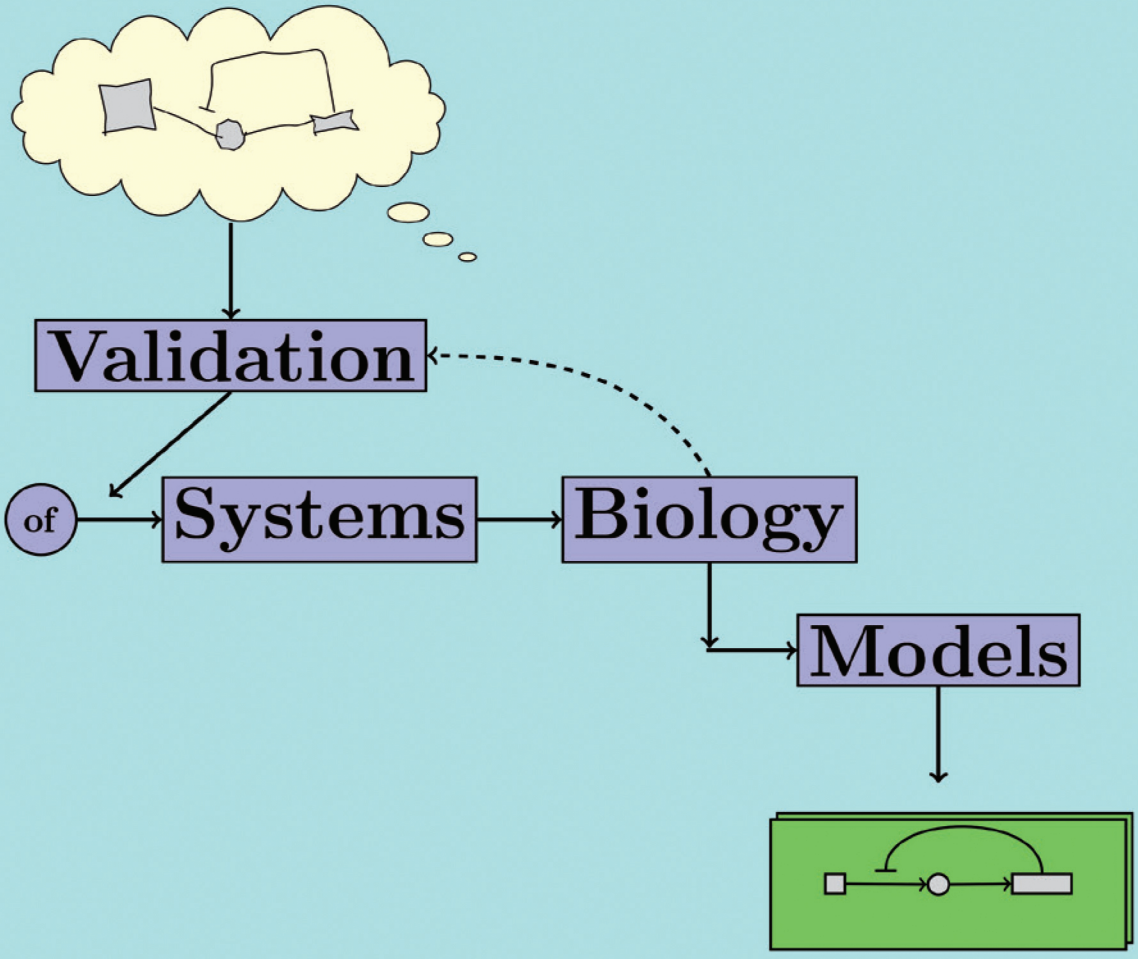

Dicle Hasdemir

Validation of systems biology models

Dicle Hasdemir

Validation of systems biology models
Dicle Hasdemir
PhD thesis
ISBN–978-94-6299-092-0

Printed by: Ridderprint BV
Cover illustration by: Dicle Hasdemir
Cover design by: Ceylan Çölmekçi Öncü

Validation of systems biology models

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D. C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 9 juni 2015, te 14:00 uur

door

Dicle Hasdemir
geboren te Diyarbakır, Turkije

Promotiecommissie:

Promotor:	Prof. dr. A. K. Smilde	Universiteit van Amsterdam
Copromotor:	Dr. H. C. J. Hoefsloot	Universiteit van Amsterdam
Overige leden:	Prof. dr. L. Buydens	Radboud Universiteit
	Prof. dr. J. Heringa	Vrije Universiteit Amsterdam
	Prof. dr. A. H. C. van Kampen	Universiteit van Amsterdam
	Prof. dr. B. Kırdar	Boğaziçi University
	Dr. G. J. Smits	Universiteit van Amsterdam
	Prof. dr. H. V. Westerhoff	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The research reported in this thesis was carried out at the Swammerdam Institute for Life Sciences, Faculty of Science, University of Amsterdam and was financed by the Netherlands Metabolomics Centre, which is part of the Netherlands Genomics Initiative (NGI).

To my mother Pürnaz

Contents

1

Introduction

Systems biology is a highly inter-disciplinary modern approach to the analysis of living systems. It employs an integrative approach where the focuses are the interactions between biological entities and their operation as a system. The typical systems biology approach incorporates modeling as a fundamental tool. It utilizes various modeling approaches and computational tools from a large selection of fields such as statistics, control theory and mathematics.

With the advancements in data acquisition and the increasing recognition of the systems biology concept, a substantial number of models have been introduced that aim to explain various biological systems. Nevertheless, both the structures and the parameters of these models are prone to uncertainty. This arises from the fact that the data used to obtain the models are noisy and the biological knowledge regarding the systems under study is incomplete. This brings the need for careful model validation and selection. However, the importance of validating models is often underestimated in the field. The major reason is that the employment of the modeling concept itself has recently started to become such prevalent and the concept of validation is still one step behind. While we witness the maturation of systems biology models, guidelines that are tailored for their validation has to be introduced in the field. In this respect, resampling strategies are worth attention.

Resampling strategies have been extensively used to assess the validity and stability of models and to infer the precision of estimated model parameters in many fields. These include but are not limited to statistics, chemometrics and machine learning. Cross validation and bootstrapping, two well known resampling approaches have

been applied previously also in systems biology. However, they deserve more attention and can be exploited further. When appropriately adapted for the characteristics of the data and for the models encountered in the field, they can provide the guidelines needed for model validation and selection.

The goals of this thesis are:

- *Developing methods for the assessment of the reliability and informative levels of models,*
- *Presenting guidelines for reliable model validation and selection,*
- *Finally, incorporating resampling strategies for the purposes above regarding extensively employed systems biology models.*

In Sections 1.1 and 1.2 we present background information regarding the types of systems biology models that are the focus of this thesis and the resampling strategies that have been applied in the systems biology concept, respectively. In Section 1.1, we also detail our goals that are specific to each type of model. In Section 1.2, we have a special focus on application examples of bootstrapping and cross validation in systems biology. We exclude the application of bootstrapping for cluster analysis and the application of cross validation for the analysis of ODE based kinetic models, since we discuss them further in the introductory sections of Chapters 5 and 3, respectively.

1.1. Common modeling approaches in systems biology

Models in systems biology can be categorized in two basic classes. The first type is a top-down model which is constructed by the analysis and summarization of large scale data such as transcriptomics, proteomics and metabolomics data. Examples to these are network models inferred from large scale data or statistical models such as cluster analysis. The other type of model is a bottom-up model that is built upon physical and biochemical knowledge regarding the individual parts of the system at a molecular level [25]. Kinetic nonlinear models of biochemical systems are typical examples of this type. Either type of models are important means for investigating the biological system as a whole and hence, are pivotal to systems biology.

1.1.1. ODE based kinetic models

Modeling using differential equations is commonly applied in systems biology for deterministic systems explained by biochemical laws. Primary examples to these systems are signaling and metabolic pathways. In these models, concentration behavior of biochemical species (state variables) is governed by mathematical formulations which follow the laws of biochemical kinetics such as Michaelis-Menten, mass action or Hill kinetics. The independent variable in the model is often time, additionally complemented with location in the cell leading to partial differential equation models when the spatial distribution of the variables is also important. When the independent variable is only the time, ordinary differential equations (ODE) are sufficient to describe a system deterministically.

Model parameters are kinetic constants used in the formulation of the kinetics involved. When these parameters are known, the time dependent behavior of the state variables can be achieved by solving the ODE system and later, can be used for predictive purposes. Theoretically, these parameter values can be determined by *in vitro* experiments with isolated enzymes. However, not all of them are applicable under the *in vivo* conditions of the cell. Therefore, *in vivo* time series measurements of the species that are involved in the model are commonly used to infer the unknown parameters [10]. This is accomplished by minimizing the difference between the measurements and the concentration values obtained by the model. During parameter inference by this approach, a fixed model structure is assumed to be the correct structure. In other words, the components of the model and the governing kinetic laws are assumed to be known. However, this assumption can never be entirely fulfilled. Therefore, modelers have to keep in mind the uncertainty of the model structure such as the possibility of putative feedback loops or allosteric regulation [100, 128].

This brings the need for model validation and selection. Proper handling of the data for parameter inference and model validation tasks is essential. In Chapters 2 and 3, we develop guidelines for this purpose by exploiting cross validation.

1.1.2. Network models

Graph theory provides a framework in the systems biology perspective for the analysis of interactions between entities. The framework represents the biological system as a network. Depending on the system studied, the nodes in the network are often proteins such as transcription factors or enzymes, genes and metabolites.

Protein-protein interaction networks are one of the most studied network type which include proteins as nodes and the physical interactions between the proteins

as edges [144]. Often, the interactions in these networks are nondirectional such as the cases of protein complexes and hence, no directionality has to be imposed on the edges. Protein-protein interaction networks are usually obtained by combining information from different sources such as yeast two hybrid experiments, co-immunoprecipitation assays and computational prediction of interactions by structural modeling. There are also studies that model signaling pathways using directional protein interaction networks and employ graph theoretical analysis [8, 41].

An example network model where directionality has to be imposed is a metabolic network model. In metabolic networks, the edges usually correspond to chemical reactions between the metabolites [119]. In other example networks, both enzymes and metabolites can be the nodes. Then, the edges imply the enzyme's role in the chemical reactions producing/consuming the metabolites it is connected to [120]. Network modeling is also used to summarize top-down information and build correlation networks. In these, metabolites with similarities in terms of their steady state or time-resolved behavior can be linked [28, 62]. In a similar way, gene networks may also be constructed by linking similarly expressed genes to each other [87]. However, the term transcriptional regulatory networks or transcription networks have traditionally been reserved for another type of network model of genes which we explain later.

