fig. 3.2C in the main manuscript, Fig. B.5 shows the model fit after treatment with PDBu with and without inhibitor and steady state fold changes in PKD and DLC1 phosphorylation in the Rho ca experiments. Overall, the response of the system is very well described. Experimental data and model trajectories agree well for pPKD and pDLC1 in the control case and in case cells were treated with the inhibitor kb-NB prior to PDBu treatment. Also the fold changes in pPKD and pDLC1 induced by constitutively active Rho are well captured by our model, though the variance in the data is quite large here, especially for pDLC1.

B.2.4. Model validation via bootstrapping

Plausibility of the model was tested with a parametric bootstrapping approach (see e.g. [54]), in which we generated many datasets from the inferred stochastic model, which were subsequently used to estimate a distribution of likelihood function values J (Fig. 3.2D). For this purpose, we resampled D_i, $i = 1, \ldots, 10000$ datasets that mimic experimental data used in the study (i.e. same number of replicates, same conditions etc.). Then we calculated $p(D_i|\hat{\theta})$ with σ_{ij} from $\hat{\theta}$ and used these values to estimate a probability density $p(J_{\text{opt}}) = p(D_i|\hat{\theta})$ via kernel density estimation, which was compared to $p(y|\hat{\theta})$, the likelihood value for the real experimental data.

B.2.5. Profile likelihood analysis

The profile likelihood shown in Fig. 3B was obtained by setting the feedback parameter θ_6 to the indicated value and re-optimizing all other parameters. For computational efficiency, we did not initialize all parameters from scratch in each of these optimization runs, but used the results from the previous run as a starting point for the next run.

B.3. Negative feedback model (model 2): Data pre-processing

B.3.1. Normalization of experimental data

For model calibration we used the normalization from Table B.1. Additional experiments were normalized analogously, which is indicated in Table B.5.

B.3.2. Selection of an error model

Analogous to model 1 we compared 12 different error models for the model 2 (Table B.6). As before, AIC_c, BIC and Akaike weights agree on the choice of the best model, which is also with the additional

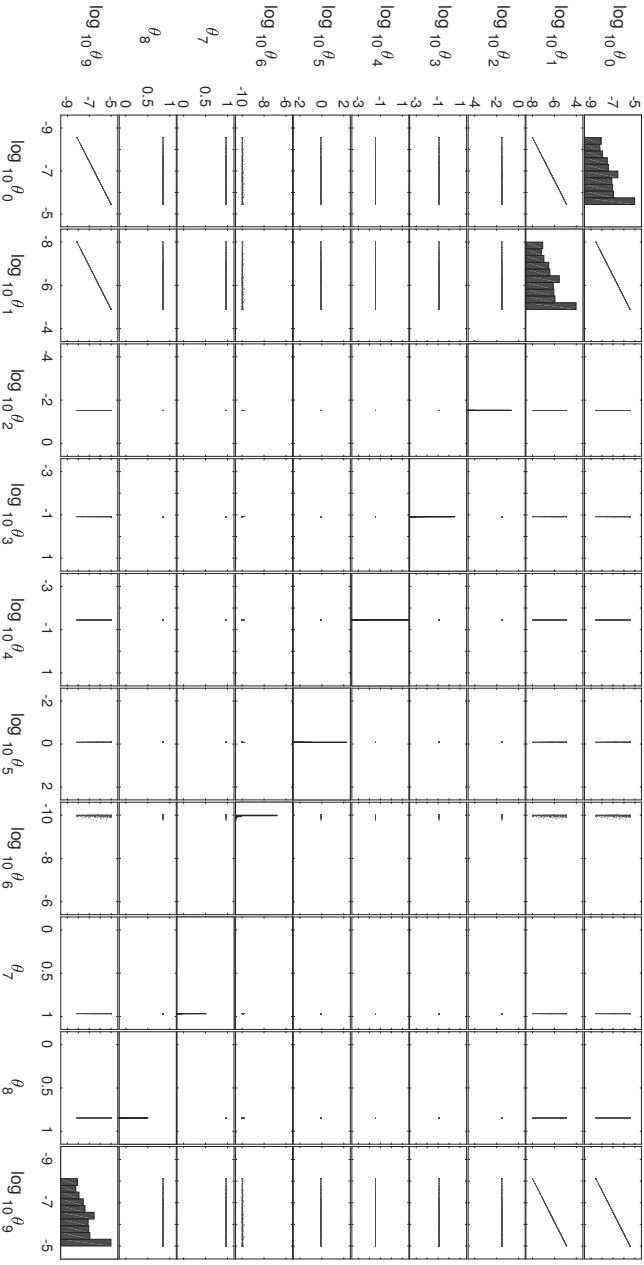

log10θ0
log10θ1
log10θ2
log10θ3
log10θ4
log10θ5
log10θ6
θ7
θ8
log10θ9

Figure B.3.: Scatterplot matrix for 929 parameter sets extracted from Figure B.2.

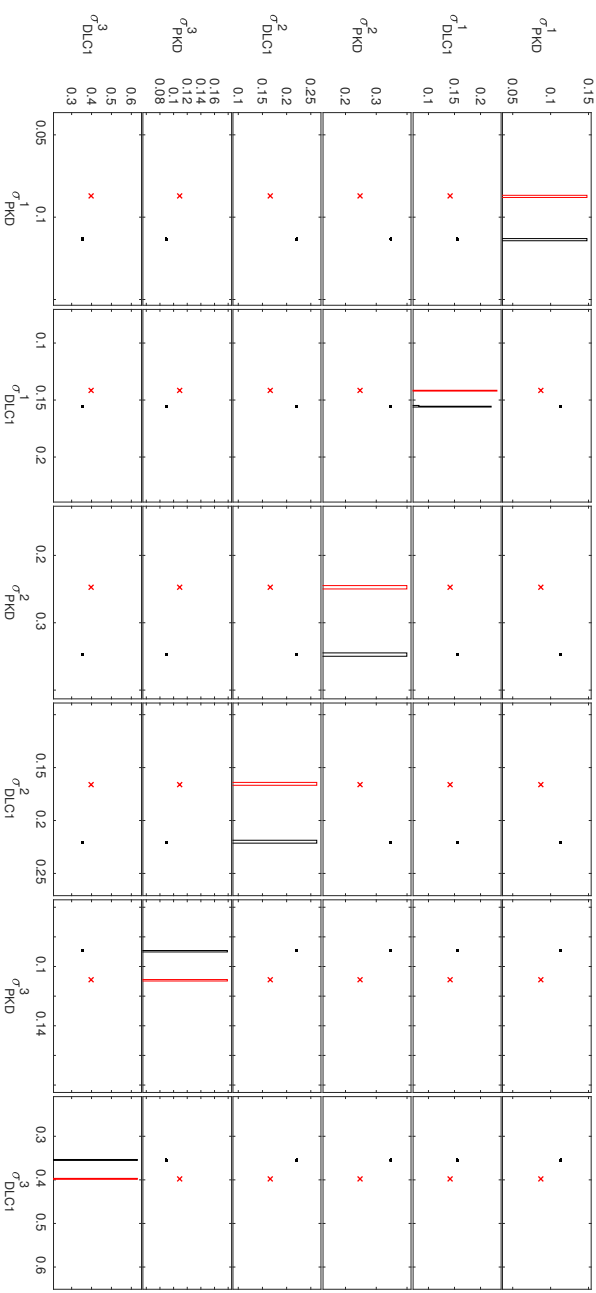

σ^1_{PKD}
σ^1_{DLC1}
σ^2_{PKD}
σ^2_{DLC1}
σ^3_{PKD}
σ^3_{DLC1}

Figure B.4.: Scatterplot matrix of 929 standard deviations σ (black) in comparison with unbiased empirical estimates (red).

Experiment 8: -Dox (Fig. 3.4A)			
			pPKD
	0.38	1.08	0.89
Experiment 9: +Dox (Fig. 3.4A)			
			pPKD
	1	1	1
Experiment 10: Input nocodazole, siNT (Fig. 3.4B)			
time			pPKD
0 min	0.28	0.46	0.50
10 min	0.76	0.54	0.55
30 min	1	1	1
Experiment 11: Input nocodazole, siDLC1 (Fig. 3.4B)			
time			pPKD
0 min	0.09	0.13	0.17
10 min	0.45	0.37	0.54
30 min	0.72	0.53	0.82
Experiment 12: siNT, -H1152 (Fig. 3.4D)			
			pPKD
	1	1	1
Experiment 13: siDLC1, -H1152 (Fig. 3.4D)			
			pPKD
	0.26	0.53	0.37
Experiment 14: siDLC1, +H1152 (Fig. 3.4D)			
			pPKD
	0.76	0.71	0.45

Table B.5.: Experimental data, normalized to the highest signal value

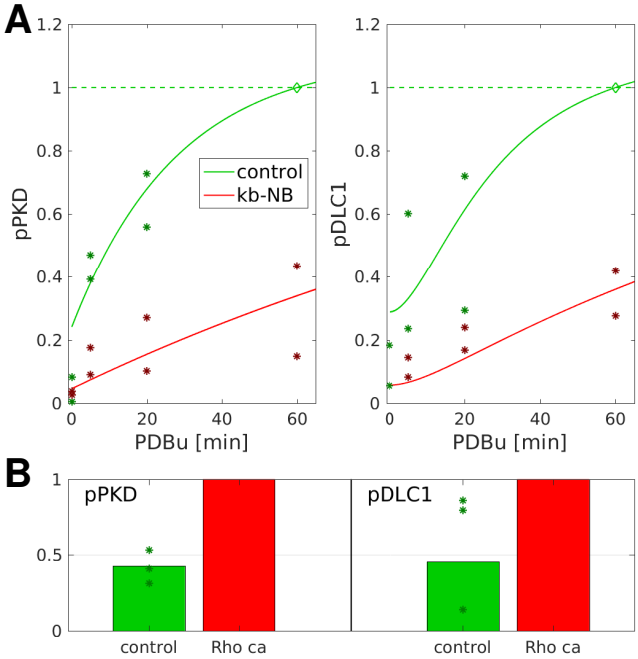


Figure B.5.: Fit for model 1. Dots indicate re-normalized experimental data from Fig. B.1, with normalization points indicated by diamonds, together with 929 estimated model trajectories that lie all on top of each other.

experiments a normally distributed error model with standard deviations pooled over experiments and outputs. This results in 9 different standard deviations, six of which are the same as in Table B.3, the other three ones are listed in Table B.7.

B.4. Model 2: Modeling and model calibration

In the following we provide details on the negative feedback modeling approach and calibration to experimental data.

Error model	# parameters	$\mathcal{L}$	AIC	AIC_c	BIC	AW
$\mathcal{N}$, unbiased, σ_i	70	3.7e33	-14.61	286.60	170.49	2.2e-52
$\mathcal{N}$, unbiased, $\bar{\sigma}_{17}$	52	7.7e27	-24.42	83.65	113.08	2.6e-08
$\mathcal{N}$, unbiased, $\bar{\sigma}_9$	44	8.3e22	-17.56	49.56	98.79	0.6560
$\mathcal{N}$, biased, σ_i	70	3.4e35	-23.61	277.61	161.50	2.0e-50
$\mathcal{N}$, biased, $\bar{\sigma}_{17}$	52	4.7e28	-28.05	80.03	109.46	1.6e-07
$\mathcal{N}$, biased, $\bar{\sigma}_9$	44	4.3e22	-16.25	50.87	100.10	0.3407
log $\mathcal{N}$, unbiased, σ_i	70	1.6e34	-17.56	283.65	167.55	9.6e-52
log $\mathcal{N}$, unbiased, $\bar{\sigma}_{17}$	52	2.1e26	-17.26	90.82	120.25	7.2e-10
log $\mathcal{N}$, unbiased, $\bar{\sigma}_9$	44	3.7e20	-6.72	60.40	109.63	0.0029
log $\mathcal{N}$, biased, σ_i	70	1.5e36	-26.55	274.66	158.55	8.6e-50
log $\mathcal{N}$, biased, $\bar{\sigma}_{17}$	52	4.9e26	-18.92	89.16	118.59	1.7e-09
log $\mathcal{N}$, biased, $\bar{\sigma}_9$	44	4.0e19	-2.26	64.86	114.09	0.0003

Table B.6.: Error model selection procedure based on information criteria. In total $N = 104$ measurements were used. Standard deviations were either estimated individually for each condition (σ_i), or by pooling over time series and outputs ($\bar{\sigma}_{17}$) or over experiments and outputs ($\bar{\sigma}_9$). Color encoding was chosen equivalent to Table B.2.

variance parameter	pooled experiments (state variable)
σ^4_{PKD}	8 (pPKD)
σ^5_{PKD}	10,11 (pPKD)
σ^6_{PKD}	13,14 (pPKD)

Table B.7.: Description of the three additional variance pools used for the modeling study, depending on the experimental conditions given in Table B.5.

B.4.1. Modeling approach and normalization

Similar to model 1, we eliminated PKD and DLC1 by assuming mass conservation of respective total amounts

$$\text{PKD}_{\text{tot}} = \text{PKD} + \text{pPKD} \tag{B.4.1a}$$

$$\text{DLC1}_{\text{tot}}(1 + \alpha_5\theta_{12} - \alpha_6\theta_{13}) = \text{DLC1} + \text{pDLC1}. \tag{B.4.1b}$$

Normalization of both state variables to total concentrations,

$$x_1 = \frac{\text{pDLC1}}{\text{DLC1}_{\text{tot}}(1 + \alpha_5\theta_{12} - \alpha_6\theta_{13})} \tag{B.4.2a}$$

$$x_2 = \frac{\text{pPKD}}{\text{PKD}_{\text{tot}}} \tag{B.4.2b}$$

leads to

$$\begin{aligned}\dot{x}_1 = &(\theta_0(1 + \tilde{\theta}_6(1 + \alpha_5\theta_{12} - \alpha_6\theta_{13})(1 - x_2)(1 - \alpha_4\theta_{10})(1 + \alpha_4\theta_{11})) \\ &+ \theta_4 u_1 + \theta_5 u_2 + \alpha_3\theta_9)(1 - \alpha_1\theta_7 - \alpha_2\theta_8)(1 - x_1) - \theta_1 x_1\end{aligned} \tag{B.4.3a}$$

$$\dot{x}_2 = \tilde{\theta}_2(1 - x_2)x_1 - \theta_3 x_2 \tag{B.4.3b}$$

where $\tilde{\theta}_2 = \theta_2\text{DLC1}_{\text{tot}}$ and $\tilde{\theta}_6 = \theta_6\text{DLC1}_{\text{tot}}$. Analogously to the first part, $\tilde{\theta}_2$ and $\tilde{\theta}_6$ will be called θ_2 and θ_6 in the following.

Changes in the DLC1 mediated PKD phosphorylation rate and the total DLC1 amount compared to model 1 are listed in Table B.8. We note here that all experiments were conducted with doxycycline

for model 1, while this is not the case for some of the additional experiments, such that we introduced an additional parameter to describe the effect of doxycycline addition on DLC1 total amounts.

B.4.2. Likelihood function

As before, the likelihood function for model 2 is given by

$$\begin{aligned}\mathcal{L}(\theta) &= p(y|\theta) \\ &= \prod_{i=1}^{14}\prod_{j=1}^{2}\prod_{t_k}\prod_{l=1}^{4}\frac{1}{\sigma_{ij}\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{y_{ijl}(t_k)-z_{ij}(\theta,t_k)}{\sigma_{ij}}\right)^2\right]\end{aligned} \tag{B.4.4}$$

with unknown parameter vector

$$\boldsymbol{\theta} = (\theta_0,\ldots,\theta_{13},\,\sigma^1_{\mathrm{PKD}},\ldots,\sigma^6_{\mathrm{PKD}},\sigma^1_{\mathrm{DLC1}},\ldots,\sigma^3_{\mathrm{DLC1}}) \in \mathbb{R}^{23}.$$

B.4.3. Optimization details

For optimization we used again `Matlab 2016b` (64 bit) and the toolboxes `SBPD` and `SBTOOLBOX2` and the internal `Matlab` solver `ode15i` for integration with options set to `options.abstol = 1e-10` and `options.reltol = 1e-10`. Optimization was performed with the gradient-based optimizer `fmincon` with the interior point method and settings `OptionsFmincon.TolFun = 1e-6`, `OptionsFmincon.TolCon = 1e-6` and `OptionsFmincon.MaxFunEvals = 5000`.

As for model 1 we used a logarithmic scale for the parameters $\theta_0,\ldots,\theta_{13}$. The final boundaries of four orders of magnitude for these parameters were set after several optimization runs in order to avoid results on the boundaries. Following (B.4.3) the boundaries for θ_7, θ_8, θ_{10} and θ_{13}, the efficiencies of the two PKD inhibitors, Rho ca in the feedback and the silencing RNA for the total DLC1, were set to $[0,1]$ without using a log-scale. Boundaries for the standard deviations σ^i_j were set around the empirical estimates, $\left[\frac{1}{2}\sigma_{\mathrm{emp}},\,2\sigma_{\mathrm{emp}}\right]$. Optimization was performed with 1000 starting values from a latin hypercube sample. For all experiments the initial conditions $x_1(0)$ and

Treatment (u, α)	Effect on $k(\text{DLC1}, \theta)$ or DLC1	Remark
–	$k = \begin{cases} \theta_0(1 + \theta_6 \text{DLC1}) & \theta_6 \text{DLC1} \in (0, 1] \\ 0 & \text{otherwise} \end{cases}$	unphosphorylated DLC1 increases the basal PKD phosphorylation rate
doxycycline	$\text{DLC1} + \text{pDLC1} = \text{DLC1}_{\text{tot}}(1 + \theta_{12})$	doxycycline induces DLC1 expression and thus increases DLC1_{tot}
siDLC1	$\text{DLC1} + \text{pDLC1} = \text{DLC1}_{\text{tot}}(1 - \theta_{13})$	siDLC1 reduces DLC1 expression
Rho ca	$k + \theta_9$	constitutively active Rho increases PKD phosphorylation
	$\theta_0(1 + \theta_6 \text{DLC1}(1 - \theta_{10}))$	feedback via Rock is negative
H1152	$\theta_0(1 + \theta_6 \text{DLC1}(1 + \theta_{11}))$	H1152 inhibits Rock activity and thus increases the overall PKD phosphorylation rate

Table B.8.: Effect of different experimental treatments on the PKD phosphorylation rate $k(\text{DLC1}, \theta)$ and on total DLC1 amounts

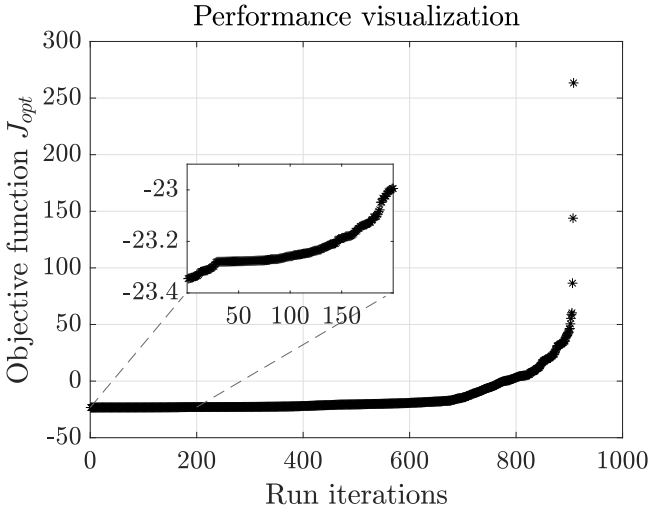


Figure B.6.: Sorted objective function values of converged parameter sets for model 2

$x_2(0)$ were assumed to be in a steady state and calculated accordingly. 908 starting points converged according to the step size criterion[2] for the parameter vector θ. As before, we evaluated the results according to the sorted J_{opt} values and took the first 150 values as a reasonable estimate for the global optimum, see Fig. B.6.

A comparison of experimental data and model fits which are not shown in the main manuscript can be seen in Fig. B.7. Shown are the time courses of pPKD and pDLC1 after stimulation with PDBu in the control case and with the PKD inhibitor kb-NB as well as the fold change in both variables induced by constitutively active Rho. The dynamic response of both variables is well captured in the control experiments, while the effect of the PKD inhibitor is underestimated, which is the opposite to the time series experiments after stimulation with nocodazole (Fig. 3.5) and thus a compromise model fit with

[2]Ensuring that the change in the parameter vector is less than the specified step size tolerance and constraint violation is less than those specified in `options.ConstraintTolerance`

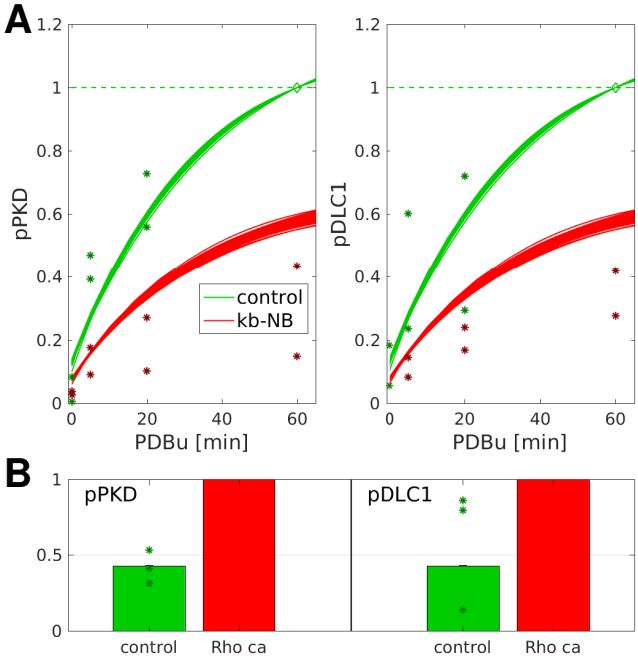


Figure B.7.: Fit for model 2. Dots indicate re-normalized experimental data from Figs. 3.1 and B.1, with normalization points indicated by diamonds, together with 150 estimated model trajectories.

respect to the efficiency of the inhibitor. The effect of Rho ca on both output variables is well captured. Compared to the model fit of model 1 trajectories show a larger variability.

Fig. B.8 shows the parameter scatterplots for model 2. Compared to that of model 1 parameters are more spread and less correlated. The parameters $\theta_1, \theta_7, \theta_8$ and θ_{12} are well identifiable. These are the PKD dephosphorylation rate, the effects of the two PKD inhibitors and the influence of doxycycline addition onto DLC1 total amounts. In contrast, the parameters θ_2 and θ_3, phosphorylation and dephosphorylation rates of DLC1, have a broad distribution. Some of the parameters also show correlations such as θ_4, θ_5 and θ_9, which are the

influence of PDBu, nocodazole and RhoA on PKD phosphorylation.

Fig. B.9 shows that all standard deviations are well identifiable within the given bounds about the empirical values.

B.4.4. Model prediction

It is known that strong negative feedback can cause robustness of activity states to variations in total protein amounts for proteins in the feedback loop [62]. However, in our case it is unclear how to define the strength of the feedback loop. One possibility is to investigate the strengths of the individual links between the two components in our model. Thus, we simulated pDLC1 and pPKD fold changes implied by variations in PKD and DLC1 total amounts, respectively (top row in Fig. B.10). pDLC1 increases in a perfect linear way with the PKD amount due to mass action kinetics for PKD mediated phosphorylation of DLC1. Similarly, pPKD is a linear function of the DLC1 amount, with an offset that corresponds to the basal and DLC1 independent PKD phosphorylation rate. Together, these results suggest that pPKD is highly influenced by DLC1 and vice versa. However, the conclusion that the overall feedback makes the system robust is not true in this case, as can be seen in Fig. B.10 (bottom row). Using mass action kinetics for a single molecule that is reversibly phosphorylated and not subject to feedback regulation, the phosphorylated protein amount is a linear function of the total amount. At the PKD level the system behaves exactly as such a decoupled system (left). This is different for DLC1 phosphorylation, were the effect of the negative feedback is visible in the deviation from the diagonal line (right). Thus, although we have strong influences between both output variables, this does not result in an overall strong effect of the negative feedback. An explanation for this paradigm is shown in Fig. B.10. When pPKD and pDLC1 fractions are small, fold changes in pDLC1 implied by variations in PKD amounts might be large, but at the same time, the fold change in unphosphorylated DLC1, which feeds back to PKD, is so small that the effect of changes in PKD amounts are not propagated by the feedback.

B.5. Additional experiments

Fig. B.11 shows that DLC1 depletion by independent siRNAs reduces PKD activation in HEK293T cells and endogenous DLC1 localizes to focal adhesions in U20S cells.

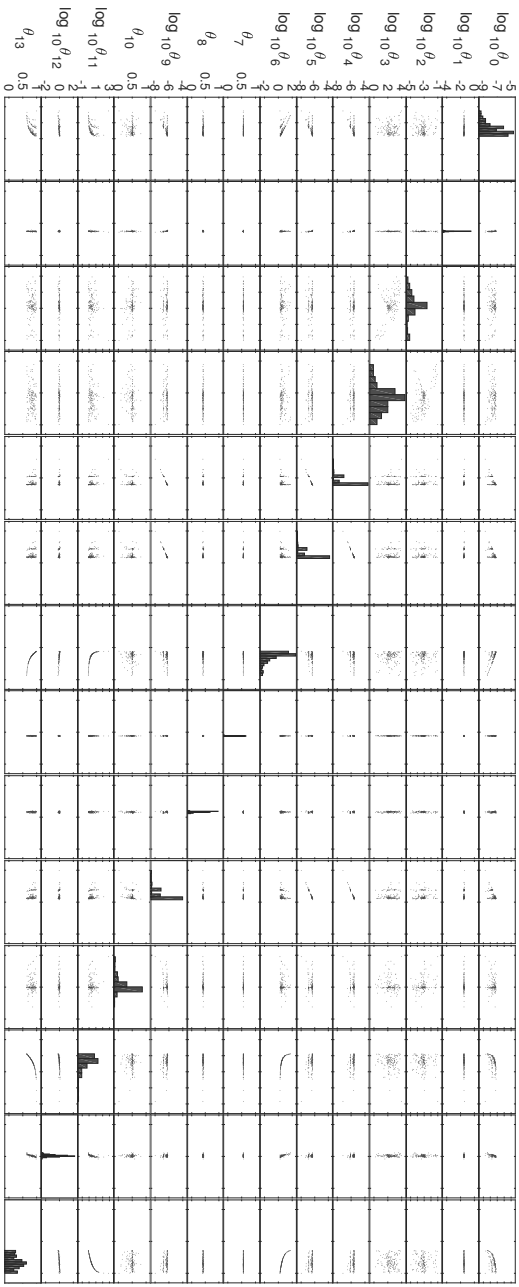

log10θ0
log10θ1
log10θ2
log10θ3
log10θ4
log10θ5
log10θ6
θ7
θ8
log10θ9
θ10
log10θ11
log10θ12
θ13

Figure B.8.: Scatterplot matrix for 929 parameter sets extracted from Fig. B.6 of model 2.

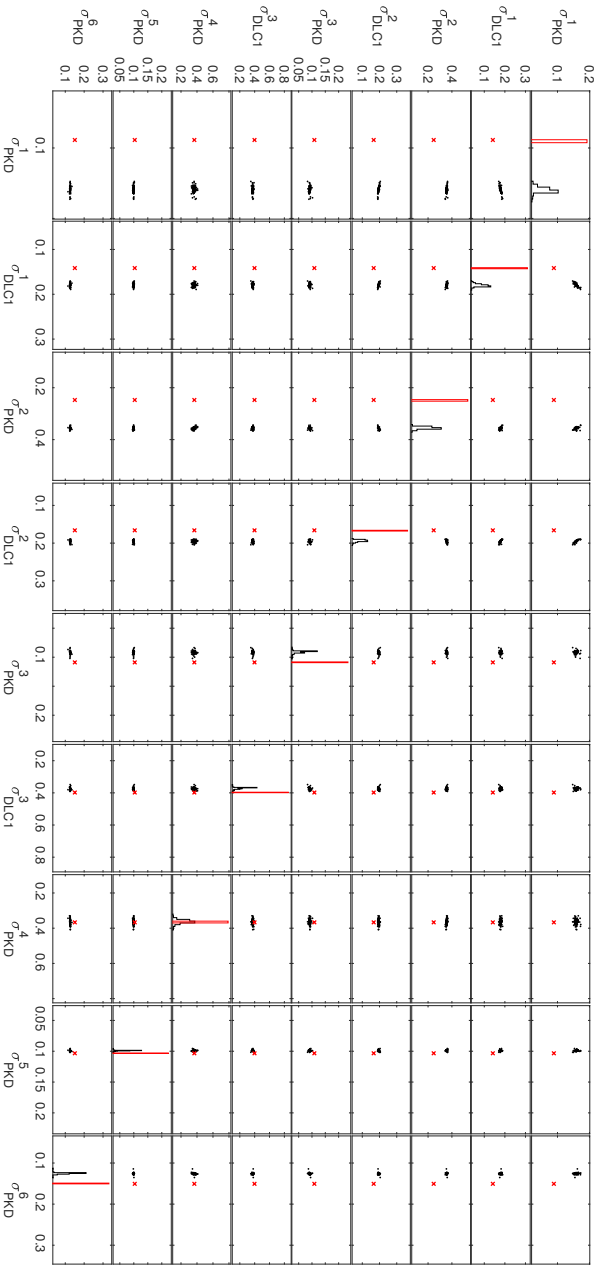

σ^1_{PKD}
σ^1_{DLC1}
σ^2_{PKD}
σ^2_{DLC1}
σ^3_{PKD}
σ^3_{DLC1}
σ^4_{PKD}
σ^5_{PKD}
σ^6_{PKD}

Figure B.9.: Scatterplot matrix of 150 estimated standard deviations σ_j^i (black) of model hypothesis 2 in comparison to unbiased empirical estimates σ_{emp} (red).