Biological networks have certain properties that discriminate them from any random network. Most importantly, they are scale-free, a property that characterizes the degree distribution of the nodes [3]. Degree of a node means the number of nodes it is linked to, in other words, its neighbors. As a result of being scale-free, biological networks typically contain a small number of nodes with very high degree, called the hubs and a large number of nodes with low degree. Many studies focus on identifying the topological properties of biological networks such as the hubs of the network, the mean degree in the network, average distance between the nodes (diameter) and the distribution of the clustering coefficients. In such studies where the focus is on static analysis, the network motifs are also crucial. Network motifs are local patterns in the network that occur significantly common compared to other patterns [96]. For example, motifs in metabolic networks may correspond to abundant reaction types or motifs in a protein interaction network can reveal abundant protein complex structures. Identifying the topological features of a network increases our understanding of the big picture of the biological systems that we investigate. Network based modeling framework is not restricted to static analysis, though. A Boolean network model is an example of network modeling approach that links time points and allows discrete dynamic modeling [163].

Transcriptional regulatory networks

Transcriptional regulatory networks consist of transcription factors, genes and the regulatory interactions between those [12, 96]. Transcription factors are proteins that bind to the promoter regions upstream of the coding regions of the genes. They regulate the expression of genes by either promoting or repressing the transcription of the mRNA. Post-transcriptional modification of gene expression is also possible by micro-RNAs which bind to the transcribed mRNA itself [45]. However, post-transcriptional modification is not a part of these models since the focus here is in the regulation of transcription.

Protein-DNA interactions can be represented by adjacency matrices. Such matrices can be binary in which a 1 is assigned in the respective place if a certain transcription factor regulates the transcription of a certain gene. Adjacency matrices can also consist of numbers which correspond to the strength of the interaction between transcription factors and genes. The strength of the interaction affects the degree of the influence that the transcription factor has on the regulation of the transcription [99].

Physical interactions between the DNA and the binding proteins are detected by a series of different techniques. Most importantly, chromatin immunoprecipitation experiments coupled with microarrays (ChIP-Chip) or coupled with high throughput sequencing (ChIP-seq) allow the genome wide detection of DNA binding. Other techniques involve yeast one hybrid systems or X-ray crystallography. Sequence based computational prediction is also possible. Prior information such as co-clustering of genes can be incorporated in such an approach. As an example, sequence based analysis in the upstream regions of co-clustered genes was shown to reveal putative binding sites in the yeast genome [145].

High throughput approaches employed in the discovery of transcriptional regulatory interactions lead to substantial amount of uncertainty in the proposed network structures due to experimental artifacts. Therefore, selection between different networks and refinement of the topological structures both computationally and experimentally are of high importance. In Chapter 4, we deal with these challenges by exploiting the detection of the inconsistencies between the network topology and the associated expression data of the genes in the network.

1.1.3. Cluster analysis

Cluster analysis is a data analysis method used for grouping similar objects together in an unsupervised manner. Similarity is measured by employing a distance metric. An appropriate distance metric has to be chosen depending upon the aim of the cluster analysis and the structure of the data. Distance metrics are either geomet-

rically defined such as the Euclidean distance, corrected for correlation between the variables as the Mahalanobis distance or based on correlation such as the Pearson correlation coefficient.

A well known clustering algorithm encountered in the analysis of biological data is hierarchical clustering [44, 84]. Hierarchical clustering depend on the pairwise similarity between the object across different variables. Hierarchical clustering allows the formation of different number of clusters at different hierarchical levels: in other words, at different levels of distances between the objects. At the highest level of the hierarchy, there is a single cluster that contains all the objects. Visual inspection of the hierarchy between the clusters is possible with a dendrogram presentation, making it popular for biological applications.

Many other clustering algorithms fall under the partitioning algorithms class where the distance of an object to the cluster centroid is important rather than the pairwise similarity of hierarchical clustering. A major example is k-means clustering in which k clusters are formed through the minimization of an objective function. The objective function is often the sum of squares of the distances of the objects to the centroids of the clusters. Further assumptions about the clusters can be made regarding the distribution of the objects in the clusters, leading to model based-clustering. Gaussian mixture models clustering is one of the most fundamental model-based clustering algorithms in which data is assumed to consist of Gaussian clusters defined by a mean vector and a covariance matrix.

Fuzzy clustering is another type of algorithm applied in systems biology applications. In this framework, clusters are not necessarily distinct. Therefore, objects can be assigned to multiple clusters with varying levels of membership [51]. Network based clustering approaches have also been popularly applied for the detection of densely connected nodes in network models [11].

Clustering is extensively used in the analysis of genome wide gene expression data. Genes are grouped together based on the similarities of their expression levels across different experimental conditions, patients or time points. Co-clustering genes give hints on the common regulatory rules acting on them, shedding light on their functional characteristics. A cluster analysis is often followed by a functional enrichment analysis which identifies the most dominant functional categories in each cluster.

Usually, the cluster analysis is assumed to be useful if the results from the following enrichment analysis lead to biologically meaningful explanations. However, the unbiased assessment of the validity of a cluster analysis without depending upon biological information is essential. In Chapter 5, we deal with the validation of k-means clustering analysis through the assessment of its stability.

1.2. Resampling strategies

1.2.1. Bootstrapping

Bootstrapping which was originally proposed by Efron [42] is employed to infer the precision of statistical estimates such as confidence intervals, variance and bias. It provides answers in cases where analytically derived formulas do not exist due to complex estimation procedures. It is also helpful in situations where the assumptions regarding the underlying distribution of the data needed for traditional statistical inference can not be fulfilled. The assumptions behind bootstrapping are rather relaxed [43, 76, 115], making its application suitable for such a situation. There is substantial literature regarding bootstrapping that constructs guidelines for applied statisticians on its implementation [42, 43, 134, 165].

Bootstrapping can be summarized as follows. Data actually consist of observed random samples from an unknown probability distribution. Parameters of a model (e.g. kinetic parameters in an ODE model) are estimated using the data. The variability of a parameter estimate around the true value of that parameter is mimicked by the variability of the parameter estimates inferred from the bootstrap samples around the parameter estimate inferred from the original data (observed value of the parameter). This is the basic idea behind using bootstrapping to calculate the precision of statistical estimates. In other words, bootstrapping provides us a means of repeating the random sampling process with replacement. How do we then construct the so-called bootstrap samples? In nonparametric bootstrapping, bootstrap samples are drawn from an empirical distribution of the samples which represents the true unknown probability distribution of the data. In parametric bootstrapping, a certain underlying distribution is assumed to be true and bootstrap samples are drawn from this parametric estimate of the population.

The most critical step in bootstrapping is obtaining the appropriate empirical distribution or the appropriate parametric estimate for an unknown probability distribution. This can be stated as obtaining the appropriate bootstrap samples. Bootstrap sampling needs more tedious work and is more error-prone when complicated and dependent data structures are involved [43, 64]. An example is time series data which is widely encountered in biological modeling. In the rest of Section 1.2.1, we summarize the application areas of bootstrapping in systems biology with a focus on the techniques used for obtaining bootstrap samples of time series data.

1

Bootstrapping for obtaining parameter confidence intervals in ODE based models

Data taken at different time points depend on each other. Therefore, bootstrapping has to be adapted accordingly. One approach is to bootstrap from the residuals in a parametric way [76] since unlike the data, the residuals are not essentially dependent. In this approach, the residuals are bootstrapped from a parametric distribution with zero mean and constant standard deviation. The standard deviation of the residuals at a data point is equal to the standard deviation inferred from the replicates of data at that point. Then, the bootstrapped residuals are added to the means of the replicate data to obtain a bootstrap data point. This procedure gives a set of new data points for the biochemical species over time which constitutes one single bootstrap sample. In this way, a sufficient number of bootstrap samples which would be enough for convergence are created. Convergence is detected by plotting the mean of parameters. If the means of parameters which are obtained by different bootstrap samples are not affected anymore by increasing the number of samples, then the samples are assumed to have converged. 1000 to 3000 samples were reported to be sufficient in a number of studies, but it depends on the complexity of the model [76, 82, 130]. Later, models are fitted to the bootstrap datasets to obtain the bootstrap estimates. The confidence intervals (CI) of parameters are then calculated from the percentiles of the bootstrap estimates. Another alternative for the last step is the use of the bias-corrected accelerated (BCa) intervals since they exhibit better compatibility with exact intervals [33]. The details are outlined in Figure 1.1. In [76], the approach was proved to be valid by showing narrower confidence intervals resulting from a traditional D-optimal experimental design that aims at better estimation of the parameters.

An essential step in the approach outlined above is to identify the type of the error in the data. For example, in [33], an error model in which error is dependent on the log transformed data is used. This is apparently dependent on the knowledge of the experimental procedure by which the data was obtained. In general, an error model where the error is dependent on the mean of the data would be appropriate for biochemical assays.

The bootstrapping approach is superior to the traditional Fisher information matrix (FIM) based approach from certain aspects. Calculation of the FIM requires linearization of the system which can result in inaccurate findings for a nonlinear ODE model. Theoretical confidence intervals obtained by the FIM method are usually unrealistically smaller and essentially symmetric [76, 129]. However, bootstrapping gives more realistic intervals.

Another way of bootstrapping time series data is based on modeling the depen-

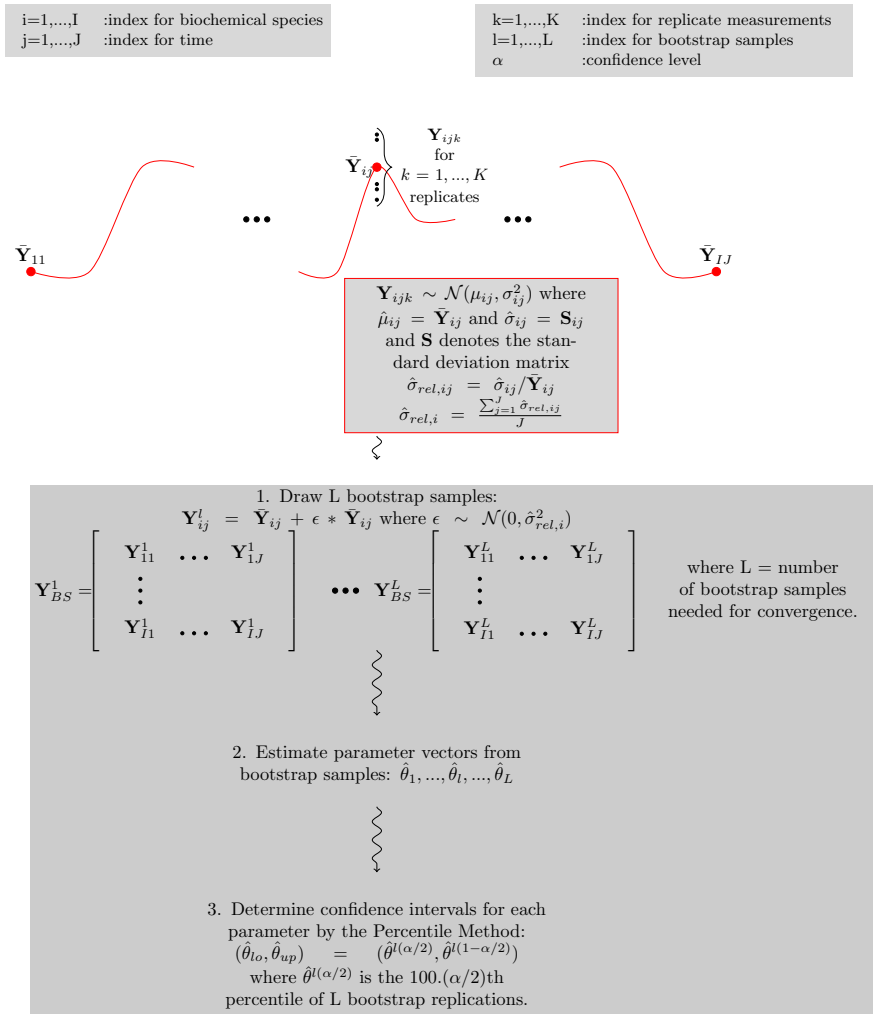

Figure 1.1: **Calculation of confidence intervals by bootstrapping.** The approach that is visually outlined here can be considered as a basic application guideline of bootstrapping tailored for nonlinear ODE models and time series data.

dence in the data. One way to deal with it is using the Gaussian process regression method [82]. This nonlinear regression method from the class of Bayesian techniques does not assume a fixed form of the regression function. It models the underlying time dependent function as a collection of random variables with a mean and a covariance function. The priors for the covariance function is especially important in using such a method since it presents the knowledge on how the data at different

time points co-vary with each other. Once the underlying function is estimated, bootstrap samples can be drawn from this distribution and confidence intervals can be calculated by repeated parameter inference on the bootstrap samples.

Bootstrapping for model comparison in ODE based models

The following major step after parameter estimation is to test whether the model explains the data well enough. χ^2 tests can be used to assess the quality of the data-model fit [29]. The hypothesis test will help to decide if the residuals from a model are small enough for the model to be accepted as adequate. This is the first step in which a model can be rejected and the modeler is forced to come up with an alternative model structure. Here, the test statistic that is expected to come from a χ^2 distribution is the likelihood. Likelihood of a model is the probability of the model to be the true data generating process given the observed data. Model residuals may usually be used instead of likelihood for practical purposes. This replacement is valid also for likelihood ratio tests which are used to compare two models.

Likelihood ratio tests are used to assess the superiority of a given model against an alternative model. The test statistic is the observed difference of the residuals from the two models and follows a χ^2 distribution if the two models are nested and linear. However, the models that are being compared might not be nested. More fundamentally, dynamic ODE based models are typically nonlinear. There are also additional issues such as limited data availability, lack of full parameter identifiability and positivity constraints on the parameters that violate the Gaussian distribution of the parameters. These issues are encountered commonly in this model type [115]. Therefore, there is no clue on the nature of the distribution from which the observed test statistic comes. Sometimes, the distribution is even a mixture of multiple distributions. Therefore, it has to be determined empirically. This can be achieved by bootstrapping.

There is a consensus guideline for the application of bootstrapping in constructing the empirical distribution [65, 148, 160, 169]. However obtaining the empirical distribution comes at a price. Unlike non-bootstrapped likelihood ratio test, the bootstrapped version can not provide a full model selection process because both models can be selected or both can be rejected following the test [29]. Consequently, the model comparison task turns into two parallel model rejection tasks. However the test is still advantageous. It can have more power compared to a single χ^2 test for the rejection of a single model if the alternative model has certain characteristics such as being not too flexible and not too rigid [74]. The alternative model is called a 'help model' and it favors the testing of the model that is primarily in question.

Simulations using a two dimensional χ^2 test in [74] demonstrates this issue.

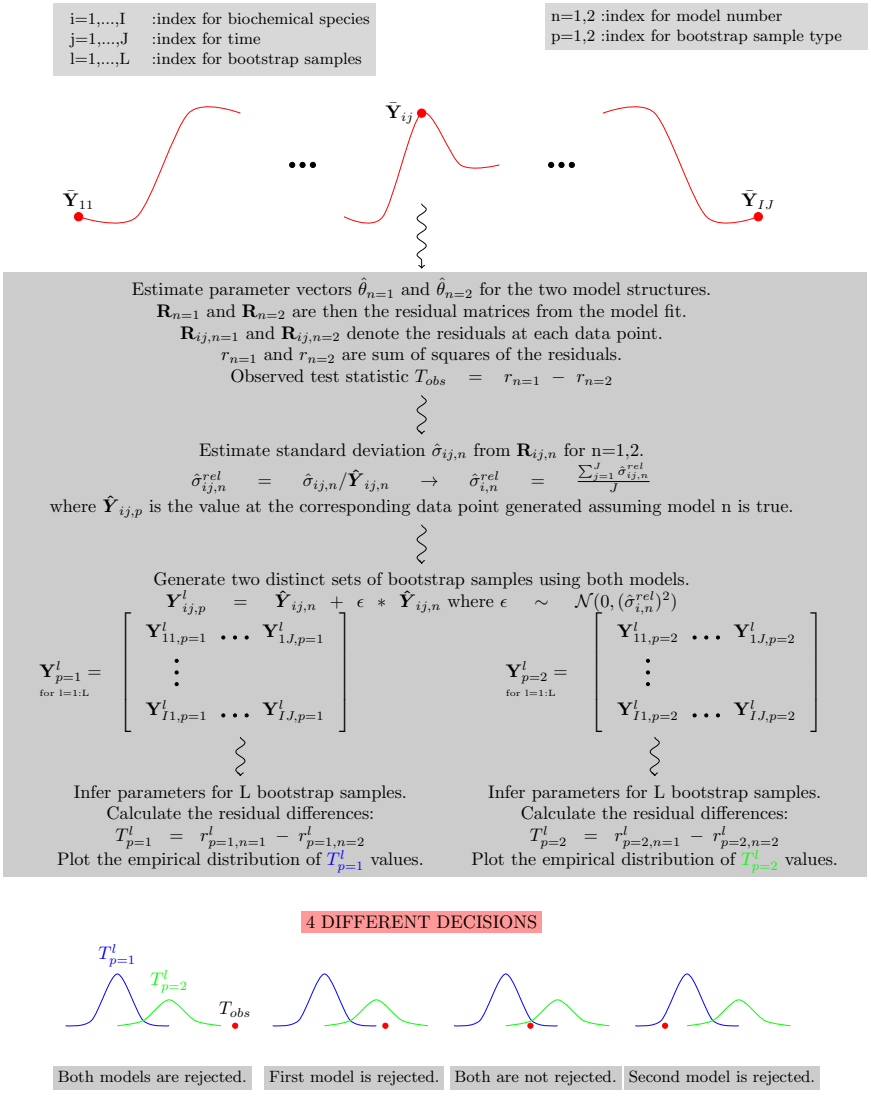

Figure 1.2: **Model rejection/comparison using bootstrapping.** The approach that is visually outlined here can be considered as a basic application guideline of bootstrapping for the comparison/rejection of ODE based models.

According to the consensus approach, two distinct sets of bootstrap samples are generated by assuming each model is the true model in turn. Fitting both models to both sets of samples and calculating the residual differences of two models for each

bootstrap sample gives us the two empirical distributions of residual differences. The final decision can then be given by comparing the observed residual difference which was inferred from the original data to the two distributions. In Figure 1.2, we describe this approach which has been reported to perform better than alternative bootstrapping strategies [57, 74].

Bootstrapping for detecting the reliability of network models

Bootstrapping has been applied to assess the reliability of network models, particularly networks of genes [69, 81, 82]. An example in this respect [82] deals with two types of correlation networks. The first type is a gene relevance network which include edges based on the correlation between genes' expression profiles. The second type is a graphical Gaussian model which include edges based on partial correlation. Both networks are inferred from time series data. Nevertheless, the inferred topology is not completely robust to uncertainty in the data that results from experimental noise. Gaussian process regression bootstrapping on time series data is employed to quantify the uncertainty in the topology by calculating the confidence intervals for the inferred edges, as previously explained also for the ODE based models. Another example[69] employs employ bootstrapping for robust estimation of regulatory interactions in a Bayesian network which can be considered as a probabilistic gene network.

1.2.2. Cross validation

Cross validation has been used in systems biology studies in which large scale data are analyzed using supervised methods. Compared to bootstrapping, its application area has been traditionally narrower in the field but it has been more commonly applied. We see primary applications in gene expression data analysis.

Cross validation for classification

Building classifiers using gene expression data is extensively applied in especially cancer research. In this research area, genome wide mRNA levels are selectively used as features for classifiers which are trained to classify sample tissues in terms of prognosis and future possibility of metastasis. Testing the validity of these classifiers on independent data leads to reliable biomarkers for diseases. This can be accomplished by using cross validation. Various cross validation techniques have been employed such as k-fold, leave-one-out, hold-out and Monte-Carlo cross validation.

In leave-one-out cross validation (LOOCV), each sample is left out of the training step once and the discrimination performance of the classifier is tested on the left

out sample. LOOCV has been reported to contribute to the selection of biomarkers and classification of tissues, following a support vector machines classification to select lymphoma subsets [121], following a random forests classification [35] or to predict metastases in breast cancer [156].

Guidelines for the appropriate application of cross validation have also been elaborately discussed. The authors of [139] present adjustments to the technique to correct for the problems arising from small sample size. Furthermore, the bias-variance behavior of different cross validation methods have been documented with applications on various types of supervised classification [113].

In addition to gene expression, large scale proteomics and metabolomics studies have benefited from cross validation based approaches. Double cross validation utilizes an additional cross validation loop within the outer cross validation loop and thus provides the complete separation of the test set samples from those used to train the meta-parameters such as the number of components needed in a principal component analysis model [141]. It has been applied in proteomics and metabolomics studies following supervised classification by using principal component discriminant analysis [141] and partial least squares discriminant analysis [167], respectively.