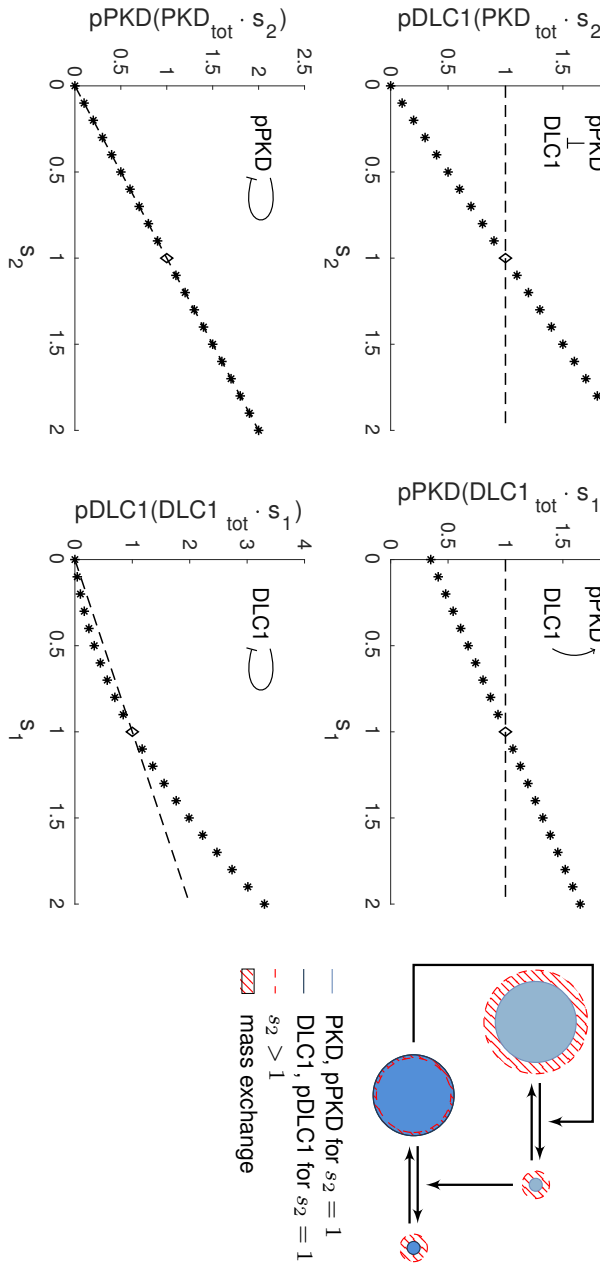

pDLC1(PKD tot · s2)
pPKD ⊥ DLC1
s2
pPKD(PKD tot · s2)
pPKD
s2
pPKD(DLC1 tot · s1)
pPKD ⊤ DLC1
s1
pDLC1(DLC1 tot · s1)
DLC1
s1
PKD, pPKD for s2 = 1
DLC1, pDLC1 for s2 = 1
s2 > 1
mass exchange

Figure B.10.: Interaction strengths between variables in the feedback loop. Normalization to nominal value denoted by diamonds. Top: Mutual influences between pPKD and pDLC1 are strong. Bottom: This does not result in a strong overall feedback and robustness of phosphorylated amounts against variations in respective total amounts. An explanation for this putative contradiction are small fractions of phosphorylated amounts.

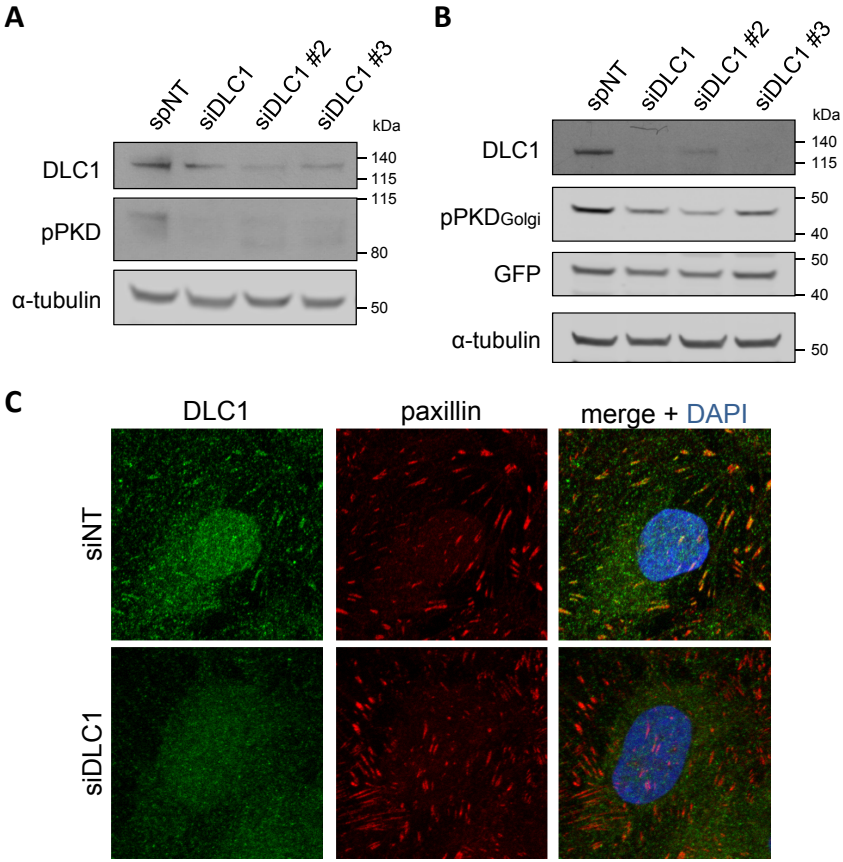

A
spNT
siDLC1
siDLC1 #2
siDLC1 #3
kDa
DLC1
140
115
pPKD
115
80
α-tubulin
50
B
spNT
siDLC1
siDLC1 #2
siDLC1 #3
kDa
DLC1
140
115
pPKDGolgi
50
40
GFP
50
40
α-tubulin
50
C
DLC1
paxillin
merge + DAPI
siNT
siDLC1

Figure B.11.: (A) HEK293T cells were transfected with the indicated siRNAs. After 3 days, cells were lysed and lysates were analyzed by immunoblotting. (B) Two days after transfection with the indicated siRNAs, HEK293T cells were transfected with the construct encoding the Golgi PKD activity reporter. The next day, cells were lysed and lysates analyzed by immunoblotting. (C) U2OS cells were transfected with the indicated siRNAs. After 3 days, cells were fixed and stained with DLC1 and paxillin specific antibodies, followed by fluorescently labeled secondary antibodies. Nuclei were counterstained with DAPI. The images shown are representative maximum intensity projections of several confocal sections.

C. Supporting Material for Chapter 4

C.1. Classification methods

For a short formal description of each method, let the data consist of a sample set $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, the predictor, with n samples $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})$, $i = 1, \ldots, n$ and p features as well as the response vector $\boldsymbol{y}$, a binary vector of length n that encodes the class membership.

C.1.1. Sparse robust discriminant analysis with sparse partial robust M regression (SPRM-DA)

Sparse robust discriminant analysis with sparse partial robust M regression [98] (SPRM-DA, in the following SPRM) classifies samples $\boldsymbol{x}_i$ by maximizing the distance between group means and minimizing the variance within groups on a projected hyperplane. Therefore, the response vector $\boldsymbol{y}$ is given in centered and scaled form and further treated as continuous variable. Likewise, the matrix $\boldsymbol{X}$ is column-wise centered. SPRM consists of two parts. In the first part, based on partial least squares [244], the data $\boldsymbol{X}$ is reduced by projecting it to a lower dimensional subspace $\mathbf{XW} \in \mathbb{R}^{n \times H}$, $H < p$, where $\mathbf{W} \in \mathbb{R}^{p \times H}$ is characterized by direction vectors $\boldsymbol{w}_h \in \mathbb{R}^p$, $h = 1, \ldots, H$ and $H < p$ defines the reduced dimensionality of the subspace. This is achieved by maximizing the squared covariance between the predictor projection $\boldsymbol{Xw}$ and the response vector $\boldsymbol{y}$,

$$\boldsymbol{w}_h = \arg\max_{\boldsymbol{w}} \operatorname{cov}^2(\boldsymbol{Xw}, \boldsymbol{y}) \tag{C.1.1}$$

for $h \in \{1, \ldots, H\}$ subject to $\|\boldsymbol{w}_h\| = 1$ and $\boldsymbol{w}_h^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w}_i = 0$ for $1 \leq i < h$. Many algorithms have been suggested to solve problem (C.1.1). The one that is used in [98] makes use of the standard covariance estimator

$$\widehat{\mathrm{cov}}^2 = \left(\frac{1}{n-1} \boldsymbol{y}^T \boldsymbol{X} \boldsymbol{w} \right)^2 . \tag{C.1.2}$$

In a second step, the transformed data is classified using Fisher's linear discriminant analysis (LDA).

Robust DA with partial robust M regression uses the concept of M estimation as a powerful tool in robust statistics to identify outliers. Therefore, weights between 0 and 1 are assigned to each sample to regulate its influence on model estimation. Weights are chosen such that samples $\boldsymbol{x}_i$ with large distances with respect to the center and covariance of its assigned class, quantified by a weighting function on the robust squared Mahalanobis distance, are downweighted. These weights enter both steps of the procedure, i.e., covariance maximization, where the weights are determined iteratively, and LDA, where the optimized weights are used to perform a weighted, robust LDA. The procedure how optimal weights are achieved is explained in [98]. Integration of these weights modifies the optimization problem (C.1.1) to

$$\hat{\boldsymbol{w}}_h = \arg\max_{\boldsymbol{w}} \mathrm{cov}^2(\boldsymbol{X}_{\boldsymbol{\Omega}} \boldsymbol{w}, \boldsymbol{y}_{\boldsymbol{\Omega}}), \tag{C.1.3}$$

where $\boldsymbol{\Omega} = \mathrm{diag}(\omega_1, \ldots, \omega_n)$ downweights samples, leading to weighted data matrix and response vector $\boldsymbol{X}_{\boldsymbol{\Omega}} = \boldsymbol{\Omega} \boldsymbol{X}$ and $\boldsymbol{y}_{\boldsymbol{\Omega}} = \boldsymbol{\Omega} \boldsymbol{y}$, respectively. Constraints for $\boldsymbol{w}_h$ apply accordingly.

In addition, sparsity is ensured by penalizing the estimation of the direction vectors $\boldsymbol{w}_{\boldsymbol{h}}$ with an ℓ_1 norm penalty η. This regularization forces complete rows of the weight matrix $\boldsymbol{W}$ to become zero, and the respective features have no influence. Thus, only features with nonzero weights are selected. The resulting optimization problem is described in [98] (equations (18)).

Hyperparameters of SPRM include the number H of latent components and the sparsity parameter η, which are determined by a

cross-validation procedure as described in [98].

We employ the sample weights, as provided by the classification output, as a measure for outlier ranking in the outlier ranking step in Figure 4.1A in the main manuscript. Samples are thus ranked in ascending order of their weights. The selected features are extracted as the set of features that have at least one non-zero entry in the corresponding row of the weight matrix $\boldsymbol{W}$.

C.1.2. Robust and sparse K-means clustering (RSK-means)

The second employed classifier is Robust and sparse K-means clustering (RSK-means) [132]. RSK-means is based on the standard K-means clustering [217], which searches for a partition of the dataset into K clusters by minimizing the within-cluster sum of squares or, equivalently, maximizing the between-cluster sum of squares. If a large fraction of features is not related to the response variables and only few features contribute to the differences between samples in different clusters, K-means often fails. This problem was first addressed by Witten and Tibshirani [243], who proposed sparse K-means to simultaneously find clusters and a small number of features which are sufficient to unravel the cluster structure. This is achieved by assigning weights $\boldsymbol{w} = (w_1, \ldots, w_p), w_j \geq 0,\ j = 1, \ldots, p$ to each feature that are constraint in their norms to enforce sparsity, leading to the optimization problem

$$\max_{C_1,\ldots,C_k,\boldsymbol{w}} \sum_{j=1}^{p} w_j \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{i,i',j} \right\} \qquad \text{(C.1.4)}$$

subject to $\|\boldsymbol{w}\|_2 \leq 1$ and $\|\boldsymbol{w}\|_1 \leq l$. Here, $l > 1$ determines the degree of sparsity in terms of non-zero weights of the solution, and $C_1, \ldots, C_K$ denotes the partition into K different clusters. In equation (C.1.4), the expression in the brackets describes the between-cluster sum of squares, with additive dissimilarity measure $d_{i,i'} = \sum_{j=1}^{p} d_{i,i',j}$ between samples i and i', which can, e.g., be chosen as squared Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i'}$, i.e., $d_{i,i',j} = (x_{i,j} - x_{i',j})^2$.

The variable n_k denotes the number of individuals in cluster k.

Like K-means clustering, problem (C.1.4) is in practice solved by iterating the following steps:

1. Given weights $\boldsymbol{w}$ and cluster centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$, assign samples to the cluster with the closest center in terms of weighted Euclidean squared distances.

2. Based on this partitioning, update cluster centers to the weighted sample means of the samples in the respective clusters.

3. Choose weights subject to constraints such that the weighted between cluster sum of squares is maximized.

Following the idea of Cuesta–Albertos and Gordaliza [47] to achieve a clustering that is robust to outliers by trimming α 100% of the samples with the largest distances to their cluster centers in step 2, Kondo et al. [132] introduced a modified algorithm, RSK-means, which combines SK-means with a trimming procedure which finally returns a set of selected features as well as a set of outliers. Thereby, the final set of outliers O is obtained as the union of the sets O_W and O_E, which are calculated with and without weights, respectively. Optimal weights $\boldsymbol{w}$ are determined by maximizing the between-cluster sum of squares under exclusion of observations flagged as outliers in the set O,

$$\max_{\substack{\|\boldsymbol{w}\|_2 \leq 1, \\ \|\boldsymbol{w}\|_1 \leq l}} \sum_{j=1}^{p} w_j \left[\frac{1}{n-|O|} \sum_{i=1}^{n-|O|} \sum_{i'=1}^{n-|O|} d_{i,i',j} - \sum_{k=1}^{K} \frac{1}{n_{k,O}} \sum_{i,i' \in C_{k,O}} d_{i,i',j} \right]. \tag{C.1.5}$$

Here, $C_{k,O}$ represents the truncated k-th cluster and $n_{k,O}$ the corresponding number of samples.

RSK-means requires the selection of three hyperparameters, the L_1 bound l, which determines the degree of sparsity and can be chosen to achieve a desired number of selected features, the trimming proportion α, and the number of clusters K. In order to select l and α, classification runs for different combinations of parameter values

were performed. Final parameters were selected as best combination with respect to the classification error rate as provided by the CER function from the RSKC [131] package. Ranges for each parameter are given in Table C.1. We defined $K = 2$ in accordance to the binary response vector $\mathbf{y}$ supplied in the classification process.

For the outlier ranking step in Figure 1A in the main manuscript, we calculate the Euclidean distance of cases from the cluster center using the cluster partition obtained by the classifier. Cluster centers are determined without identified outliers O and including feature weights

$$D_i(\boldsymbol{x}_i) = \sum_{j=1}^{p} w_j (x_{i,j} - \mu_{k,j})^2, \tag{C.1.6}$$

for $i \in C_k$ and $\mu_{k,j} = \dfrac{1}{n_{k,O}} \sum_{i \in C_{k,O}} x_{i,j}$, $k \in \{1, 2\}$. Since larger distances from the cluster center correspond to a higher chance of being an outlier, ranking is assigned in descending order of distance D_i. Furthermore, the features weights w_j, $j = 1, \ldots, p$, are evaluated. Features with corresponding non-zero weights constitute the set of selected features.