2

Assessing the informative levels of kinetic models

Kinetic models can present mechanistic descriptions of molecular processes within a cell. They can be used to predict the dynamics of metabolite production, signal transduction or transcription of genes. Although there has been tremendous effort in constructing kinetic models for different biological systems, not much effort has been put into their validation. In this study, we introduce the concept of resampling methods for the analysis of kinetic models and present a statistical model invalidation approach. We based our invalidation approach on the evaluation of a kinetic model's predictive power through cross validation and forecast analysis. As a reference point for this evaluation, we used the predictive power of an unsupervised data analysis method which does not make use of any biochemical knowledge, namely Smooth Principal Components Analysis (SPCA) on the same test sets. Through a simulations study, we showed that too simple mechanistic descriptions can be invalidated by using our SPCA-based comparative approach until high amount of noise exists in the experimental data. We also applied our approach on an eicosanoid production model developed for human and concluded that the model could not be invalidated using the available data despite its simplicity in the formulation of the reaction kinetics. Furthermore, we analyzed the high osmolarity glycerol (HOG) pathway in yeast to question the validity of an existing model as another realistic demonstration of our method. With this study, we have successfully presented the potential of two resampling methods, cross validation and forecast analysis in the analysis of kinetic

models' validity. Our approach is easy to grasp and to implement, applicable to any ordinary differential equation (ODE) type biological model and does not suffer from any computational difficulties which seems to be a common problem for approaches that have been proposed for similar purposes. Matlab files needed for invalidation using SPCA cross validation and our toy model in SBML format are provided at http://www.bdagroup.nl/content/Downloads/software/software.php. [1]

2.1. Background

With the concept of 'systems biology' coming to the stage of biological research, construction of kinetic models has been the primary focus in a substantial number of studies [37, 54, 86, 146]. Kinetic models are mechanistic representations of biological systems. They include information on two main levels. The first level of information includes the metabolites, enzymes, signaling molecules and chemical reactions involved in the model together with the formulation of the reaction kinetics such as Michaelis-Menten kinetics. Knowledge about inhibition, activation and allosteric regulation of enzymes are also a part of this level. The second level of information consists of numerical values of all different parameters defined in the first level of information. Those parameters include but are not limited to rate parameters for chemical reactions such as production of new metabolites in metabolic models, post-translational modifications of proteins in signaling pathways and transcription processes in genetic regulatory circuits.

As of present, kinetic models are usually restricted to small scale systems. The median of the number of the reactions and species that 462 curated kinetic models in Biomodels Database [94] included are only 12 and 11, respectively. Yet the information they provide at both levels increases very rapidly. This is usually accomplished by *in vitro* experiments which give insight into appropriate formulations of enzyme kinetics. Also values of the parameters can be determined by *in vitro* experiments with isolated enzymes. Another common way towards this aim is the use of *in vivo* experiments in which metabolite concentrations are measured. Optimal values of the parameters can then be estimated by using concentration data [148]. However, *in vitro* and *in vivo* kinetics can be very different, not only in the values of the parameters but more importantly, also in the formulation [146]. This points to the need for careful investigation of the model's validity on the first information level that we defined above.

[1] This chapter is based on:
D. Hasdemir, Huub CJ Hoefsloot, Johan A Westerhuis, Age K Smilde. How informative is your kinetic model?: using resampling methods for model invalidation. *BMC Systems Biology*, 8(1):61, 2014.

Most of the time, models are assessed qualitatively based on the goodness of their fit to concentration data [54]. In some other cases, new datasets in different biological conditions are generated and a qualitative analysis is made based on the model's ability to predict new datasets [38]. However, most of the time multiple candidate models with different structures can show very similar goodness of fit and also prediction in another experimental condition. This stems from high levels of adaptability in these models. One could argue that all candidate models are good as long as they perform reasonably well in prediction. However, rapid elimination of less informative models would be very beneficial to the metabolic modeling community. It would ease the way to trustworthy libraries of models providing the researchers with speed and accuracy for larger scale models. To this aim, model selection and invalidation algorithms supply a quantitative framework.

Model selection criteria borrowed from statistical literature such as Akaike and Bayesian Information Criteria (AIC and BIC respectively) are among the most popular approaches introduced for the selection of systems biology models [2, 100, 154]. Model selection based on AIC have also been successfully implemented in software packages which aim to select the best model within a family of automatically generated models derived from one master model by adding/removing species or interactions [47, 61].

However, those criteria always support in favor of one model without providing any significance to their decisions [29] and can not produce clear results when many parameters are involved [61]. An alternative which is capable of ranking different models according to their plausibility was introduced within a Bayesian perspective using Bayes Factors [159]. This family of Bayesian methods unfortunately still remain unemployed in the field due to the need for smart assumptions on parameters' prior distributions and their costliness in calculation of bulky integrals despite promising effort regarding the second obstacle [111, 149]. In some studies robustness based measures were proposed for model selection [13, 56]. For oscillating systems, robustness of the model can support its preference over different models. However, this might not hold true for the whole family of kinetic models in systems biology.

Although not employed regularly, the systems biology community has been provided with tools to select between different model structures. However, invalidation of a model structure without an alternative to compare with has not been considered much in the related literature. An analytical approach suggests use of barrier certificates which are functions whose existence proves that the model behavior can never intersect the experimental data [6]. The existence of the barrier certificates proves the invalidity of the models. However the approach is purely analytical and very complex so it is not easily applicable by biologists. Another drawback is the

difficulty in the construction of the barrier certificates for complicated system descriptions as the authors also elaborate in their paper.

In this study, we introduce a statistical measure for the invalidation of kinetic models which suffers neither from complex model descriptions nor large scale models. We use the predictive power of Smooth Principal Components Analysis (SPCA), an unsupervised data analysis method as a threshold to assess the predictive power of kinetic metabolic models. By using this threshold value, we can determine which model structures are informative enough to deserve further attention and which model structures should be abandoned. Our approach stands on a basic assumption: If a totally unsupervised data analysis method without any prior biochemical knowledge predicts better than a kinetic model can do, that points to an inaccuracy or incompleteness in the information which the kinetic model provides us with.

With this study, we also want to bring the attention of the systems biology community to the idea of using resampling methods which have proven to be very useful in machine learning and data analysis. To our knowledge these methods' potential has not been exploited fully in the analysis of kinetic systems biology models.

Using synthetic data generated from metabolic models has been adopted widely in literature as a way of testing algorithms in a controlled context [109]. Here, we also employed this approach and used a toy metabolic model and a real signaling model for the generation of data. By using this data, we tested models also with lower complexity than the true model to assess the sensitivity and specificity of our approach.

We applied our method also on an eicosanoid production model in human white blood cells. Eicosanoid is a subclass of fatty acyls. Fatty acyls constitute one of the six major classes of lipids and are related to inflammation, rheumatoid arthritis, sepsis and asthma. Eicosanoids are divided into different groups one of which is prostaglandin family. Prostaglandins have been found to be related to many symptoms of inflammation like fever and pain [26, 54, 171]. That makes the eicosanoids important targets for modeling studies which can be used for predictive purposes in response to treatment with anti-inflammatory drugs. A kinetic model describing the production of prostaglandins from arachidonic acid has been published in [54]. The model includes the substrate arachidonic Acid, 8 downstream metabolites, signaling molecules and 4 different enzymes. All reactions were formulated by mass action kinetics. Due to the scarcity of information on enzyme activity regulation, rate parameters for enzymatic reactions were formulated as linear functions of enzyme-regulator molecules. Given the simplicity of the kinetics in the model and limited number of components, we wanted to assess its informative level and our results

showed that the model could not be invalidated with the available data.

The other benchmark pathway we analyzed was the well known high osmolarity glycerol (HOG) pathway in yeast. Osmo-adaptation in yeast has started to receive increasing attention with the discovery of the associated mitogen-activated protein kinase (MAPK) cascade [55, 63]. Since then, the HOG pathway proved to be a well studied model system to study the principles of signal transduction due to MAPK cascades being conserved eukaryotic signal transduction pathways. The pathway is in charge of regulating the glycerol accumulation in the cell in response to the changing osmotic pressure in the environment. It has been widely accepted that the upstream pathway consists of two redundant paths starting with two different transmembrane osmosensor proteins Sho1p and Sln1p. The cascade proceeds with the phosphorylation of Pbs2p, Pbs2p-Sho1p complex and Hog1p towards the transcriptional regulation of glycerol production [123, 128]. However, there is still active debate on post-translational interactions and transient feedback mechanisms involved in the signal transduction [103, 128]. Therefore we analyzed a recently published comprehensive model of the HOG pathway to check its predictive properties given part of the experimental data used to build the model [103, 128]. We also used the model as a basis for our simulation studies in which we generated data according to the published level of complexity and questioned a simplified version for its validity.

2.2. Methods

2.2.1. Toy metabolic model

We used an unbranched toy metabolic pathway for the generation of synthetic data (Figure 2.1). It included one substrate and four downstream metabolites whose production followed Michaelis-Menten kinetics. Equation 2.1 shows the set of ordinary differential equations constituting the true model (ODE_T) which we used for the generation of the data. We used the dynamic part of the time series data in the first 22 time points as the data without experimental noise. We stored the data in a matrix with metabolites in the columns and time points in the rows.

We introduced homogeneous experimental noise to the data in the form of Gaussian noise with zero mean and varying standard deviation. We varied the standard

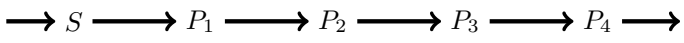

Figure 2.1: Layout of the toy model. Unbranched toy model consisted of one substrate and 4 downstream products.

deviation of noise between 0.001 and 0.05. At each degree of experimental noise, we repeated the simulations with 100 different realizations of the data.

$$\begin{aligned}
\frac{dS}{dt} &= k_{in} - \frac{v_{max_1}[S]}{Km_1 + [S]} \\
\frac{dP1}{dt} &= \frac{v_{max_1}[S]}{Km_1 + [S]} - \frac{v_{max_2}[P1]}{Km_2 + [P1]} \\
\frac{dP2}{dt} &= \frac{v_{max_2}[P1]}{Km_2 + [P1]} - \frac{v_{max_3}[P2]}{Km_3 + [P2]} \\
\frac{dP3}{dt} &= \frac{v_{max_3}[P2]}{Km_3 + [P2]} - \frac{v_{max_4}[P3]}{Km_4 + [P3]} \\
\frac{dP4}{dt} &= \frac{v_{max_4}[P3]}{Km_4 + [P3]} - k_{out}[P4]
\end{aligned} \tag{2.1}$$

2.2.2. Comparison of predictive power by cross validation

One of the key features of our approach is using cross validation, a resampling technique as we mentioned in our introduction. Cross validation is a very commonly used validation method in statistics and machine learning [4, 23] for determining the optimal level of complexity in models. In cross validation, a data set is divided into two parts: training and test sets. Only the training set is used for the parameter inference whereas the test set is only used for assessing the performance of the model on parts of the data that have not been associated with parameter inference. The procedure is repeated with alternating training and test sets several times and the performance results are averaged over all repetitions. In classification problems, that performance measure is the accuracy in classification of the test objects. In regression or dimension reduction problems, it is the prediction error. Throughout this chapter we refer to the residuals in the prediction of only the test set data points by using the term 'prediction error'. In this study, we inferred the parameters of both the kinetic and the SPCA model using the training data and we used prediction error as a measure of the predictive power of both modeling approaches.

We used a diagonal cross validation scheme in which 10% of the data was used as the test set. This kind of stratified cross validation scheme provided us with diverse test sets which were homogeneous both in metabolites and time points (Figure 2.2). With this scheme, every element -excluding the first and the last time points- in the data matrix belonged to a test set once and the sum of the prediction error over all test sets gave the total prediction error. The first time points were not included in the test sets because initial concentrations of the metabolites were also unknown parameters of the kinetic model as we will touch upon also in the proceeding sections.

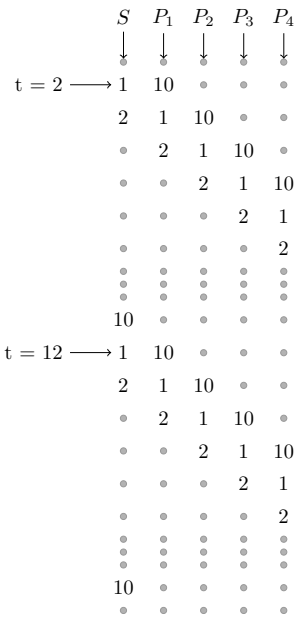

Figure 2.2: Stratified diagonal cross validation scheme. 1 denotes the elements in the first test set whereas 2 denotes those in the second test set and so on. Elements of 10 different test sets were diagonally selected as shown in the figure.

That is why these points could not be used as test points in cross validation. The reason for excluding the last time points was related to the fact that it is more challenging to predict the end points with SPCA compared to the interior time points. Due to this fact, we approached the prediction of last time points in the forecast analysis context where we could adjust the smoothing penalty parameter of SPCA accordingly.

2.2.3. Comparison of predictive power by forecast analysis

Forecasting refers to predicting the future outcome of a variable of interest. It is commonly used in a lot of disciplines ranging from economics to meteorology where modeling is crucial. In forecast analysis, models are established using past data and extrapolated to the future. Variations on forecast analysis exist depending on the types of the models, the needs of the field, partitioning of the training and test sets and the types of the measures that are used to assess the amount of prediction error [20].

Here, we used a basic scheme which fits for both SPCA and kinetic modeling. In each run, we left out approximately the last 20% of the time points of one metabolite

as the test set. By this way, we could assign a certain percentage of the end time profiles to a test set once and the total prediction error on those time points gave us a measure for the predictive power of the models.

2.2.4. Kinetic modeling

We estimated the rate parameters ($\vec{k}$) and the initial metabolite concentrations at t=0 ($\vec{X}_0$) from the training data. We carried out the optimization with a nonlinear solver in Matlab, namely lsqnonlin function which implements the trust-region-reflective algorithm [32]. The objective function was to minimize the square of the difference between the noisy synthetic data and the model values of the training set elements. In Equation 2.2, the weighting matrix $\mathbf{W_{tr}}$ is a binary matrix with 0's corresponding to the test set elements in the data matrix and 1's corresponding to the training set elements. We excluded test set elements from the parameter inference process by element-wise multiplication by $\mathbf{W_{tr}}$. This multiplication is denoted by the Hadamard Product, $\circ$, whereas the model concentration values ($\hat{\mathbf{X}}$) are given as a function of the unknown parameter vectors $\vec{k}$ and $\vec{X}_0$.

$$\underset{\vec{k},\vec{X}_0}{min}\|\mathbf{W_{tr}} \circ (\mathbf{X} - \hat{\mathbf{X}}(\vec{k}, \vec{X}_0))\|^2 \tag{2.2}$$

We estimated the model concentration values by numerically integrating the set of differential equations defining the model in question at every iteration step throughout the optimization. We repeated the procedure with two different models: the true model (Equation 2.1) and the simplified model (Equation 2.4). The true model (ODE_T) is the model we had used for data generation. In the simplified model (ODE_S), the production of the first metabolite was formulated with linear kinetics with only one rate parameter.

$$\|\mathbf{W_{test}} \circ (\mathbf{X} - \hat{\mathbf{X}}(\vec{k}, \vec{X}_0))\|^2 \tag{2.3}$$

The prediction error for one test set was then calculated as in Equation 2.3 where $\mathbf{W_{test}}$ has 0's for training set elements and 1's for test set elements.

$$\begin{aligned}
\frac{dS}{dt} &= k_{in} - k[S] \\
\frac{dP1}{dt} &= k[S] - \frac{v_{max_2}[P1]}{Km_2 + [P1]} \\
\frac{dP2}{dt} &= \frac{v_{max_2}[P1]}{Km_2 + [P1]} - \frac{v_{max_3}[P2]}{Km_3 + [P2]} \\
\frac{dP3}{dt} &= \frac{v_{max_3}[P2]}{Km_3 + [P2]} - \frac{v_{max_4}[P3]}{Km_4 + [P3]} \\
\frac{dP4}{dt} &= \frac{v_{max_4}[P3]}{Km_4 + [P3]} - k_{out}[P4]
\end{aligned} \tag{2.4}$$

2.2.5. Smooth principal components analysis

The other key feature of our approach is its comparative nature. The reference method we used for comparison was Smooth Principal Components Analysis (SPCA) [157]. SPCA is an extension of the well known dimension reduction method Principal Components Analysis (PCA) [4, 75] with roughness penalties on the scores.

The reference method is completely unsupervised, making no use of the kinetic model structure nor of any prior biochemical knowledge. Smooth Principal Components Analysis penalizes the non-smoothness of the scores and thus can make use of the underlying time profile in predicting the missing points in the data [157]. This makes it more efficient than normal PCA to be used as a prediction method when the scores are expected to have smoothness as in the case of time series data.

We have estimated the smooth scores ($\mathbf{Z}$) and loadings ($\mathbf{P}$) within a Weighted Principal Components Analysis (WPCA) formulation. WPCA is a special variety of PCA in which data points are weighted proportional to the measurement accuracy at those points by using a weighting matrix [72]. WPCA can also be used to handle PCA on data with missing points using a binary weighting matrix where the entries corresponding to missing points are 0 [80]. That allows it to be employed as a favorite analysis method in multivariate statistics when there are missing points in the data [77] and also for performing cross validation where some of the data points are excluded as test set elements [23]. Our application in this study follows the latter.

We have minimized the objective function in Equation 2.5 by using the same nonlinear solver as we have used for kinetic modeling. The objective function in Equation 2.5 is comprised of two terms. The first term is the sum of squares of the difference between the measured ($\mathbf{X}$) and model values of the training set elements by the SPCA model ($\mathbf{ZP^T}$). Here, $\mathbf{W_{tr}}$ is the same binary matrix as we used in

the kinetic modeling section. The second term is the penalty term scaled with the smoothing parameter λ where $\mathbf{D_2}$ represents a second order difference matrix. With a second order penalty, scores are penalized for the change in slope [157] which is appropriate in our case since we deal with time series data.

$$\underset{\mathbf{Z},\mathbf{P}}{min}\|\mathbf{W_{tr}} \circ (\mathbf{X} - \mathbf{ZP^T})\|^2 + \lambda\|\mathbf{D_2Z}\|^2 \tag{2.5}$$

Prior to using SPCA, the number of principal components (PCs) and the value of the smoothing parameter (λ) have to be calibrated for each specific problem. We used cross validation also for this purpose. After the test set elements (outer test sets) which we used also in the kinetic modeling section were removed from the dataset, the remaining part was again subjected to a division of test (inner test sets) and training sets for a 10-fold cross validation with 10 repetitions. We applied SPCA using a particular value for λ and a particular number of PCs on every training set. The average prediction error on all different inner test sets from 10 different repetitions gave us a measure of how well the inner test set points could be predicted using that particular parameter combination. We repeated the same procedure by using increasing λ values and increasing number of PCs until the predictions on the inner test sets could not improve with increasing number of PCs and started to deteriorate with increasing λ after certain limits. These limits gave us the optimal values for the parameters. This approach is known as "Double Cross Validation" since it makes use of cross validation at two different levels and it leads to unbiased prediction errors [141].

Once the optimal λ the optimal number of PCs were determined, they were used for the estimation of the scores ($\mathbf{Z}$) and the loadings ($\mathbf{P}$). Equation 2.6 shows how we calculated the prediction error for a single test set whether an inner or an outer test set. In Equation 2.6, $\mathbf{W_{test}}$ has 1's for test set elements and 0's for training set elements as we used in the kinetic modeling section. In Figure 2.3, we give a detailed flowchart of our approach.

$$\|\mathbf{W_{test}} \circ (\mathbf{X} - \mathbf{ZP^T})\|^2 \tag{2.6}$$

In forecast analysis we followed the same approach with a small variation. There, we left out windows of data which consisted of 5 consecutive time points from the same metabolite as inner test sets, in each run. This helped us to infer the optimal parameters better for the accurate prediction of the end time points. This was because, also in forecast analysis, the purpose was to predict consecutive time points, in opposition to cross validation where the outer test set points were not consecutive.

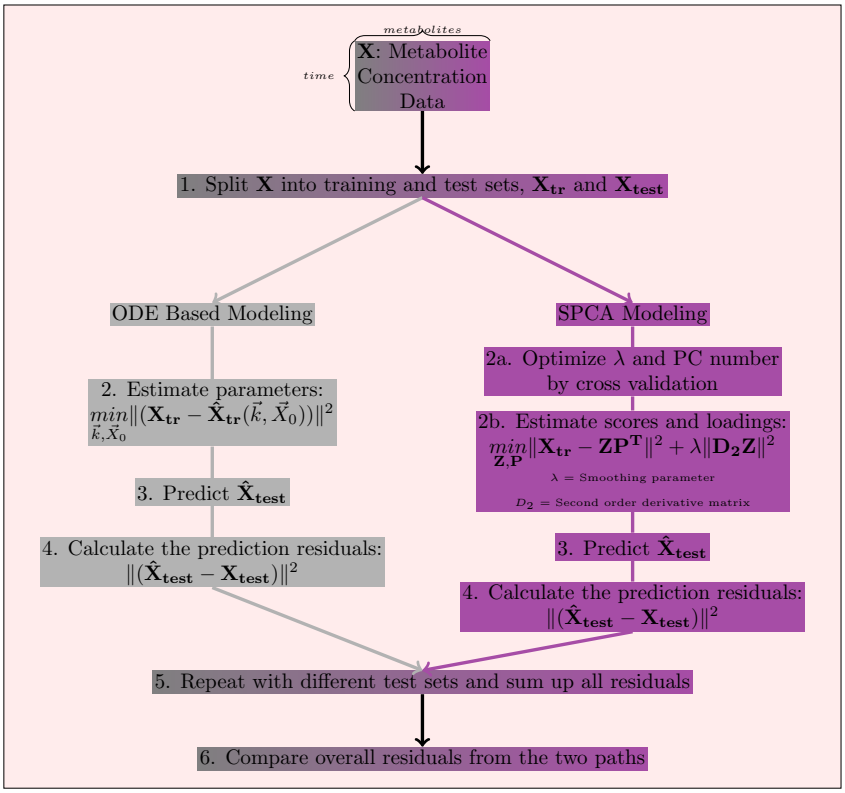

Figure 2.3: Flowchart for the approach. The figure summarizes graphically our comparative model invalidation approach.