C.1.3. Robust and sparse logistic regression with elastic net penalty (enetLTS)

Robust and sparse logistic regression with elastic net penalty (enetLTS) [137] uses a logistic regression model to determine a regression hyperplane between the groups. Therefore, the logistic regression model $y_i = \pi_i + \varepsilon_i$, for $i = 1, \ldots, n$, is used to describe the relation between the predictor $\boldsymbol{X}$ and the response $\boldsymbol{y}$. The term ε_i describes a binomially distributed error, and π_i denotes the conditional probability for the i-th individual to belong to class one,

$$\pi_i = P(y_i = 1 | \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}, \tag{C.1.7}$$

with regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$. In case $n > p$ optimal regression coefficients $\hat{\boldsymbol{\beta}}$ are identified by minimizing a deviance function

$d(\boldsymbol{x}_i^T\boldsymbol{\beta}, y_i)$,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} d(\boldsymbol{x}_i^T\boldsymbol{\beta}, y_i) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} -y_i\boldsymbol{x}_i^T\beta + \log\left(1 + e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right). \tag{C.1.8}$$

The method is adjusted for multicollinearity among the predictors and cases of $n < p$ by adding an elastic net penalty term

$$P_\alpha(\boldsymbol{\beta}) = (1-\alpha)\frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}\left[(1-\alpha)\frac{1}{2}\beta_j^2 + \alpha|\beta_j|\right] \tag{C.1.9}$$

to equation (C.1.8),

$$\hat{\boldsymbol{\beta}}_{\text{enet}} = \arg\min_{\boldsymbol{\beta}}\left\{\sum_{i=1}^{n} d(\boldsymbol{x}_i^T\boldsymbol{\beta}, y_i) + \lambda P_\alpha(\boldsymbol{\beta})\right\}. \tag{C.1.10}$$

The tuning parameter $\lambda \geq 0$ determines the strength of the penalty and thus sparsity, and $\alpha \in [0, 1]$ defines the mixing proportion of the ℓ_1 and ℓ_2 norm. In addition, the method becomes robust against outliers by iteratively trimming the sample set to an optimal subset

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{enetLTS}} &= \arg\min_{\boldsymbol{\beta}, H} Q(H, \boldsymbol{\beta}) \\ &= \arg\min_{\boldsymbol{\beta}, H}\left\{\sum_{i\in H} d(\boldsymbol{x}_i^T\boldsymbol{\beta}, y_i) + h\lambda P_\alpha(\boldsymbol{\beta})\right\},\end{aligned} \tag{C.1.11}$$

where $H \subseteq \{1, 2, \ldots, n\}$ with $|H| = h$. This subset is supposed to be outlier free, and hence all individuals that are not contained in the subset are defined as outliers. Solving problem (C.1.11) is in general a difficult problem, which is solved in an iterative way. First, an optimal set $H_{opt} = \arg\min_{H\subseteq\{1,\ldots,n\},|H|=h} Q(H, \hat{\boldsymbol{\beta}}_H)$ is found as explained in [137]. Then, optimal regression parameters $\hat{\boldsymbol{\beta}}_{\text{enetLTS}}$ are found via optimizing the objective function $Q(H_{opt}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$.

Altogether, enetLTS requires the selection of three hyperparameters. These consist of α, which describes the mixing proportion of the

two penalty terms in equation (C.1.9), λ defining the strength of the penalty and thus the degree of sparsity, as well as the subset proportion h_p. For the optimization, the original set of n samples is reduced by the proportion h_p, resulting in the trimmed set of size $h = h_p \cdot n$. Cross validation for different combinations of values of α, λ and h_p is performed. The range of values is presented in Table C.1.

In order to obtain an outlier ranking, we first calculate the absolute value of the Pearson residual r_i, $i \in \{1, \ldots, n\}$,

$$r_i = \left| \frac{y_i - \pi_i}{\sqrt{\pi_i(1-\pi_i)}} \right| \tag{C.1.12}$$

with π_i from Equation (C.1.7). Following the same reasoning as for RSK-means, we then rank from largest to smallest residual. Finally, the set of selected features is formed by all features with non-zero coefficient $\hat{\beta}_{\text{enetLTS},j}$, $j = 1, \ldots, p$.

C.2. Classification setup

All computations were performed on R version 3.3.3 [183].

We applied the three classification methods on the reduced dataset using default settings for enetLTS and RSK-means. However, for SPRM we specified scale = standard deviation, center = mean and fun = Fair as weighting function for the case weights.

Optimal parameters were chosen according to cross validation of a range of values for all parameters for enetLTS and SPRM, as implemented in the corresponding packages.

C.3. Simulation study

From the entire set of genes of the breast cancer dataset, a subset consisting of 3200 genes was selected as features. Since the simulation study was conducted after applying ROSIE to the breast cancer dataset, we decided to include the 54 commonly selected genes. Thus, the 54 commonly selected genes were chosen a priori and the remaining

Method	Parameter	Range for parameter selection
SPRM	α	$\{1, 2, \ldots, 5\}$
	η	$\{0.2, 0.3, \ldots, 0.9\}$
RSK-means	α	$\{0.05, 0.1, 0.15, 0.2\}$
	l	$\{15, 16, \ldots, 20\}$
enetLTS	α	$\{0.1, 0.2, \ldots, 0.8\}$
	λ	$\{0, 0.05, \ldots, 0.2\}$
	h_p	$\{0.7, 0.75, \ldots, 0.9\}$

Table C.1.: Parameter ranges for parameter selection

number was filled up randomly. Means and covariances were then computed for the resulting dataset for each the TNBC and non-TNBC groups. In order to achieve a similar class ratio as in the original dataset (160 TNBC vs. 859 non-TNBC, which corresponds to about 16% TNBC), we drew about 16% (31) samples from a multivariate normal distribution with mean and covariance matrix calculated for the TNBC group and about 84% (169) samples from a multivariate normal distribution with mean and covariance matrix calculated for the non-TNBC group. These datasets were corrupted by outliers as description in the main manuscript.

	SPRM	RSK-means	enetLTS
5% switched labels	$\alpha = 2$ $\eta = 0.5$	$\alpha = 0.05$ $l = 14$	$\alpha = 0.2$ $\lambda = 0.05$ $h_p = 0.85$
15% switched labels	$\alpha = 1$ $\eta = 0.7$	$\alpha = 0.05$ $l = 18$	$\alpha = 0.2$ $\lambda = 0.05$ $h_p = 0.7$
5% outliers in 15% of features	$\alpha = 1$ $\eta = 0.5$	$\alpha = 0.05$ $l = 14$	$\alpha = 0.1$ $\lambda = 0.05$ $h_p = 0.7$

Table C.2.: **Hyperparameters.** Optimal parameters found by cross-validation for the simulation study.

	SPRM	RSK-means	enetLTS
Parameters	$\alpha = 1$ $\eta = 0.4$	$\alpha = 0.05$ $l = 16$	$\alpha = 0.6$ $\lambda = 0.05$ $h_p = 0.75$

Table C.3.: **Hyperparameters.** Optimal parameters found by cross-validation for the breast cancer dataset.

C.4. Additional tables and figures

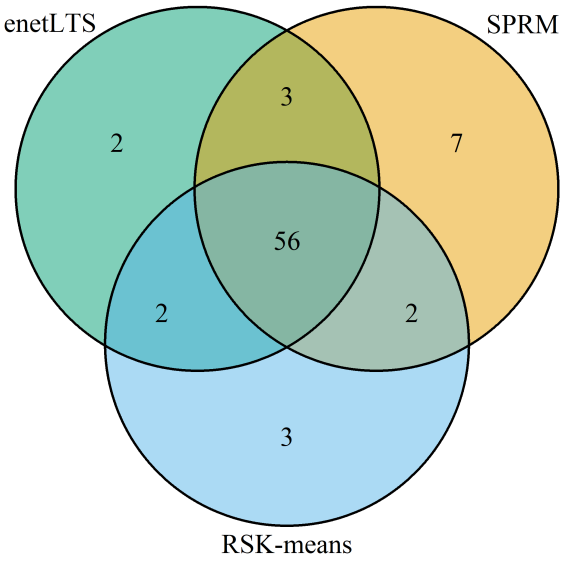


Figure C.1.: Venn diagram of samples misclassified by SPRM, RSK-means and enetLTS.

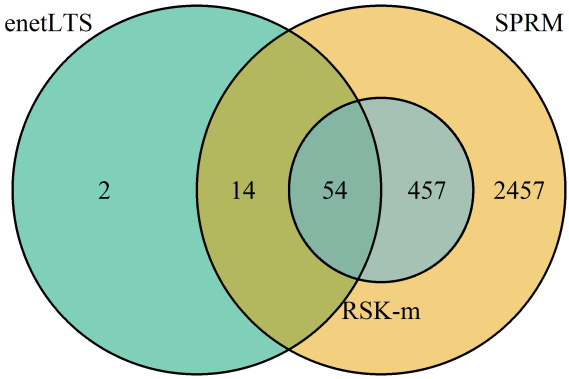


Figure C.2.: Venn diagram of the genes selected by SPRM, RSK-means and enetLTS.

5×	4×	3×	2×	1×	0×
FOXA1	***GATA3***	***CA12***	***TBC1D9***	*RPIA*	***TGFB3***
SPDEF	*DLX6*	***CAPN13***	***GALNT10***	*FABP7*	***AGR3***
MLPH	*SOX8*	*OTX1*	***CMBL***	*C16orf95*	*FAM136A*
CXXC5	*SOX6*	*GCNT2*	*STAC*		
AGR2	*CHRM3*	*PPP1R14C*	*MELTF*		
OCA2	*TMCC2*		*MICALL1*		
VGLL1	*A2ML1*		*TTLL4*		
ROPN1B	*UGT8*				
ROPN1	*CDCA2*				
FOXC1	*LEMD1*				
PAPSS1	*SMOC1*				
HORMAD1	*POU5F1*				
ZIC1	*SFT2D2*				
SRSF12	*NKX1-2*				
CHODL					
ART3					
EN1					
TTYH1					
COL9A3					
FAM19A3					
FZD9					
CT83					

Table C.4.: List of commonly selected genes sorted according to the number of common selections in bootstrap runs. Bold names represent genes at least partially downregulated in TNBC samples (smaller block of positively correlated genes in Figure 4.3 in the main manuscript), while remaining genes are at least partially upregulated (larger block of positively correlated genes in Figure 4.3 in the main manuscript).

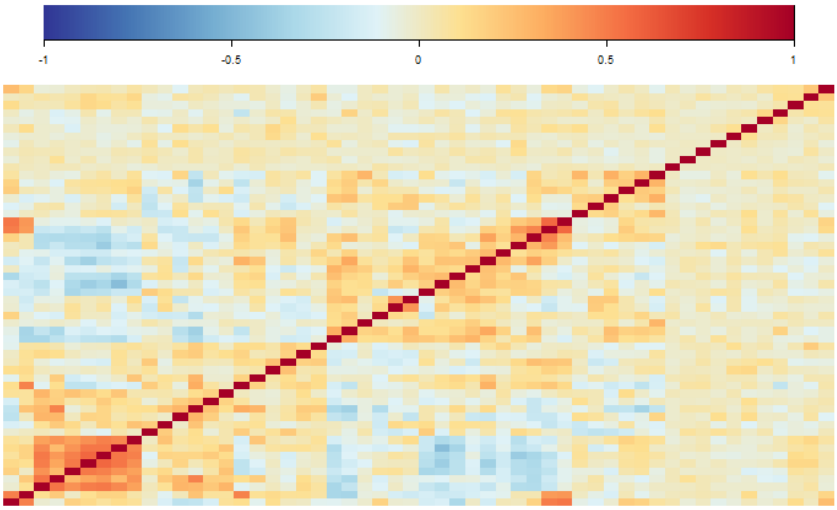


Figure C.3.: Heatmap of correlation values of 54 randomly selected features.

	SPRM	RSK-means	enetLTS	# of commonly selected features	# of influential samples
Block 1				71	13
# of selected genes	3206	505	210		
Misclassifications	67	65	50		
Block 2				83	11
# of selected genes	2902	519	164		
Misclassifications	50	57	41		
Block 3				81	13
# of selected genes	3193	511	187		
Misclassifications	62	61	169		
Block 4				82	13
# of selected genes	2857	522	222		
Misclassifications	68	67	160		
Block 5				82	11
# of selected genes	3095	517	211		
Misclassifications	60	61	44		

Table C.5.: Summary of classification results for bootstrapped data including number of selected features and number of misclassifications for SPRM, RSK-means and enetLTS.

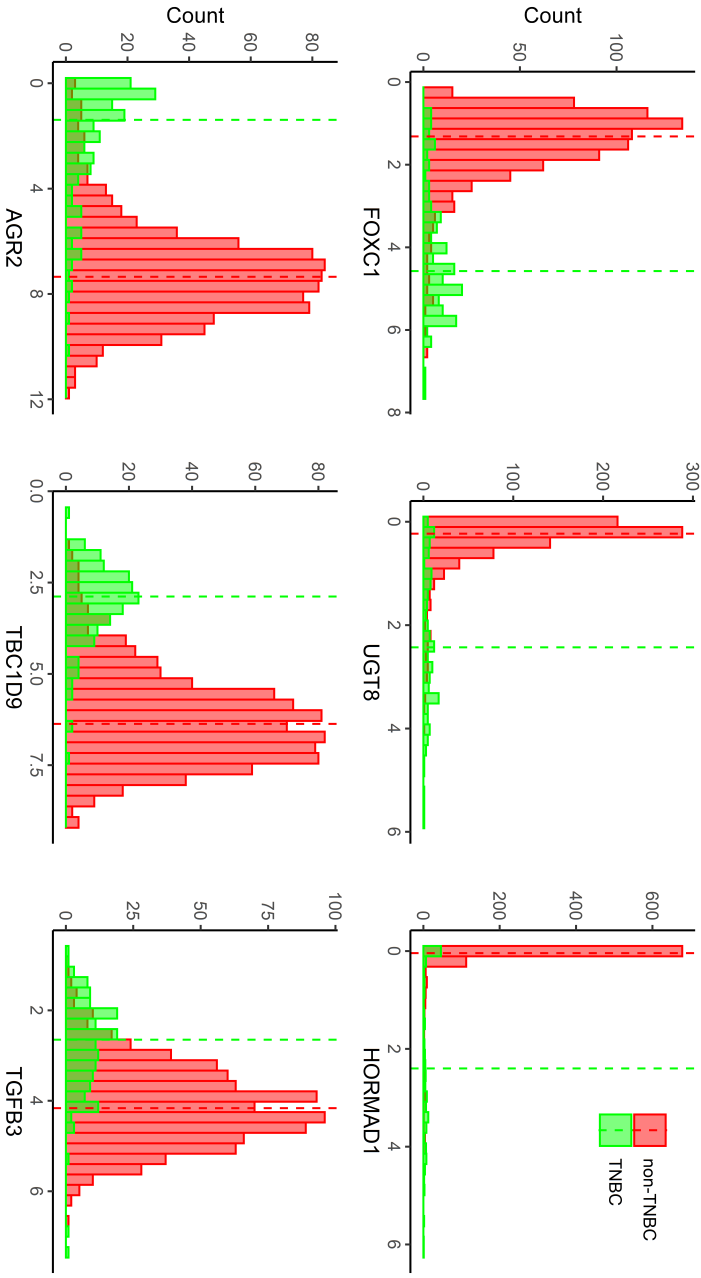

Count
FOXC1
UGT8
HORMAD1
AGR2
TBC1D9
TGFB3
non-TNBC
TNBC

Figure C.4.: Groupwise histograms of TNBC (green), and non-TNBC (red) samples. Vertical lines represent respective group median.

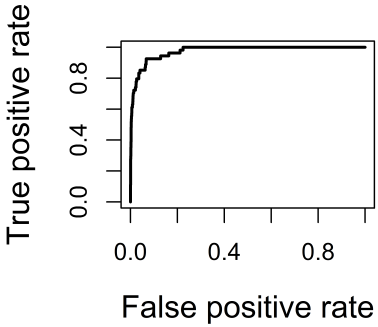


Figure C.5.: ROC curve analysis comparing differentially expressed genes found by edgeR with commonly selected genes from ROSIE. Cutoff values for ROC analysis were taken from the false discovery rate.