2.3. Results and discussion

2.3.1. Toy model

We carried out simulations at different experimental noise levels. At the lowest noise level we tested, the experimental noise was drawn from a normal distribution with a standard deviation (σ_{noise}) of 0.001. At this level of standard deviation, the mean relative noise in all of the 100 different realizations of the data was below 1%. At the maximum noise level ($\sigma_{noise} = 0.05$), the mean relative noise at a single realization of the data could increase up to 13%. Mean Relative Noise (MRN) is a measure of the noise in the data calculated as the mean of individual relative noise levels for each element in the data matrix (Equation 2.7). In Equation 2.7, x_{ij}^m denotes the values generated by the model according to Equation 2.1 whereas x_{ij} denotes the synthetic data with experimental noise added.

$$MRN = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|x_{ij}^{m}-x_{ij}|}{x_{ij}^{m}}}{n \times m}$$
$$\text{n} = \#\text{ time points}$$
$$\text{m} = \#\text{ metabolites} \tag{2.7}$$

Table 2.1 and 2.2 show all the invalidation decisions made in the simulations study. Results show that our SPCA-based comparative approach performs very well in invalidating simplified models, indicating the method's high sensitivity. The low number of invalidation decisions made for the true model relate to the high specificity of our approach.

Table 2.1: All the invalidation decisions made by using cross validation. The numbers show in how many of the 100 different noise realizations an invalidation decision was made for the models being questioned by using cross validation. ODE_S and ODE_T denote the simplified model and the true model, respectively. Mean Relative Noise at each noise level is given as the mean of the MRN values in 100 different realizations of the data, calculated based on Equation 2.7.

σ_{noise}	MRN (%)	ODE_S	ODE_T
0.001	< 1	100	0
0.01	2.2	100	0
0.025	5.4	100	4
0.03	6.5	100	8
0.05	10.8	75	14

At low noise levels (up to $\sigma_{noise} = 0.01$), the difference between the prediction error levels of the true (ODE_T) and the simplified (ODE_S) kinetic models was very high, around two orders of magnitude (Figure 2.4). At these simulations, SPCA always performed better than ODE_S and worse than ODE_T, in the cross validations. At that level, forecast analysis resulted in very similar performance with very high sensitivity and specificity.

At medium noise level ($\sigma_{noise} = 0.025$), the difference between prediction error levels of ODE_T and ODE_S became smaller due to noise interference. At that point, the reconstructed metabolite profile by ODE_S (green line in Figure 2.5) pointed to a reasonable model for the data (blue stars) from a qualitative point of view. However, our quantitative analysis showed that ODE_S predictions were worse than SPCA in the cross validations. This showed that SPCA predictions could be used to invalidate ODE_S with very high sensitivity. Decision for not invalidating ODE_T in most of the cases showed the specificity of the method. The number of noise realizations at which SPCA cross validation invalidated ODE_T or ODE_S can be seen in Table 2.1 for each noise level.

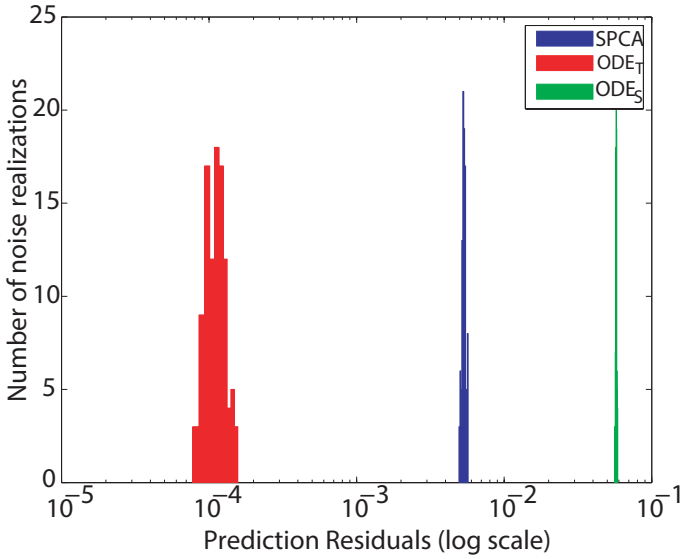

Figure 2.4: Prediction errors in the simulations with very low level of noise. The figure shows the residuals obtained in the cross validation simulations at the lowest noise level. At very low noise levels, there is a clear difference between prediction errors of the true and the simplified models both in cross validation and forecast analysis. The figure has a logarithmic x-axis.

Table 2.2: All the invalidation decisions made by using forecast analysis. The numbers show in how many of the 100 different noise realizations an invalidation decision was made for the models being questioned by using forecast analysis. ODE_S and ODE_T denote the simplified model and the true model, respectively. Mean Relative Noise at each noise level is given as the mean of the MRN values in 100 different realizations of the data, calculated based on Equation 2.7.

σ_{noise}	MRN (%)	ODE_S	ODE_T
0.001	< 1	100	0
0.01	2.2	100	3
0.025	5.4	86	17
0.03	6.5	81	17

Up to this noise level, we determined the optimal value of the λ parameter as 0.005 by cross validation for all different realizations of the data. Cross validation gave also the optimal number of principal components as 4 in all of the cases covering more than 99% of the variance in the data. We estimated the optimal values of the parameters to be the same in different noise realizations due to the low amount of noise in the data. However, starting with this noise level, we had to determine the values of the SPCA parameters differently for each noise realization. This clearly showed that the datasets in 100 different noise realizations had different characteristics due to the increasing difference in the realization of the added noise. The

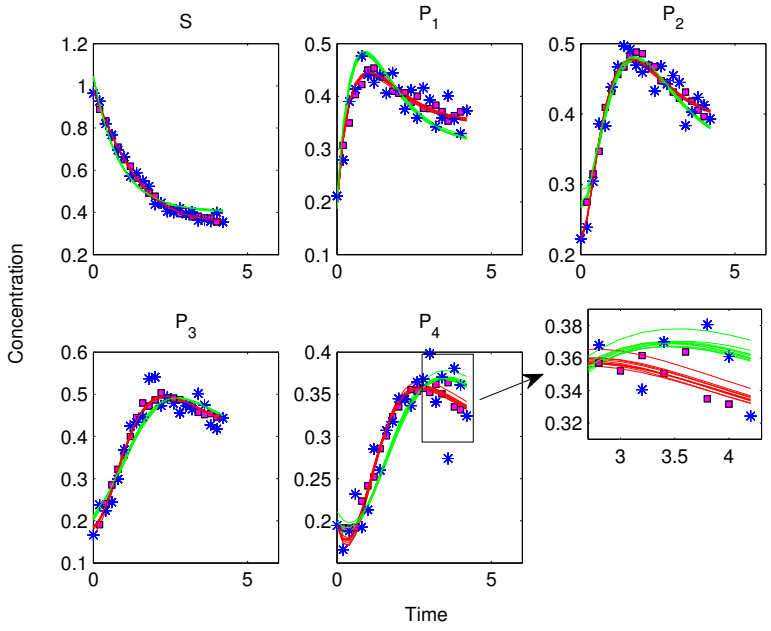

Figure 2.5: Predictions by different models at medium noise level - Cross validation. The blue stars denote the data points whereas the magenta squares show the SPCA predictions when the corresponding data points were excluded as test set elements. The red and the green lines show the reconstructed time profiles of the metabolites by using the true and the simplified models for different test sets, respectively. The magnifying window in the lower right hand side of the figure shows the differences of the reconstructed time profiles for different test sets. There, the deviation between the lines of the same color (obtained by using different test sets but the same model description each time) can be seen in great detail.

difference in the parameters were more apparent for the forecast analysis than for the cross validation.

At this noise level, invalidation by forecasting started to drag behind the cross validations. Apparently, noise interfered more when consecutive time points in the end of the time profiles were removed from the training data. This held true for both the SPCA and the kinetic modeling. Due to worsening predictions of SPCA, ODE_S could not be invalidated in 14% of the noise realizations (see Table 2.2). However, the predictions by the ODE_T got also worse, resulting in an incorrect invalidation decision in 17% of the realizations. Predictions of an example simulation at this noise level can be seen in Figure 2.6.

At high noise levels ($\sigma_{noise} = 0.05$), ODE_T started to lose its predictive power compared to SPCA in 14% of the realizations (see Table 2.1). This could have

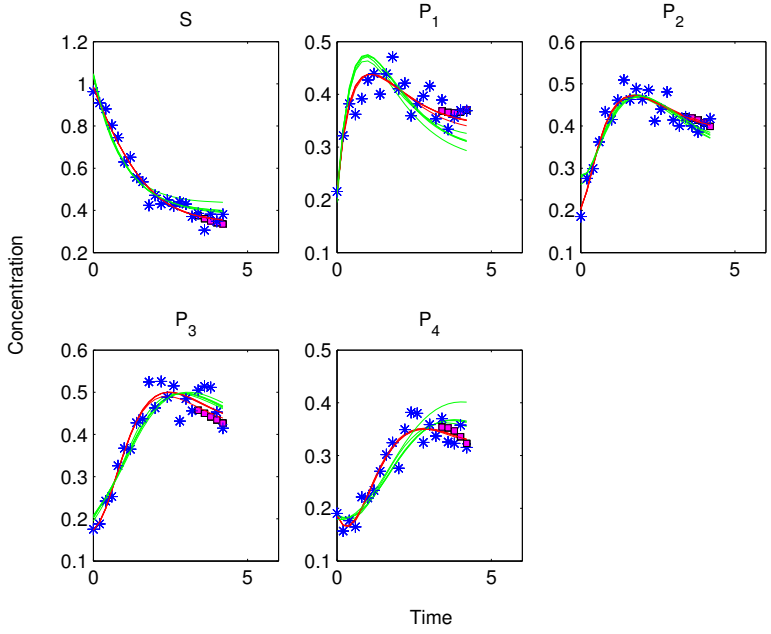

Figure 2.6: Predictions by different models at medium noise level - Forecast analysis. The color coding for this figure follows the one in Figure 2.5. In the forecast analysis approach, a window of the data which consisted of a significant number of consecutive time points were left out as test sets for each metabolite. The figure shows the SPCA predictions with magenta squares which are better than the ODE_S and worse than the ODE_T.