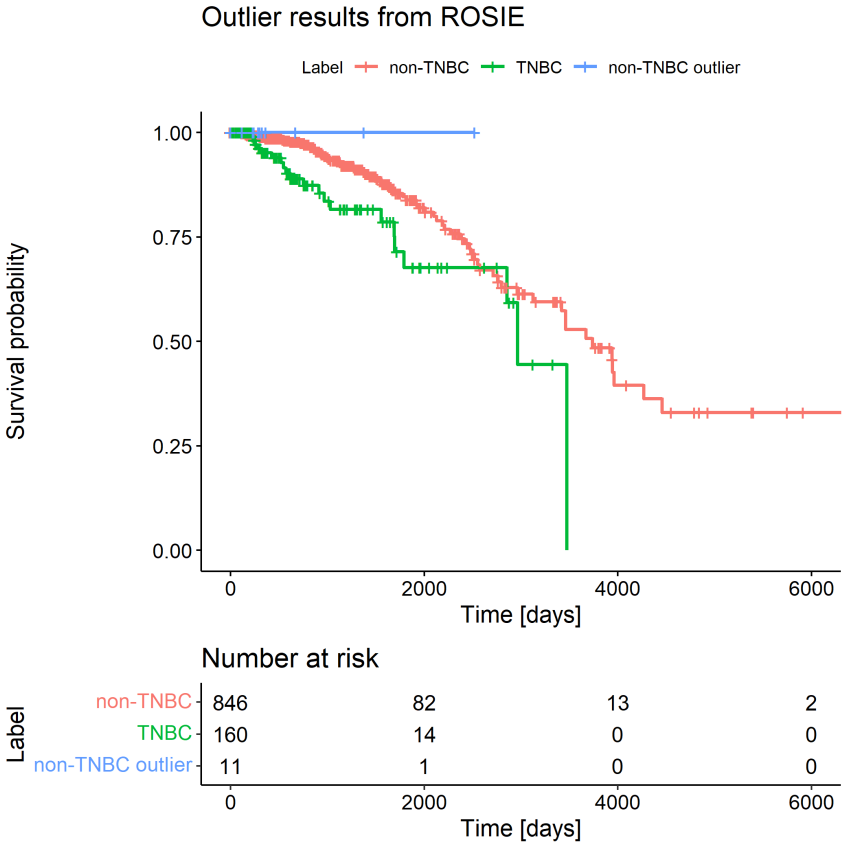


Figure C.6.: Kaplan–Meier curves and numbers at risk at different time points of TNBC, non-TNBC and outliers

D. Bibliography

[1] C. C. Aggarwal. *Data mining: the textbook*. Springer, 2015.

[2] S. Ahmed, K. G. Grant, L. E. Edwards, A. Rahman, M. Cirit, M. B. Goshe, and J. M. Haugh. "Data-driven modeling reconciles kinetics of ERK phosphorylation, localization, and activity states". In: *Mol Syst Biol* 10.1 (2014), p. 718.

[3] M. E. Ahsen, T. P. Boren, N. K. Singh, B. Misganaw, D. G. Mutch, K. N. Moore, F. J. Backes, C. K. McCourt, J. S. Lea, D. S. Miller, et al. "Sparse feature selection for classification and prediction of metastasis in endometrial cancer". In: *BMC Genomics* 18.3 (2017), pp. 1–12.

[4] H. Akaike. "Information theory and an extension of the maximum likelihood principle". In: *Proc of the Second International Symposium on Information Theory*. Ed. by N. Petrov. Budapest, 1973, pp. 267–81.

[5] C. Alabert et al. "Domain Model Explains Propagation Dynamics and Stability of Histone H3K27 and H3K36 Methylation Landscapes". In: *Cell Rep* 30.4 (2020), 1223–1234.e8. DOI: 10.1016/j.celrep.2019.12.060.

[6] S. Alford, R. Robinett, L. Milechin, and J. Kepner. "Pruned and Structurally Sparse Neural Networks". In: *2018 IEEE MIT URTC*. 2018, pp. 1–4. DOI: 10.1109/URTC45901.2018.9244787.

[7] U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. CRC Press, 2006.

[8] U. Alon. "Network motifs: theory and experimental approaches". In: *Nature Reviews Genetics* 8.6 (2007), pp. 450–461.

[9] F. Alpy and C. Tomasetto. “START ships lipids across interorganelle space”. In: *Biochimie* 96 (2014). Lipids in Metabolic Diseases, pp. 85–95. DOI: 10.1016/j.biochi.2013.09.015.

[10] M. Amano, M. Nakayama, and K. Kaibuchi. “Rho-kinase/ROCK: A key regulator of the cytoskeleton and cell polarity”. In: *Cytoskeleton (Hoboken)* 67.9 (2010), pp. 545–554. DOI: 10.1002/cm.20472.

[11] G. de Anda-Jáuregui and E. Hernández-Lemus. “Computational Oncology in the Multi-Omics Era: State of the Art”. In: *Front Oncol* 10 (2020), p. 423. DOI: 10.3389/fonc.2020.00423.

[12] S. Andorf, T. Gärtner, M. Steinfath, H. Witucka-Wall, T. Altmann, and D. Repsilber. “Towards systems biology of heterosis: a hypothesis about molecular network structure applied for the Arabidopsis metabolome”. In: *EURASIP J Bioinform Syst Biol* 2009 (2008), pp. 1–12.

[13] D. Angeli, J. E. Ferrell, and E. D. Sontag. “Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems”. In: *Proc Natl Acad Sci* 101.7 (2004), pp. 1822–1827.

[14] A. Askari, A. d’Aspremont, and L. E. Ghaoui. “Naive Feature Selection: Sparsity in Naive Bayes”. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1813–1822.

[15] N. Bairagi and D. Adak. “Global analysis of HIV-1 dynamics with Hill type infection rate and intracellular delay”. In: *Appl Math Model* 38.21 (2014), pp. 5047–5066. DOI: 10.1016/j.apm.2014.03.010.

[16] F. Bard et al. “Functional genomics reveals genes involved in protein secretion and Golgi organization”. In: *Nature* 439.7076 (2006), p. 604. DOI: 10.1038/nature04377.

[17] N. Barkai and B.-Z. Shilo. "Variability and robustness in biomolecular systems". In: *Mol Cell* 28.5 (2007), pp. 755–760.

[18] C. Baron and V. Malhotra. "Role of Diacylglycerol in PKD Recruitment to the TGN and Protein Transport to the Plasma Membrane". In: *Science* 295.5553 (2002), pp. 325–328. DOI: 10.1126/science.1066759.

[19] L. Bast, F. Calzolari, M. K. Strasser, J. Hasenauer, F. J. Theis, J. Ninkovic, and C. Marr. "Increasing Neural Stem Cell Division Asymmetry and Quiescence Are Predicted to Contribute to the Age-Related Decline in Neurogenesis". In: *Cell Rep* 25.12 (2018), 3231–3240.e8. DOI: 10.1016/j.celrep.2018.11.088.

[20] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. "The protein data bank". In: *Nucleic Acids Res* 28.1 (2000), pp. 235–242.

[21] D. Berrar. "Cross-Validation". In: *Encyclopedia of Bioinformatics and Computational Biology*. Ed. by S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach. Oxford: Academic Press, 2019, pp. 542–545. DOI: 10.1016/B978-0-12-809633-8.20349-X.

[22] J. Berro. "'Essentially, all models are wrong, but some are useful'—a cross-disciplinary agenda for building useful models in cell biology and biophysics". In: *Biophys Rev* 10.6 (2018), pp. 1637–1647.

[23] X. A. Bi, Y. Wang, Q. Shu, Q. Sun, and Q. Xu. "Classification of Autism Spectrum Disorder Using Random Support Vector Machine Cluster". In: *Front Genet* 9 (2018), p. 18. DOI: 10.3389/fgene.2018.00018.

[24] M. R. Birtwistle, J. Rauch, A. Kiyatkin, E. Aksamitiene, M. Dobrzyński, J. B. Hoek, W. Kolch, B. A. Ogunnaike, and B. N. Kholodenko. "Emergence of bimodal cell population responses from the interplay between analog single-cell signaling and protein expression noise". In: *BMC Syst Biol* 6.1 (2012), pp. 1–12.

[25] F. Blanchini and E. Franco. “Multistability and robustness of the MAPK pathway”. In: *2011 50th IEEE Conf Decis Control European Control Conf*. IEEE. 2011, pp. 2214–2219.

[26] F. Blanchini, E. Franco, and G. Giordano. “Determining the structural properties of a class of biological models”. In: *2012 51st IEEE CDC*. IEEE. 2012, pp. 5505–5510.

[27] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel. “Variable selection for Naïve Bayes classification”. In: *Comput Oper Res* 135 (2021), p. 105456. DOI: 10.1016/j.cor.2021.105456.

[28] J. von Blume, J. M. Duran, E. Forlanelli, A.-M. Alleaume, M. Egorov, R. Polishchuk, H. Molina, and V. Malhotra. “Actin remodeling by ADF/cofilin is required for cargo sorting at the trans-Golgi network”. In: *J Cell Biol* 187.7 (2009), pp. 1055–1069.

[29] N. Blüthgen. “Signaling output: it’s all about timing and feedbacks”. In: *Mol Syst Biol* 11.11 (2015), p. 843. DOI: https://doi.org/10.15252/msb.20156642.

[30] S. Bolte and F. P. Cordelières. “A guided tour into subcellular colocalization analysis in light microscopy”. In: *J Microsc* 224.3 (2006), pp. 213–232. DOI: 10.1111/j.1365-2818.2006.01706.x.

[31] J. Bootkrajang and A. Kabán. “Classification of mislabelled microarrays using robust sparse logistic regression”. In: *Bioinformatics* 29.7 (2013), pp. 870–877. DOI: 10.1093/bioinformatics/btt078.

[32] J. L. Bos, H. Rehmann, and A. Wittinghofer. “GEFs and GAPs: critical elements in the control of small G proteins”. In: *Cell* 129.5 (2007), pp. 865–877.

[33] G. E. P. Box and N. R. Draper. *Empirical model-building and response surfaces*. John Wiley & Sons, 1987.

[34] A. C. Braun and M. A. Olayioye. “Rho regulation: DLC proteins in space and time”. In: *Cell Signal* 27.8 (2015), pp. 1643–1651.

[35] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk. “Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments”. In: *FEBS Lett* 573.1-3 (2004), pp. 83–92. DOI: 10.1016/j.febslet.2004.07.055.

[36] F. A. Brightman and D. A. Fell. “Differential feedback regulation of the MAPK cascade underlies the quantitative differences in EGF and NGF signalling in PC12 cells”. In: *FEBS Lett* 482.3 (2000), pp. 169–174.

[37] S. P. Brooks and G. O. Roberts. “Assessing convergence of Markov chain Monte Carlo algorithms”. In: *Stat Comput* 8.4 (1998), pp. 319–335.

[38] J. E. Cavanaugh. “Unifying the derivations of the Akaike and corrected Akaike information criteria”. In: *Stat Probab Lett* 31 (1997), pp. 201–8. DOI: 10.1016/s0167-7152(96)00128-9.

[39] B. Chen, H. Tang, X. Chen, G. Zhang, Y. Wang, X. Xie, and N. Liao. “Transcriptomic analyses identify key differentially expressed genes and clinical outcomes between triple-negative and non-triple-negative breast cancer”. In: *Cancer Manag Res* 11 (2019), pp. 179–190. DOI: 10.2147/CMAR.S187151.

[40] C. Chen, W. T. Baumann, R. Clarke, and J. J. Tyson. “Modeling the estrogen receptor to growth factor receptor signaling switch in human breast cancer cells”. In: *FEBS Lett* 587.20 (2013), pp. 3327–3334. DOI: 10.1016/j.febslet.2013.08.022.

[41] Y.-T. Chen, C. A. Venditti, G. Theiler, B. J. Stevenson, C. Iseli, A. O. Gure, C. V. Jongeneel, L. J. Old, and A. J. G. Simpson. “Identification of CT46/HORMAD1, an immunogenic cancer/testis antigen encoding a putative meiosis-related protein”. In: *Cancer Immun* 5.1 (2005).

[42] P. Cheng, Y. Yang, and Y. Liu. “Interlocked feedback loops contribute to the robustness of the Neurospora circadian clock”. In: *Proceedings of the National Academy of Sciences* 98.13 (2001), pp. 7408–7413.

[43] T.-C. Chou. “Derivation and properties of Michaelis-Menten type and Hill type equations for reference ligands”. In: *J Theor Biol* 59.2 (1976), pp. 253–276. DOI: https://doi.org/10.1016/0022-5193(76)90169-7.

[44] S. Clodong, U. Dühring, L. Kronk, A. Wilde, I. Axmann, H. Herzel, and M. Kollmann. “Functioning and robustness of a bacterial circadian clock”. In: *Mol Syst Biol* 3.1 (2007), p. 90.

[45] S. R. Cole, H. Chu, and S. Greenland. “Maximum likelihood, profile likelihood, and penalized likelihood: a primer”. In: *Am J Epidemiol* 179.2 (2014), pp. 252–260. DOI: 10.1093/aje/kwt245.

[46] C. F. Cowell, I. K. Yan, T. Eiseler, A. C. Leightner, H. Döppler, and P. Storz. “Loss of cell–cell contacts induces NF-κB via RhoA-mediated activation of protein kinase D1”. In: *J Cell Biochem* 106.4 (2009), pp. 714–728.

[47] J. A. Cuesta-Albertos, A. Gordaliza, and C. Matrán. “Trimmed k-means: an attempt to robustify quantizers”. In: *Ann Stat* 25.2 (1997), pp. 553–576. DOI: 10.1214/aos/1031833664.

[48] P. Dai, Z. Yan, S. Ma, Y. Yang, Q. Wang, C. Hou, Y. Wu, Y. Liu, and Q. Diao. “The herbicide glyphosate negatively affects midgut bacterial communities and survival of honey bee during larvae reared in vitro”. In: *J Agric Food Chem* 66.29 (2018), pp. 7786–7793.

[49] J. Davis and M. Goadrich. “The Relationship between Precision-Recall and ROC Curves”. In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML ’06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 233–240. DOI: 10.1145/1143844.1143874.

[50] A. Degasperi, M. Birtwistle, N. Volinsky, J. Rauch, W. Kolch, and B. Kholodenko. “Evaluating strategies to normalize biological replicates of Western Blot data”. In: *PLoS One* 9.1 (2014), e87293.

[51] G. Dong and H. Liu. *Feature engineering for machine learning and data analytics*. CRC Press, 2018.

[52] H. Döppler, P. Storz, J. Li, M. J. Comb, and A. Toker. “A phosphorylation state-specific antibody recognizes Hsp27, a novel substrate of protein kinase D”. In: *J Biol Chem* 280.15 (2005), pp. 15013–15019.

[53] N. Durand, L. I. Bastea, J. Long, H. Döppler, K. Ling, and P. Storz. “Protein Kinase D1 regulates focal adhesion dynamics and cell adhesion through Phosphatidylinositol-4-phosphate 5-kinase type-l γ”. In: *Sci Rep* 6 (2016), p. 35963.

[54] B. Efron. “Bayesian inference and the parametric bootstrap”. In: *Annal Appl Stat* 6.4 (2012), p. 1971.

[55] I. Eisenkolb, A. Jensch, K. Eisenkolb, A. Kramer, P. C. Buchholz, J. Pleiss, A. Spiess, and N. E. Radde. “Modeling of biocatalytic reactions: A workflow for model calibration, selection, and validation using Bayesian statistics”. In: *AIChE J* (2019), e16866.