stemmed from inefficient estimation of the model parameters because of possible local minima in the optimization. In order to check that, optimization was repeated in those problematic cases with multiple starting points. This revealed that the problem was not due to sub-optimal parameters but due to the fact that data was too deteriorated to be explained well even by ODE_T (Figure 2.7). However, still in 75% of the realizations, SPCA predictions invalidated ODE_S successfully. At this noise level, inference of the optimal SPCA parameters in the cross validations started to be affected by the noise as well. The value of the smoothing parameter λ and the number of PCs determined by cross validation using other test set patterns were not always optimal. That is why we adopted a grid search approach for this noise level in which we varied the parameter λ in a small range around the value determined by cross validation. As long as we could find better predictions by SPCA than the model in question, we could conclude that we could invalidate that model. Here, we have to emphasize that during the grid search in the small neighborhood of the estimated λ, SPCA predictions changed very little. This showed that prediction

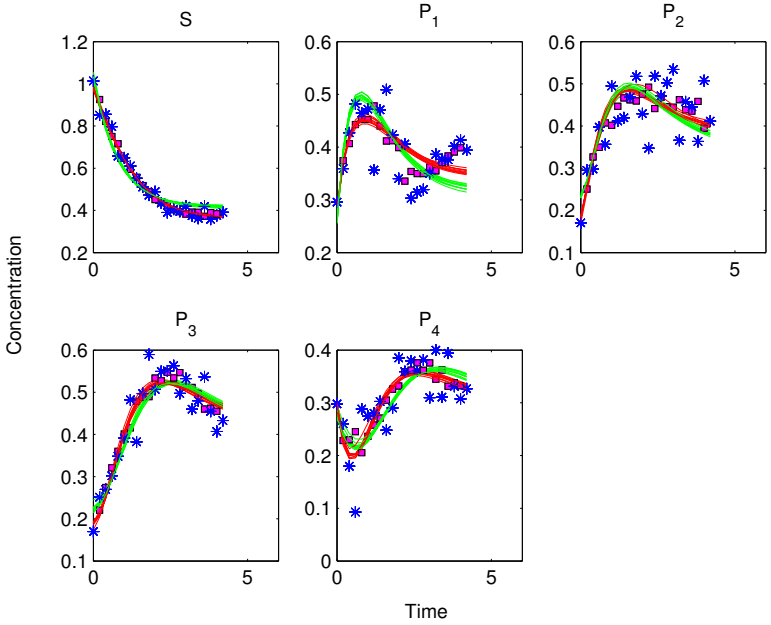

Figure 2.7: Predictions by different models at very high noise level. The color coding for this figure follows the one in Figure 2.5. At this noise level, the data seems very deteriorated by noise especially for metabolites with lower concentration ranges since the added noise is homogeneous.

error from SPCA was very stable. As we use it as a threshold for invalidation of models, proving to be robust against small changes in the parameters is very important.

The overall results of our simulations study with the toy model suggest that SPCA predictions within a traditional stratified cross validation framework perform very well as a threshold measure which can be used to invalidate too simple models. It meets the essential criteria of being totally unsupervised and providing a good description of the data. Even at very high levels of noise (Figure 2.7), it can serve as an invalidating measure. SPCA predictions within a forecasting framework also serve well for the invalidation purpose. However, it performs worse in high noise levels. On the other hand, we think that for many kinetic modelers, forecasting seems more intuitive and biologically meaningful. Therefore, it is of high importance to include it in our study.

Noise level affects the plausibility of model simplifying approximations: As a small demonstration of a specific research question for which our approach can be used, we investigated the plausibility of model simplifying approximations

in kinetic modeling.

We used a moderate value (0.33) for the first Michaelis constant (Km_1) while generating the data. Its value was well within the range of the substrate concentration ($[S] \in [0.2, 1]$). If it was much higher than the substrate concentration, the substrate concentration term in the denominator of the first rate equation (see Equation 2.1) could have been neglected. Therefore, the model simplification from ODE_T to ODE_S could have been performed with very low information loss. This approximation is widely employed in many model fitting studies to justify the simplification of Michaelis-Menten Kinetics to linear kinetics which helps to decrease the number of parameters in the model. However, the ranges of the parameter values in which this approximation will be plausible are never clear.

Table 2.3: The number of cases where the model simplification was acceptable. The numbers show in how many of the 100 different noise realizations, invalidation decision could not be made for the simplified model, ODE_S. Lack of invalidation decision showed the validity of the model simplification. The table shows that at different Michealis constant (Km_1) values there is different level of support for the validity of the linear kinetics assumption.

Km_1	$\sigma_{noise} = 0.01$	$\sigma_{noise} = 0.02$	$\sigma_{noise} = 0.04$
0.33	0	0	10
1.4	44	70	82
3	100	97	94

By using our SPCA-based invalidation approach, we could investigate how the invalidation decisions changed for the simplified model with respect to the value of the Michaelis constant. This helped us to assess the plausibility of the approximation based on the degree of support by the available data. We could also observe how that assessment became difficult by increasing noise in the data. For this purpose, we used three different Km_1 values in data generation. We performed the simulations with noise levels between $\sigma_{noise} = 0.01$ and $\sigma_{noise} = 0.04$.

We could see the expected relationship between the value of the Michaelis constant and the plausibility of the model simplifying approximation by using our approach. When the Michaelis constant was 0.33, well within the range of the substrate concentration, the simplifying approximation was never supported by the data until high amount of noise in the data (See Table 2.3). However, when its value was increased nearly 9-fold, well above the substrate concentration range, in all of the realizations, the data supported the simplifying approximation.

The change in the accuracy of the plausibility assessment proved to be an even more important observation. Table 2.3 shows that under low levels of noise, when the Michaelis constant was only slightly above the range of the substrate concentration at 1.4, in some 40 of the realizations, ODE_S was not invalidated. This means that

the simplification was supported in nearly half of the realizations. The number of realizations at which ODE_S could not be invalidated could increase to 82 when the measurements were more erroneous at $\sigma_{noise} = 0.04$ (Mean Relative Noise $\approx$ 8%). This clearly shows that noise is an important factor that interferes with the plausibility of model simplification. At low noise levels, it is easier to pull out the correct kinetic mechanism from the rest of the simpler candidates. When higher noise is existent in data, detection of poorer predictions by simpler mechanisms become more difficult by the noise. Models that are in fact too simple to explain the mechanistic behavior can be wrongly regarded as plausible candidates when the measurement accuracy is low in the experiments.

2.3.2. Eicosanoid production model

Data belonging to the biological system under study were time series concentration data (0,0.5,1,2,4,8,12,24 hours) of 8 metabolites (arachidonic Acid, 11-HETE, PGE2, PGF2a, PGD2, PGJ2, dPGD2, dPGJ2) from 3 different experiments with 3 technical replicates (9 replicates in total) in response to treatment of human macrophage cells with KDO_2-lipidA (an LPS analog) [54]. The model describing the system included 22 first order reaction rate parameters. The topology of the pathway is as shown in Figure 2.8.

We used the mean of all replicates in the calculations. However, replicates in data allowed us to estimate the noise level and we calculated the mean relative noise (MRN) in the data as 8%. That level of noise in the data corresponded to the medium to high noise level that we have covered in our simulations study. Based on the results we achieved in our simulations study, we could expect high sensitivity and specificity of our approach in that noise range.

A weighted objective function was needed in kinetic modeling to overcome the risk of it being dominated by the metabolites with higher concentrations. The weight matrix $\mathbf{W}$ we used included the reciprocal of the maximum concentration of the corresponding metabolite in all the time points. Equation 2.8 shows the entries of this weight matrix $\mathbf{W}$ for the training set elements. For the test elements, entries were 0 as in the case of the calculations for simulated data.

$$\mathbf{W}_{ij} = \frac{1}{max(\mathbf{X}_j)} \tag{2.8}$$

In SPCA, we preprocessed the data in accordance with the kinetic modeling approach. Therefore, we first scaled every concentration value in the data matrix by the maximum concentration of the corresponding metabolite in all the time points and carried out SPCA on that scaled data matrix. It is highly recommended to

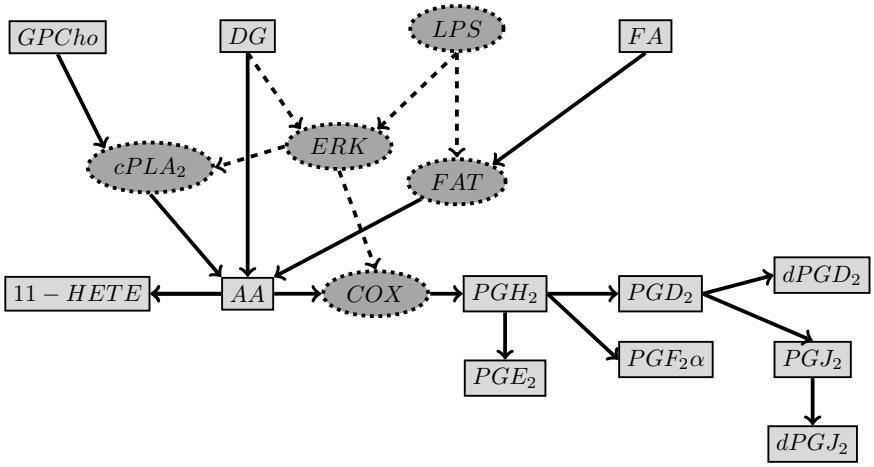

Figure 2.8: Topology of the signaling and metabolic pathway of eicosanoid production in human. The known pathway topology was simplified by Gupta *et al.* [54] based on the availability of metabolite concentration data in their experiments. The rectangles show the metabolites and the solid arrows indicate the metabolic transformations involved. The ellipses with dashed borders show the enzymes catalyzing the metabolic transformations between two metabolites that are neighboring it in the graph. The dashed lines denote the effect of enzymes and molecules on the activity of enzymes.

scale the data prior to any type of PCA application if the order of magnitude of the data values change substantially between columns, since that will allow a more fair distribution of the loadings of the variables in the most important principal components. Then, the smoothing parameter applies more equally for every metabolite and we can achieve better smoothing of all the time profiles.

We used an 8-fold diagonal cross validation scheme with 5 repetitions. In the first repetition, the test sets involved consecutive time points from consecutive metabolites as was shown in Figure 2.2. In the other four repetitions, the test sets involved time points with increasing intervals from different metabolites. By this approach we could achieve very diverse test sets and all data points except the first and last time points of each metabolite were included in a test set five times. We also weighted the resulting residuals by the maximum concentration before summing up to the final value and averaged by the number of repetitions.