[56] A. Fiedler, S. Raeth, F. J. Theis, A. Hausser, and J. Hasenauer. “Tailored parameter optimization methods for ordinary differential equation models with steady-state constraints”. In: *BMC Syst Biol* 10.1 (2016), pp. 1–19.

[57] R. Fisher. “The Use of Multiple Measurements in Taxonomic Problems”. In: *Ann Eugen* 7.2 (1936), pp. 179–188.

[58] G. Fletcher, S. Patel, K. Tyson, P. Adam, M. Schenker, J. Loader, L. Daviet, P. Legrain, R. Parekh, A. Harris, et al. “hAG-2 and hAG-3, human homologues of genes involved in differentiation, are associated with oestrogen receptor-positive breast tumours and interact with metastasis gene C4. 4a and dystroglycan”. In: *Br J Cancer* 88.4 (2003), pp. 579–585.

[59] M. L. Forister, E. M. Pelton, and S. H. Black. “Declines in insect abundance and diversity: We know enough to act now”. In: *Conserv Sci Pract* 1.8 (2019), e80. DOI: 10.1111/csp2.80.

[60] E. Franco and F. Blanchini. “Structural properties of the MAPK pathway topologies in PC12 cells”. In: *J Math Biol* 67.6 (2013), pp. 1633–1668.

[61] M. Freeman. “Feedback control of intercellular signalling in development”. In: *Nature* 408.6810 (2000), pp. 313–319.

[62] R. Fritsche-Guenther, F. Witzel, A. Sieber, R. Herr, N. Schmidt, S. Braun, T. Brummer, C. Sers, and N. Blüthgen. “Strong negative feedback from Erk to Raf confers robustness to MAPK signalling”. In: *Mol Syst Biol* 7.1 (2011), p. 489.

[63] Y. Fu and C. S. Rubin. “Protein kinase D: coupling extracellular stimuli to the regulation of cell physiology”. In: *EMBO Rep* 12.8 (2011), pp. 785–796.

[64] Y. F. Fuchs, S. A. Eisler, G. Link, O. Schlicker, G. Bunt, K. Pfizenmaier, and A. Hausser. “A Golgi PKD Activity Reporter Reveals a Crucial Role of PKD in Nocodazole-Induced Golgi Dispersal”. In: *Traffic* 10.7 (2009), pp. 858–867. DOI: 10.1111/j.1600-0854.2009.00918.x.

[65] T. Fugmann, A. Hausser, P. Schöffler, S. Schmid, K. Pfizenmaier, and M. A. Olayioye. “Regulation of secretory transport by protein kinase D–mediated phosphorylation of the ceramide transfer protein”. In: *J Cell Biol* 178.1 (2007), pp. 15–22.

[66] L. Gao, M. Ye, X. Lu, and D. Huang. “Hybrid Method Based on Information Gain and Support Vector Machine for Gene Selection in Cancer Classification”. In: *Genomics Proteomics Bioinformatics* 15.6 (2017), pp. 389–395. DOI: 10.1016/j.gpb.2017.08.002.

[67] L. A. García-Escudero, A. Mayo-Iscar, and M. Riani. “Constrained parsimonious model-based clustering”. In: *Stat Comput* 32.1 (2022), pp. 1–15.

[68] L. A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. "A general trimming approach to robust cluster Analysis". In: *Ann Stat* 36.3 (2008), pp. 1324–1345. DOI: 10.1214/07-AOS515.

[69] L. A. García-Escudero, A. Mayo-Iscar, and M. Riani. "Model-based clustering with determinant-and-shape constraint". In: *Stat Comput* 30 (2020), pp. 1363–1380.

[70] GBIF: The Global Biodiversity Information Facility. *What is GBIF?* Available from https://www.gbif.org/what-is-gbif. 2021.

[71] H. Gehart, A. Goginashvili, R. Beck, J. Morvan, E. Erbs, I. Formentini, M. A. De Matteis, Y. Schwab, F. T. Wieland, and R. Ricci. "The BAR domain protein Arfaptin-1 controls secretory granule biogenesis at the trans-Golgi network". In: *Dev Cell* 23.4 (2012), pp. 756–768.

[72] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. 2nd ed. Chapman & Hall/CRC, 2006.

[73] R. Gesztelyi, J. Zsuga, A. Kemeny-Beke, B. A. Varga, B. Juhász, and Á. Tósaki. "The Hill equation and the origin of quantitative pharmacology". In: *Arch Hist Exact Sci* 66 (2012), pp. 427–438.

[74] K. K. Ghosh, S. Ghosh, S. Sen, R. Sarkar, and U. Maulik. "A two-stage approach towards protein secondary structure classification". In: *Med Biol Eng Comput* 58 (2020), pp. 1723–1737.

[75] Y. Gotoh, E. Nishida, T. Yamashita, M. Hoshi, M. Kawakami, and H. Sakai. "Microtubule-associated-protein (MAP) kinase activated by nerve growth factor and epidermal growth factor in PC12 cells: Identity with the mitogen-activated MAP kinase of fibroblastic cells". In: *Eur J Biochem* 193.3 (1990), pp. 661–669.

[76] D. Goulson, E. Nicholls, C. Botías, and E. L. Rotheray. "Bee declines driven by combined stress from parasites, pesticides, and lack of flowers". In: *Science* 347.6229 (2015), p. 1255957. DOI: 10.1126/science.1255957.

[77] D. Goulson, J. Thompson, and A. Croombs. "Rapid rise in toxic load for bees revealed by analysis of pesticide use in Great Britain". In: *PeerJ* 6 (2018), e5255.

[78] J.-L. Gouzé. "Positive and negative circuits in dynamical systems". In: *J Biol Syst* 6.01 (1998), pp. 11–15.

[79] L. Grieco, L. Calzone, I. Bernard-Pierrot, F. Radvanyi, B. Kahn-Perles, and D. Thieffry. "Integrative modelling of the influence of MAPK network on cancer cell fate decision". In: *PLoS Comput Biol* 9.10 (2013).

[80] F. Gross and M. MacLeod. "Prospects and problems for standardizing model validation in systems biology". In: *Prog Biophys Mol Biol* 129 (2017). Validation of Computer Modelling, pp. 3–12. DOI: 10.1016/j.pbiomolbio.2017.01.003.

[81] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. "Gene selection for cancer classification using support vector machines". In: *Mach Learn* 46 (1 2002), pp. 389–422.

[82] H. Haario, M. Laine, A. Mira, and E. Saksman. "DRAM: efficient adaptive MCMC". In: *Stat Comput* 16.4 (2006), pp. 339–354.

[83] D. Haeufle, M. Günther, A. Bayer, and S. Schmitt. "Hill-type muscle model with serial damping and eccentric force–velocity relation". In: *J Biomech* 47.6 (2014), pp. 1531–1536. DOI: 10.1016/j.jbiomech.2014.02.009.

[84] M. A. Hall. "Correlation-based feature selection for machine learning". PhD thesis. 1999.

[85] C. A. Hallmann et al. "More than 75 percent decline over 27 years in total flying insect biomass in protected areas". In: *PLoS One* 12.10 (2017), pp. 1–21. DOI: 10.1371/journal.pone.0185809.

[86] D. Hanahan and R. A. Weinberg. “The hallmarks of cancer”. In: *cell* 100.1 (2000), pp. 57–70.

[87] D. Hasdemir, H. C. Hoefsloot, and A. K. Smilde. “Validation and selection of ODE based systems biology models: how to arrive at more reliable decisions”. In: *BMC Syst Biol* 9.1 (2015), pp. 1–19.

[88] J. Hasenauer, C. Hasenauer, T. Hucho, and F. J. Theis. “ODE constrained mixture modelling: a method for unraveling subpopulation structures and dynamics”. In: *PLoS Comput Biol* 10.7 (2014), e1003686.

[89] H. Hass, C. Loos, E. Raimúndez-Álvarez, J. Timmer, J. Hasenauer, and C. Kreutz. “Benchmark problems for dynamic modeling of intracellular processes”. In: *Bioinformatics* 35.17 (2019), pp. 3073–3082. DOI: 10.1093/bioinformatics/btz020.

[90] W. K. Hastings. “Monte Carlo sampling methods using Markov chains and their applications”. In: *Biometrika* 57.1 (1970), pp. 97–109. DOI: 10.1093/biomet/57.1.97.

[91] A. Hausser, G. Link, L. Bamberg, A. Burzlaff, S. Lutz, K. Pfizenmaier, and F. J. Johannes. “Structural requirements for localization and activation of protein kinase C mu (PKC mu) at the Golgi compartment”. In: *J Cell Biol* 156.1 (2002), pp. 65–74.

[92] A. Hausser, S. Märtens, G. Link, K. Pfizenmaier, P. Storz, and A. Toker. “Protein kinase D regulates vesicular transport by phosphorylation and activation of phosphatidylinositol-4 kinase III β at the Golgi complex”. In: *Nat Cell Biol* 7.9 (2005), p. 880.

[93] A. Hausser, P. Storz, G. Link, H. Stoll, Y.-C. Liu, A. Altman, K. Pfizenmaier, and F.-J. Johannes. “Protein kinase C μ is negatively regulated by 14-3-3 signal transduction proteins”. In: *J Biol Chem* 274.14 (1999), pp. 9258–9264.

[94] D. M. Hawkins. *Identification of outliers*. Vol. 11. Springer, 1980.

[95] K. D. Healy, L. Hodgson, T.-Y. Kim, A. Shutes, S. Maddileti, R. L. Juliano, K. M. Hahn, T. K. Harden, Y.-J. Bang, and C. J. Der. "DLC-1 suppresses non-small cell lung cancer growth and invasion by RhoGAP-dependent and independent mechanisms". In: *Mol Carcinog* 47.5 (2008), pp. 326–337.

[96] T. Heskes, R. Eisinga, and R. Breitling. "A fast algorithm for determining bounds and accurate approximate p-values of the rank product statistic for replicate experiments". In: *BMC Bioinformatics* 15.1 (2014), p. 367. DOI: 10.1186/s12859-014-0367-1.

[97] A. E. Hoerl and R. W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67.

[98] I. Hoffmann, P. Filzmoser, S. Serneels, and K. Varmuza. "Sparse and robust PLS for binary classification". In: *J Chemom* 30.4 (2016), pp. 153–162. DOI: 10.1002/cem.2775.

[99] G. Holeiter, J. Heering, P. Erlmann, S. Schmid, R. Jähne, and M. A. Olayioye. "Deleted in liver cancer 1 controls cell migration through a Dia1-dependent signaling pathway". In: *Cancer Res* 68.21 (2008), pp. 8743–8751.

[100] J. H. Holland et al. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. 2nd ed. MIT press, 1992.

[101] Y. Homma and Y. Emori. "A dual functional signal mediator showing RhoGAP and phospholipase C-delta stimulating activities." In: *EMBO J* 14.2 (1995), p. 286.

[102] R. Hooke and T. A. Jeeves. "'Direct Search' Solution of Numerical and Statistical Problems". In: *J ACM* 8.2 (1961), pp. 212–229.

[103] J. J. Hornberg, F. J. Bruggeman, H. V. Westerhoff, and J. Lankelma. "Cancer: A Systems Biology disease". In: *Biosystems* 83.2 (2006). 5th International Conference on Systems Biology, pp. 81–90. DOI: 10.1016/j.biosystems.2005.05.014.

[104] X. Hu, C. Rudin, and M. Seltzer. “Optimal sparse decision trees”. In: *Adv Neural Inf Process Syst* (2019).

[105] C.-Y. Huang and J. E. Ferrell. “Ultrasensitivity in the mitogen-activated protein kinase cascade”. In: *Proc Natl Acad Sci* 93.19 (1996), pp. 10078–10083.

[106] A. Jabeen, N. Ahmad, and K. Raza. “Machine learning-based state-of-the-art methods for the classification of rna-seq data”. In: *Classification in BioApps*. Springer, 2018, pp. 133–172.

[107] A. B. Jaffe and A. Hall. “Rho GTPases: biochemistry and biology”. In: *Annu Rev Cell Dev Biol* 21 (2005), pp. 247–269.

[108] M. Jammal, S. Canu, and M. Abdallah. “Robust and Sparse Support Vector Machines via Mixed Integer Programming”. In: *Machine Learning, Optimization, and Data Science*. Ed. by G. Nicosia, V. Ojha, E. La Malfa, G. Jansen, V. Sciacca, P. Pardalos, G. Giuffrida, and R. Umeton. Cham: Springer International Publishing, 2020, pp. 572–585.

[109] M. K. C. from Jed Wing et al. *caret: Classification and Regression Training*. R package version 6.0-84. 2019.

[110] A. Jensch, Y. Frey, K. Bitschar, P. Weber, S. Schmid, A. Hausser, M. A. Olayioye, and N. E. Radde. “The tumor suppressor protein DLC1 maintains protein kinase D activity and Golgi secretory function”. In: *J Biol Chem* 293.37 (2018), pp. 14407–14416.

[111] A. Jensch, M. B. Lopes, S. Vinga, and N. Radde. “ROSIE: RObust Sparse ensemble for outlIEr detection and gene selection in cancer omics data”. In: *Stat Methods Med Res* 31.5 (2022), pp. 947–958.

[112] A. Jensch, C. Thomaseth, and N. E. Radde. “Sampling-based Bayesian approaches reveal the importance of quasi-bistable behavior in cellular decision processes on the example of the MAPK signaling pathway in PC-12 cell lines”. In: *BMC Syst Biol* 11.1 (2017), p. 11.

[113] K. A. Johnson and R. S. Goody. "The Original Michaelis Constant: Translation of the 1913 Michaelis–Menten Paper". In: *Biochemistry* 50.39 (2011). PMID: 21888353, pp. 8264–8269. DOI: 10.1021/bi201284u.

[114] D. R. Jones, C. D. Perttunen, and B. E. Stuckman. "Lipschitzian optimization without the Lipschitz constant". In: *J Optim Theory Appl* 79.1 (1993), pp. 157–181.

[115] H. de Jong. "Modeling and Simulation of Genetic Regulatory Systems: A Literature Review". In: *J Comput Biol* 9.1 (2002). PMID: 11911796, pp. 67–103. DOI: 10.1089/10665270252833208.

[116] D. Joshi, A. Khajuria, and P. Joshi. "An automatic non-invasive method for Parkinson's disease classification". In: *Comput Methods Programs Biomed* 145 (2017), pp. 135–145. DOI: 10.1016/j.cmpb.2017.04.007.

[117] A. Jović, K. Brkić, and N. Bogunović. "A review of feature selection methods with applications". In: *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2015, pp. 1200–1205. DOI: 10.1109/MIPRO.2015.7160458.

[118] M. Kan, M. Shumyatcher, and B. E. Himes. "Using omics approaches to understand pulmonary diseases". In: *Respir Res* 18.1 (2017), pp. 1–20.

[119] R. E. Kass and A. E. Raftery. "Bayes Factors". In: *J Am Stat Assoc* 90.430 (1995), pp. 773–95. DOI: 10.2307/2291091.

[120] H. Kato, S. Takahashi, and K. Saito. "Omics and Integrated Omics for the Promotion of Food and Nutrition Science". In: *J Tradit Complement Med* 1.1 (2011), pp. 25–30. DOI: 10.1016/S2225-4110(16)30053-0.

[121] J. Kennedy and R. Eberhart. "Particle swarm optimization". In: *Proc of ICNN'95 - Int Conf Neural Netw.* Vol. 4. 1995, 1942–1948 vol.4. DOI: 10.1109/ICNN.1995.488968.

[122] B. N. Kholodenko. “Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades”. In: *Eur J Biochem* 267.6 (2000), pp. 1583–1588.

[123] 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”. In: *Proc Natl Acad Sci* 99.20 (2002), pp. 12841–12846.

[124] B. Kim and S. J. Shin. “Principal weighted logistic regression for sufficient dimension reduction in binary classification”. In: *J Korean Stat Soc* 48.2 (2019), pp. 194–206.

[125] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. “Optimization by Simulated Annealing”. In: *Science* 220.4598 (1983), pp. 671–680. DOI: 10.1126/science.220.4598.671.

[126] H. Kitano. *Foundations of systems biology*. The MIT Press Cambridge, Massachusetts London, England, 2001.

[127] E. Klipp, W. Liebermeister, C. Wierling, and A. Kowald. *Systems biology: a textbook*. John Wiley & Sons, 2016.

[128] P. Kocieniewski and T. Lipniacki. “MEK1 and MEK2 differentially control the duration and amplitude of the ERK cascade response”. In: *Phys Biol* 10.3 (2013), p. 035006.

[129] W. Kolch. “Coordinating ERK/MAPK signalling through scaffolds and inhibitors”. In: *Nat Rev Mol Cell Biol* 6.11 (2005), pp. 827–837.

[130] W. Kolch, M. Calder, and D. Gilbert. “When kinases meet mathematics: the systems biology of MAPK signalling”. In: *FEBS Lett* 579.8 (2005), pp. 1891–1895.

[131] Y. Kondo. *RSKC: Robust Sparse K-Means*. R package version 2.4.2. 2016.

[132] Y. Kondo, M. Salibian-Barrera, R. Zamar, et al. “RSKC: an R package for a robust and sparse k-means clustering algorithm”. In: *J Stat Softw* 72.5 (2016), pp. 1–26.

[133] I. Kononenko. “Estimating attributes: Analysis and extensions of RELIEF”. In: *European conference on machine learning*. Springer. 1994, pp. 171–182.

[134] C. Kothari, M. A. Osseni, L. Agbo, G. Ouellette, M. Déraspe, F. Laviolette, J. Corbeil, J.-P. Lambert, C. Diorio, and F. Durocher. “Machine learning analysis identifies genes differentiating triple negative breast cancers”. In: *Sci Rep* 10.1 (2020), pp. 1–15.

[135] C. Kreutz, M. Rodriguez, T. Maiwald, M. Seidl, H. Blum, L. Mohr, and J. Timmer. “An error model for protein quantification”. In: *Bioinformatics* 23.20 (2007), pp. 2747–53.

[136] F. S. Kurnaz, I. Hoffmann, and P. Filzmoser. *enetLTS: Robust and Sparse Methods for High Dimensional Linear and Logistic Regression*. 2018.

[137] F. S. Kurnaz, I. Hoffmann, and P. Filzmoser. “Robust and sparse estimation methods for high-dimensional linear and logistic regression”. In: *Chemometr Intell Lab Syst* 172 (2018), pp. 211–222. DOI: 10.1016/j.chemolab.2017.11.017.

[138] S. K. Kwak and J. H. Kim. “Statistical data preparation: management of missing values and outliers”. In: *Korean J Anesthesiol* 70.4 (2017), p. 407.

[139] P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armananzas, G. Santafé, A. Pérez, et al. “Machine learning in bioinformatics”. In: *Brief Bioinform* 7.1 (2006), pp. 86–112.

[140] H. Lavoie and M. Therrien. “Regulation of RAF protein kinases in ERK signalling”. In: *Nat Rev Mol Cell Biol* 16.5 (2015), pp. 281–298.

[141] R. LeCover, T. Orfeo, K. Brummel-Ziedins, M. Bravo, A. Pusateri, and J. Varner. “Kinetic Modeling of Coagulation and Fibrinolysis”. In: *IFAC-PapersOnLine* 52.26 (2019). 8th Conference on Foundations of Systems Biology in Engineering FOSBE 2019, pp. 94–100. DOI: 10.1016/j.ifacol.2019.12.242.

[142] J. Lee, A. Tiwari, V. Shum, G. B. Mills, M. A. Mancini, O. A. Igoshin, and G. Balázsi. "Unraveling the regulatory connections between two controllers of breast cancer cell fate". In: *Nucleic Acids Res* 42.11 (2014), pp. 6839–6849.

[143] J. Lee, J. Lee, K. S. Farquhar, J. Yun, C. A. Frankenberger, E. Bevilacqua, K. Yeung, E.-J. Kim, G. Balázsi, and M. R. Rosner. "Network of mutually repressive metastasis regulators can promote cell heterogeneity and metastatic transitions". In: *Proc Natl Acad Sci* 111.3 (2014), E364–E373.

[144] S. Legewie, B. Schoeberl, N. Blüthgen, and H. Herzel. "Competing docking interactions can bring about bistability in the MAPK cascade". In: *Biophys J* 93.7 (2007), pp. 2279–2288.

[145] E. Limpert, W. A. Stahel, and M. Abbt. "Log-normal Distributions across the Sciences: Keys and Clues: On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question". In: *BioScience* 51.5 (2001), pp. 341–352. DOI: 10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2.

[146] D. Liu, Y. Shi, Y. Tian, and X. Huang. "Ramp loss least squares support vector machine". In: *J Comput Sci* 14 (2016). The Route to Exascale: Novel Mathematical Methods, Scalable Algorithms and Computational Science Skills, pp. 61–68. DOI: 10.1016/j.jocs.2016.02.001.

[147] H. Liu, H. Motoda, R. Setiono, and Z. Zhao. "Feature Selection: An Ever Evolving Frontier in Data Mining". In: *Proceedings of the Fourth International Workshop on Feature Selection in Data Mining*. Ed. by H. Liu, H. Motoda, R. Setiono, and Z. Zhao. Vol. 10. Proceedings of Machine Learning Research. Hyderabad, India: PMLR, 2010, pp. 4–13.

[148] H. Liu and R. Setiono. "Chi2: feature selection and discretization of numeric attributes". In: *Proc 7th IEEE Int Conf Tools Artif Intell*. 1995, pp. 388–391. DOI: 10.1109/TAI.1995.479783.

[149] M. B. Lopes, A. Veríssimo, E. Carrasquinha, S. Casimiro, N. Beerenwinkel, and S. Vinga. "Ensemble outlier detection and gene selection in triple-negative breast cancer data". In: *BMC Bioinformatics* 19.1 (2018), p. 168.

[150] Z. Mai and H. Liu. "Random parameter sampling of a generic three-tier MAPK cascade model reveals major factors affecting its versatile dynamics". In: *PloS one* 8.1 (2013), e54441.

[151] C. Maier, C. Loos, and J. Hasenauer. "Robust parameter estimation for dynamical systems from outlier-corrupted data". In: *Bioinformatics* 33.5 (2016), pp. 718–725. DOI: 10.1093/bioinformatics/btw703.

[152] V. Malhotra and F. Campelo. "PKD regulates membrane fission to generate TGN to cell surface transport carriers". In: *Cold Spring Harb Perspect Biol* 3.2 (2011), a005280.

[153] P. Mamoshina, M. Volosnikova, I. V. Ozerov, E. Putin, E. Skibina, F. Cortese, and A. Zhavoronkov. "Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification". In: *Front Genet* 9 (2018), p. 242. DOI: 10.3389/fgene.2018.00242.

[154] T. Manninen, H. Huttunen, P. Ruusuvuori, and M. Nykter. "Leukemia Prediction Using Sparse Logistic Regression". In: *PLoS One* 8.8 (2013), pp. 1–10. DOI: 10.1371/journal.pone.0072932.

[155] M. D. Mckay, R. J. Beckman, and W. J. Conover. "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code". In: *Technometrics* 42.1 (2000), pp. 55–61. DOI: 10.1080/00401706.2000.10485979.

[156] G. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Interscience, 2004.

[157] R. Medar, V. S. Rajpurohit, and B. Rashmi. "Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning". In: *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*. 2017, pp. 1–6. DOI: 10.1109/ICCUBEA.2017.8463779.

[158] R. Merkle et al. "Identification of Cell Type-Specific Differences in Erythropoietin Receptor Signaling in Primary Erythroid and Lung Cancer Cells". In: *PLOS Comput Biol* 12.8 (2016), pp. 1–34. DOI: 10.1371/journal.pcbi.1005049.

[159] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. "Equation of State Calculations by Fast Computing Machines". In: *J Chem Phys* 21.6 (1953), pp. 1087–1092. DOI: 10.1063/1.1699114.

[160] L. Michaelis, M. L. Menten, et al. "Die kinetik der invertinwirkung". In: *Biochem. z* 49.333-369 (1913), p. 352.

[161] M. Mishra, S. Tiwari, and A. V. Gomes. "Protein purification and analysis: next generation Western blotting techniques". In: *Expert Rev Proteomics* 14.11 (2017). PMID: 28974114, pp. 1037–1053. DOI: 10.1080/14789450.2017.1388167.

[162] A. Y. Mitrophanov and E. A. Groisman. "Positive feedback in cellular control systems". In: *Bioessays* 30.6 (2008), pp. 542–555.

[163] T. Miyasaka, M. Chao, P. Sherline, and A. Saltiel. "Nerve growth factor stimulates a protein kinase in PC-12 cells that phosphorylates microtubule-associated protein-2." In: *J Biol Chem* 265.8 (1990), pp. 4730–4735.

[164] V. Moignard, S. Woodhouse, L. Haghverdi, A. J. Lilly, Y. Tanaka, A. C. Wilkinson, F. Buettner, I. C. Macaulay, W. Jawaid, E. Diamanti, et al. "Decoding the regulatory network of early blood development from single-cell gene expression measurements". In: *Nat Biotechnol* 33.3 (2015), pp. 269–276.

[165] S. Motta and F. Pappalardo. "Mathematical modeling of biological systems". In: *Brief Bioinform* 14.4 (2012), pp. 411–422. DOI: 10.1093/bib/bbs061.

[166] S. Mousavi, D. Lee, T. Griffin, D. Steadman, and A. Mockus. "Collaborative Learning Of Semi-Supervised Clustering And Classification For Labeling Uncurated Data". In: *2020 IEEE ICIP*. 2020, pp. 1716–1720. DOI: 10.1109/ICIP40778.2020.9191300.

[167] T. Nguyen, J. Scimeca, C. Filloux, P. Peraldi, J. Carpentier, and E. Van Obberghen. "Co-regulation of the mitogen-activated protein kinase, extracellular signal-regulated kinase 1, and the 90-kDa ribosomal S6 kinase in PC12 cells. Distinct effects of the neurotrophic factor, nerve growth factor, and the mitogenic factor, epidermal growth factor". In: *J Biol Chem* 268.13 (1993), pp. 9803–9810.

[168] S. Nhek, M. Ngo, X. Yang, M. M. Ng, S. J. Field, J. M. Asara, N. D. Ridgway, and A. Toker. "Regulation of Oxysterol-binding Protein Golgi Localization through Protein Kinase D–mediated Phosphorylation". In: *Mol Biol Cell* 21.13 (2010), pp. 2327–2337.

[169] T. Nozawa, S. Sano, A. Minowa-Nozawa, H. Toh, S. Nakajima, K. Murase, C. Aikawa, and I. Nakagawa. "TBC1D9 regulates TBK1 activation through Ca 2+ signaling in selective autophagy". In: *Nat Commun* 11.1 (2020), pp. 1–16.

[170] M. A. Olayioye, S. Barisic, and A. Hausser. "Multi-level control of actin dynamics by protein kinase D". In: *Cell Signal* 25.9 (2013), pp. 1739–1747.

[171] D. V. Olivença, I. Uliyakina, L. L. Fonseca, M. D. Amaral, E. O. Voit, and F. R. Pinto. "A mathematical model of the phosphoinositide pathway". In: *Sci Rep* 8.1 (2018), pp. 1–12.

[172] R. J. Orton, O. E. Sturm, V. Vyshemirsky, M. Calder, D. R. Gilbert, and W. Kolch. "Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway". In: *Biochem J* 392.2 (2005), pp. 249–261.

[173] M. Paczkowska, J. Barenboim, N. Sintupisut, N. S. Fox, H. Zhu, D. Abd-Rabbo, M. W. Mee, P. C. Boutros, and J. Reimand. "Integrative pathway enrichment analysis of multivariate omics data". In: *Nat Commun* 11.1 (2020), pp. 1–16.

[174] S. K. Pal, B. H. Childs, and M. Pegram. "Triple negative breast cancer: unmet medical needs". In: *Breast Cancer Res Treat* 125.3 (2011), pp. 627–636. DOI: 10.1007/s10549-010-1293-1.

[175] B. Palsson. *Systems biology*. Cambridge university press, 2015.

[176] O. Pertz, L. Hodgson, R. L. Klemke, and K. M. Hahn. "Spatiotemporal dynamics of RhoA activity in migrating cells". In: *Nature* 440.7087 (2006), p. 1069.

[177] N. C. Popescu and S. Goodison. "Deleted in liver cancer-1 (DLC1): an emerging metastasis suppressor gene". In: *Mol Diagn Ther* 18.3 (2014), pp. 293–302.

[178] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, et al. "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes". In: *Genome Res* 19.7 (2009), pp. 1316–1323.

[179] G. V. Pusapati, T. Eiseler, A. Rykx, S. Vandoninck, R. Derua, E. Waelkens, J. Van Lint, G. von Wichert, and T. Seufferlein. "Protein kinase D regulates RhoA activity via rhotekin phosphorylation". In: *J Biol Chem* 287.12 (2012), pp. 9473–9483.

[180] L. Qiao, R. B. Nachbar, I. G. Kevrekidis, and S. Y. Shvartsman. "Bistability and oscillations in the Huang-Ferrell model of MAPK signaling". In: *PLoS Comput Biol* 3.9 (2007), e184.

[181] M.-S. Qiu and S. H. Green. "PC12 cell neuronal differentiation is associated with prolonged p21ras activity and consequent prolonged ERK activity". In: *Neuron* 9.4 (1992), pp. 705–717.

[182] G. Quassollo, J. Wojnacki, D. A. Salas, L. Gastaldi, M. P. Marzolo, C. Conde, M. Bisbal, A. Couve, and A. Cáceres. "A RhoA signaling pathway regulates dendritic Golgi outpost formation". In: *Curr Biol* 25.8 (2015), pp. 971–982.

[183] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018.

[184] N. Radde. "Fixed point characterization of biological networks with complex graph topology". In: *Bioinformatics* 26.22 (2010), pp. 2874–2880.

[185] E. Raimúndez, S. Keller, G. Zwingenberger, K. Ebert, S. Hug, F. J. Theis, D. Maier, B. Luber, and J. Hasenauer. "Model-based analysis of response and resistance factors of cetuximab treatment in gastric cancer cell lines". In: *PLoS Comput Biol* 16.3 (2020), e1007147.

[186] T. V. Rampisela and Z. Rustam. "Classification of Schizophrenia Data Using Support Vector Machine (SVM)". In: *J Phys: Conf Ser* 1108.1 (2018), p. 012044. DOI: 10.1088/1742-6596/1108/1/012044.

[187] A. Rao, Y. Lee, A. Gass, and A. Monsch. "Classification of Alzheimer's Disease from structural MRI using sparse logistic regression with optional spatial regularization". In: *2011 Annu Int Conf IEEE Eng Medicine Biol Soc*. 2011, pp. 4499–4502. DOI: 10.1109/IEMBS.2011.6091115.

[188] 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". In: *Bioinformatics* 25.15 (2009), pp. 1923–1929. DOI: 10.1093/bioinformatics/btp358.

[189] A. Raue et al. "Lessons Learned from Quantitative Dynamical Modeling in Systems Biology". In: *PLoS One* 8.9 (2013), pp. 1–17. DOI: 10.1371/journal.pone.0074335.

[190] A. J. Reid, A. M. Talman, H. M. Bennett, A. R. Gomes, M. J. Sanders, C. J. Illingworth, O. Billker, M. Berriman, and M. K. Lawniczak. "Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites". In: *Elife* 7 (2018), e33105.

[191] T. C. G. A. Research Network. *The Cancer Genome Atlas*. accessed December 2019.

[192] L. M. Rios and N. V. Sahinidis. "Derivative-free optimization: a review of algorithms and comparison of software implementations". In: *J Glob Optim* 56.3 (2013), pp. 1247–1293.

[193] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2010), pp. 139–140. DOI: 10.1093/bioinformatics/btp616.

[194] A. Roy, J. Ye, F. Deng, and Q. J. Wang. "Protein kinase D signaling in cancer: A friend or foe?" In: *Biochim. Biophys. Acta* 1868.1 (2017), pp. 283–294.

[195] T. A. Runkler. *Data analytics*. Springer, 2020.

[196] H. Ryu, M. Chung, M. Dobrzyński, D. Fey, Y. Blum, S. S. Lee, M. Peter, B. N. Kholodenko, N. L. Jeon, and O. Pertz. "Frequency modulation of ERK activation dynamics rewires cell fate". In: *Mol Syst Biol* 11.11 (2015), p. 838.

[197] S. B. Salvarezza, S. Deborde, R. Schreiner, F. Campagne, M. M. Kessels, B. Qualmann, A. Caceres, G. Kreitzer, and E. Rodriguez-Boulan. "LIM kinase 1 and cofilin regulate actin filament population required for dynamin-dependent apical carrier fission from the trans-Golgi network". In: *Mol Biol Cell* 20.1 (2009), pp. 438–451.

[198] F. Sánchez-Bayo and K. A. Wyckhuys. "Worldwide decline of the entomofauna: A review of its drivers". In: *Biol Conserv* 232 (2019), pp. 8–27. DOI: 10.1016/j.biocon.2019.01.020.

[199] S. D. Santos, P. J. Verveer, and P. I. Bastiaens. "Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate". In: *Nat Cell Biol* 9.3 (2007), pp. 324–330.

[200] S. Sasagawa, Y.-i. Ozaki, K. Fujita, and S. Kuroda. "Prediction and validation of the distinct dynamics of transient and sustained ERK activation". In: *Nat Cell Biol* 7.4 (2005), pp. 365–373.

[201] M. A. Savageau. "Comparison of classical and autogenous systems of regulation in inducible operons". In: *Nature* 252.5484 (1974), pp. 546–549.

[202] J. Schäfer and K. Strimmer. "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics". In: *Stat Appl Genet Mol Biol* 4 (2005). DOI: 10.2202/1544-6115.1175.

[203] M. Schilling, T. Maiwald, S. Hengl, D. Winter, C. Kreutz, W. Kolch, W. D. Lehmann, J. Timmer, and U. Klingmüller. "Theoretical and experimental analysis links isoform-specific ERK signalling to cell fate decisions". In: *Mol Syst Biol* 5.1 (2009), p. 334.

[204] R.-P. Scholz, J. Regner, A. Theil, P. Erlmann, G. Holeiter, R. Jähne, S. Schmid, A. Hausser, and M. A. Olayioye. "DLC1 interacts with 14-3-3 proteins to inhibit RhoGAP activity and block nucleocytoplasmic shuttling". In: *J Cell Sci* 122.1 (2009), pp. 92–102.

[205] G. E. Schwarz. "Estimating the dimension of a model". In: *Ann Stat* 6.2 (1978), pp. 461–64. DOI: 10.1214/aos/1176344136.

[206] P. Segaert, M. B. Lopes, S. Casimiro, S. Vinga, and P. J. Rousseeuw. "Robust identification of target genes and outliers in triple-negative breast cancer data". In: *Stat Methods Med Res* (2018). PMID: 30146936, p. 0962280218794722. DOI: 10.1177/0962280218794722.

[207] S. Serneels and I. Hoffmann. *sprm: Sparse and Non-Sparse Partial Robust M Regression and Classification*. R package version 1.2.2. 2016.

[208] A. Sharma, L. H. Boise, and M. Shanmugam. "Cancer Metabolism and the Evasion of Apoptotic Cell Death". In: *Cancers* 11.8 (2019). DOI: 10.3390/cancers11081144.

[209] X. Shi. "A Hill type equation can predict target gene expression driven by p53 pulsing". In: *FEBS Open Bio* 11.6 (2021), pp. 1799–1808. DOI: 10.1002/2211-5463.13179.

[210] S.-Y. Shin, O. Rath, S.-M. Choo, F. Fee, B. McFerran, W. Kolch, and K.-H. Cho. "Positive-and negative-feedback regulations coordinate the dynamic behavior of the Ras-Raf-MEK-ERK signal transduction pathway". In: *J Cell Sci* 122.3 (2009), pp. 425–435.

[211] M. Singla and K. Shukla. "Robust statistics-based support vector machine and its variants: a survey". In: *Neural Comput Appl* 32.15 (2020), pp. 11173–11194.

[212] P. Smolen, D. A. Baxter, and J. H. Byrne. "Bistable MAP kinase activity: a plausible mechanism contributing to maintenance of late long-term potentiation". In: *Am J of Physiol Cell Physiol* 294.2 (2008), pp. C503–C515.

[213] H. K. Sok, M. P.-L. Ooi, and Y. C. Kuang. "Sparse alternating decision tree". In: *Pattern Recognit Lett* 60-61 (2015), pp. 57–64. DOI: 10.1016/j.patrec.2015.03.002.

[214] J. Song, J. Li, A. Lulla, B. M. Evers, and D. H. Chung. "Protein kinase D protects against oxidative stress-induced intestinal epithelial cell injury via Rho/ROK/PKC-δ pathway activation". In: *Am J Physiol Cell Physiol* 290.6 (2006), pp. C1469–C1476.

[215] S. Srinivas, A. Subramanya, and R. Venkatesh Babu. "Training Sparse Neural Networks". In: *Proc IEEE CVPR Workshops*. 2017.

[216] S. J. Stehbens, M. Paszek, H. Pemble, A. Ettinger, S. Gierke, and T. Wittmann. "CLASPs link focal adhesion-associated microtubule capture to localized exocytosis and adhesion site turnover". In: *Nat Cell Biol* 16.6 (2014), p. 561.

[217] H. Steinhaus. "Sur la division des corps matériels en parties". In: *Bull. Acad. Polon. Sci* 1.804 (1956), p. 801.

[218] J. Stelling, U. Sauer, Z. Szallasi, F. J. Doyle III, and J. Doyle. "Robustness of cellular functions". In: *Cell* 118.6 (2004), pp. 675–685.

[219] J. D. Storey. "A direct approach to false discovery rates". In: *J R Stat Soc Series B Stat Methodol* 64.3 (2002), pp. 479–498.

[220] J. D. Storey and R. Tibshirani. "Statistical significance for genomewide studies". In: *Proc Natl Acad Sci* 100.16 (2003), pp. 9440–9445.

[221] P. Storz, A. Hausser, G. Link, J. Dedio, B. Ghebrehiwet, K. Pfizenmaier, and F.-J. Johannes. "Protein kinase C μ is regulated by the multifunctional chaperon protein p32". In: *J Biol Chem* 275.32 (2000), pp. 24601–24607.

[222] T. Sumner. "Sensitivity analysis in systems biology modelling and its application to a multi-scale model of blood glucose homeostasis". PhD thesis. UCL (University College London), 2010.

[223] H. Sun, Y. Cui, H. Wang, H. Liu, and T. Wang. "Comparison of methods for the detection of outliers and associated biomarkers in mislabeled omics data". In: *BMC Bioinformatics* 21.1 (2020), p. 357. DOI: 10.1186/s12859-020-03653-9.

[224] J. Sutherland, Y. Webster, J. Willy, G. Searfoss, K. Goldstein, A. Irizarry, D. Hall, and J. Stevens. "Toxicogenomic module associations with pathogenesis: a network-based approach to understanding drug toxicity". In: *Pharmacogenomics J* 18.3 (2018), pp. 377–390.

[225] A. Tharwat. "Classification assessment methods". In: *Appl Comput Inform* (2020).

[226] R. Thomas. “On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations”. In: *Numerical methods in the study of critical phenomena*. Springer, 1981, pp. 180–193.

[227] R. S. Thomas et al. “Temporal Concordance Between Apical and Transcriptional Points of Departure for Chemical Risk Assessment”. In: *Toxicol Sci* 134.1 (2013), pp. 180–194. DOI: 10.1093/toxsci/kft094.

[228] C. Thomaseth and N. Radde. “Normalization of Western blot data affects the statistics of estimators”. In: *IFAC-PapersOnLine* 49.26 (2016), pp. 56–62. DOI: 10.1016/j.ifacol.2016.12.103.

[229] D. A. Thompson and R. J. Weigel. “hAG-2, the Human Homologue of theXenopus laevisCement Gland Gene XAG-2, Is Coexpressed with Estrogen Receptor in Breast Cancer Cell Lines”. In: *Biochem Biophys Res Commun* 251.1 (1998), pp. 111–116. DOI: 10.1006/bbrc.1998.9440.

[230] R. Tibshirani. “Regression shrinkage and selection via the lasso”. In: *J R Stat Soc Series B Methodol* 58.1 (1996), pp. 267–288.

[231] K. Tomczak, P. Czerwińska, and M. Wiznerowicz. “The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge”. In: *Contemp Oncol (Pozn)* 19.1A (2015), A68–77. DOI: 10.5114/wo.2014.47136.

[232] R. Tomioka, K. Aihara, and K.-R. Müller. “Logistic regression for single trial EEG classification”. In: *Adv Neural Inf Process Syst* 19 (2007), p. 1377.

[233] D. Vaudry, P. Stork, P. Lazarovici, and L. Eiden. “Signaling pathways for PC12 cell differentiation: making the right connections”. In: *Science* 296.5573 (2002), pp. 1648–1649.

[234] G. Venter. “Review of Optimization Techniques”. In: *Encyclopedia of Aerospace Engineering*. American Cancer Society, 2010. DOI: 10.1002/9780470686652.eae495.

[235] A. Veríssimo. *brca.data: BRCA gene expression and clinical data from TCGA (with import script)*. R package version 1.0. 2019.

[236] A. Wagner. "Circuit topology and the evolution of robustness in two-gene circadian oscillators". In: *Proc Natl Acad Sci* 102.33 (2005), pp. 11775–11780.

[237] Y. Wakana, J. Van Galen, F. Meissner, M. Scarpa, R. S. Polishchuk, M. Mann, and V. Malhotra. "A new class of carriers that transport selective cargo from the trans Golgi network to the cell surface". In: *EMBO J* 31.20 (2012), pp. 3976–3990.

[238] S. Wang, Z. Zheng, P. Chen, and M. Wu. "Tumor classification and biomarker discovery based on the 5'isomiR expression level". In: *BMC cancer* 19.1 (2019), pp. 1–10.

[239] J. Watkins et al. "Genomic Complexity Profiling Reveals That HORMAD1 Overexpression Contributes to Homologous Recombination Deficiency in Triple-Negative Breast Cancers". In: *Cancer Discov* 5.5 (2015), pp. 488–505. DOI: 10.1158/2159-8290.CD-14-1092.

[240] P. Weber, M. Hornjik, M. Olayioye, A. Hausser, and N. Radde. "A computational model of PKD and CERT interactions at the trans-Golgi network of mammalian cells". In: *BMC Syst Biol* 9.9 (2015).

[241] Q. Wei and R. L. Dunbrack Jr. "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics". In: *PLoS One* 8.7 (2013), pp. 1–12. DOI: 10.1371/journal.pone.0067863.

[242] R. A. Weinberg. "How Cancer Arises". In: *Sci Am* 275.3 (1996), pp. 62–70.

[243] D. M. Witten and R. Tibshirani. "A Framework for Feature Selection in Clustering". In: *J Am Stat Assoc* 105.490 (2010). PMID: 20811510, pp. 713–726. DOI: 10.1198/jasa.2010.tm09415.

[244] S. Wold, M. Sjöström, and L. Eriksson. "PLS-regression: a basic tool of chemometrics". In: *Chemometr Intell Lab Syst* 58.2 (2001). PLS Methods, pp. 109–130. DOI: 10.1016/S0169-7439(01)00155-1.

[245] W. Xiong and J. E. Ferrell. "A positive-feedback-based bistable 'memory module'that governs a cell fate decision". In: *Nature* 426.6965 (2003), pp. 460–465.

[246] J. Yang, Y. Liu, C. Feng, and G. Zhu. "Applying the Fisher score to identify Alzheimer's disease-related genes". In: *Genet Mol Res* 15.2 (2016).

[247] A. D. Yates et al. "Ensembl 2020". In: *Nucleic Acids Res* 48.D1 (2019), pp. D682–D688. DOI: 10.1093/nar/gkz966.

[248] J. Yuan, L. W. Slice, and E. Rozengurt. "Activation of protein kinase D by signaling through Rho and the α subunit of the heterotrimeric G protein G13". In: *J Biol Chem* 276.42 (2001), pp. 38619–38627.

[249] Q. Zhai, H. Li, L. Sun, Y. Yuan, and X. Wang. "Identification of differentially expressed genes between triple and non-triple-negative breast cancer using bioinformatics analysis". In: *Breast Cancer* 26.6 (2019), pp. 784–791. DOI: 10.1007/s12282-019-00988-x.

[250] X. Zhang, Y. Yap, D. Wei, G. Chen, and F. Chen. "Novel omics technologies in nutrition research". In: *Biotechnol Adv* 26.2 (2008), pp. 169–176. DOI: 10.1016/j.biotechadv.2007.11.002.

[251] H. Zou and T. Hastie. "Regression shrinkage and selection via the elastic net, with applications to microarrays". In: *JR Stat Soc Ser B* 67 (2003), pp. 301–20.

Abstract

Biological systems are complex and diverse. Learning about and understanding these systems is nowadays not only based on experimental observations but often also involves mathematical modeling.

In this thesis a workflow for data-based modeling in the context of cancer biology, with a particular focus on sparse data, is described. Data pre-processing, system modeling, model calibration, model validation, and model analysis constitute the five workflow steps. This workflow is applied to three different biological systems. While the first project investigates a feedback mechanism of the known MAPK pathway, the second project analyzes the role of the tumor suppressor protein DLC1 in regulating PKD activity at the Golgi. Finally, the third project gives insight into the genetic composition characterizing triple-negative breast cancer in contrast to other breast cancer types. In the first two projects systems biology approaches were employed whereas the last is based on a classification approach.

All systems studied here are confronted with sparse data. In the first two systems, the sparsity is characterized by a low time resolution, measurements of only a subset of the components, large variability between replicates, and relative measurements, generating uncertainty in the model parameters when calibrating the model. This problem is addressed with a combination of statistical methods, which allow the propagation of uncertainty in the model parameters to the model predictions. The sparsity in the third system manifests in a large feature space compared to the number of samples. As most non-sparse methods assume that the number of samples exceeds the number of features they may overfit the training data or fail completely. Consequently, an ensemble integrating sparse and robust methods for feature selection and outlier identification is proposed.