The optimal λ and the number of principal components needed were estimated by using a 12 fold stratified cross validation scheme with 10 repetitions. We have found the optimal number of PCs to be 3 and the λ value between 5 and 25. Following a grid search between those lambda values, we achieved the final prediction residuals in SPCA as 6% of the sum of squares of the weighted data matrix, higher than

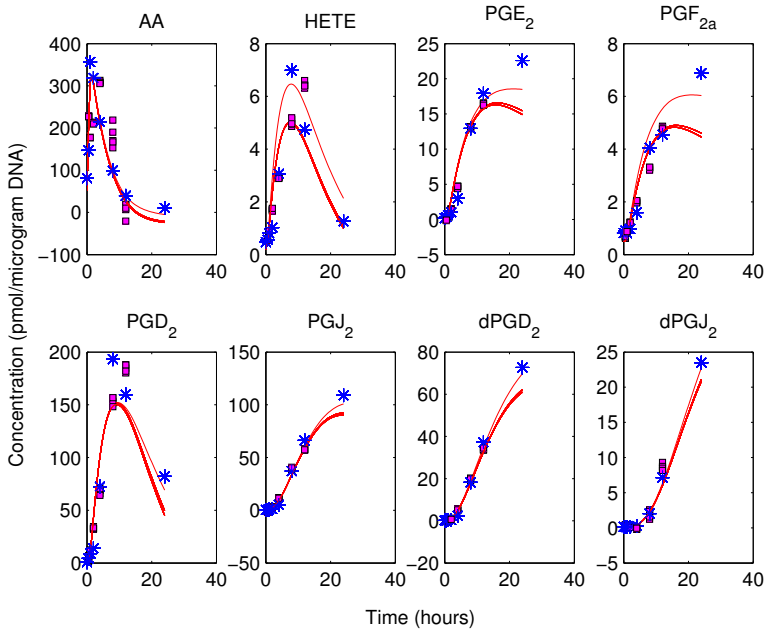

Figure 2.9: Predictions on eicosanoid production pathway. Solid red lines show the metabolite profiles constructed for the 8 metabolites by the kinetic model in question when different test sets were used. The blue stars show the mean of all 9 replicates of data at each time point whereas the magenta squares denote the predictions by SPCA for each data point when they were excluded from calculations as test set elements. There exist 5 SPCA predictions for each interior time point because they were included 5 times in different test sets.

the prediction residuals in kinetic modeling which was only 3%. These predictions can be seen in Figure 2.9. This showed that the model proposed for the eicosanoid production pathway could not be invalidated by using the available data. Despite its simplicity in enzymatic reaction kinetics, it proved to be competent in explaining the data.

2.3.3. HOG signaling model in yeast

High osmolarity glycerol signaling pathway in yeast is a well studied system since it is regarded as a model system for studying the principles of signal transduction in eukaryotic cells. The structure of the phosphorylation cascade starting from two redundant osmosensors (Sho1p and Sln1p) and leading to the transcriptional regulation of glycerol production for osmotic balance is generally agreed upon. However there are still competing hypotheses on especially the transient feedback relations involved in the cascade. These include but are not limited to the post-translational

regulation of glycerol production, Fps1p phosphorylation and Sho1p phosphorylation by the Hog1p. Schaber and coworkers carried out a comprehensive study where they compared 192 different models [128]. Here, we used their best approximating model with the accession number of MODEL1209110001 in Biomodels Database [94]. The model consisted of 15 species and 20 free parameters. 10 different variations of mass-action kinetics with either inhibitors or activators were used for the reaction kinetics in the model. Volume was also included in the model as a variable whose value changes in time. The interactions in the model can be seen in Figure 2.10.

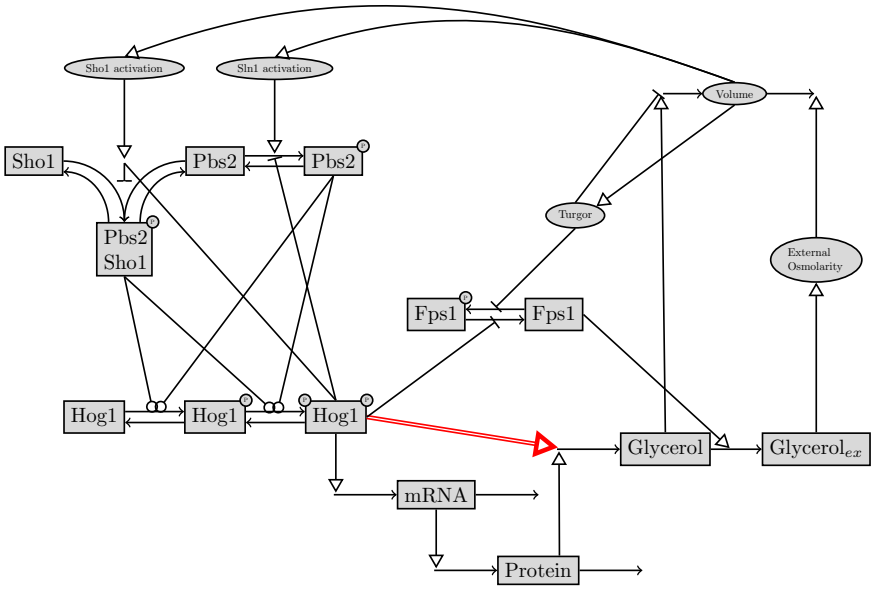

Figure 2.10: Topology of the HOG signaling pathway in yeast. This pathway topology was proposed as the best approximating model topology in [128]. We used this model as our true model ODE_T for data generation in our simulations on HOG signaling pathway. The black lines with small arrow tips depict the transition between different species in the model like production, degradation or complex formation. The black lines with circle tips depict the phosphorylation processes by kinases. The lines with open triangle tips show activating regulatory interactions where as lines with blunt ends show deactivating regulatory interactions. The red colored double arrow denotes the post translational regulation of glycerol production by the active phosphorylated Hog1 protein. We removed this regulatory interaction in our simplified model ODE_S.

Synthetic data

We used the model depicted in Figure 2.10 to generate data by using the optimal parameter values determined in [128]. Synthetic data consisted of the time profiles of 4 different species measured following two different osmotic shocks at 0.4 and 0.5

M. NaCl in wild type cells. The species were the phosphorylated Hog1p, glycerol, Hog1 dependent protein (mainly Gpd1p) and the associated mRNA. We set the number of measurement points to 43 which spans the dynamic part of the profiles between the shock and the steady state at around one hour later. Following the generation of model values, we added heterogeneous noise on the data. Noise was drawn from a standard normal distribution with two different values of standard deviation and multiplied by the concentration value of the species at that time point. The standard deviation was 0.01 and 0.2 in the low and high noise levels, respectively. We carried out kinetic modeling with the true model, ODE_T that we used to generate the data and a simplified model ODE_S which lacked the post-translational regulation of glycerol production by the phosphorylated Hog1p (see Figure 2.10). During both kinetic modeling and SPCA we used a weighting matrix which normalizes the difference between the data and the model predictions, by the mean of the concentration values of the species during all the time points. Weighting serves the purposes we explained in the previous section.

Table 2.4: All the invalidation decisions made for HOG pathway models. The numbers show in how many of the 100 different noise realizations an invalidation decision was made for the models being questioned by using forecast analysis. ODE_S and ODE_T denote the simplified model and the true model, respectively.

σ_{noise}	ODE_S	ODE_T
0.001	100	0
0.02	100	16

In this section, we employed the forecast analysis approach. In each run, we left out approximately 30% of the last time points of each species as the test set. For the determination of the optimal SPCA parameters, we followed a grid search approach and found that 2 principal components are enough with a mild smoothing penalty with λ=1. In Table 2.4 we report the number of invalidation decisions made for the two models and Figure 2.11 show the kinetic model and SPCA predictions on this dataset. Our results in this section confirmed once more that SPCA can predict well even when approximately one third of the data for a single species is left out. This can be seen especially in the upper 4 plots in Figure 2.11 where predictions not only on the steady part but also on the dynamic part of the profile are good. Even at this very high noise level (see Figure 2.11), SPCA predictions in forecast analysis can serve as an invalidating measure since ODE_S could be invalidated in all the noise realizations.

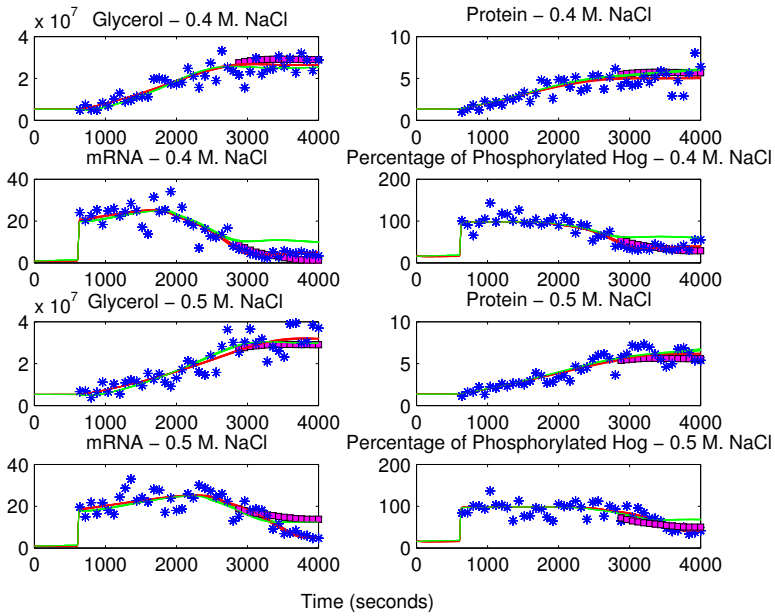

Figure 2.11: Predictions using synthetic data on HOG signaling pathway. In this figure, blue stars denote the synthetic data whereas magenta squares denote the SPCA predictions when the associated data points were left out as test set points. The red and the green solid lines show the time profiles predicted by the ODE_T and the ODE_S, respectively. The upper 4 subplots belong to the 0.4 M. NaCl shock experiment and the lower 4 subplots belong to the 0.5 M. NaCl shock experiment. The glycerol, Hog1 dependent protein (mainly Gpd1p) and mRNA amounts (in μmoles) are in absolute scale whereas we used the normalized phosphorylated Hog1p values. The Hog1p values were normalized to their maximum value measured in the corresponding experiment.

Real data

We used a part of the experimental data from [128] and[103] to question the best HOG signaling model reported in [128]. The real data included 4 different species. The first species was the phosphorylated Hog1p whose concentration values were normalized by its maximum concentration value in wild type cells at the same osmotic shock experiment. It was measured for the Sho1 and Sln1 deletion mutants at 6 different levels of osmotic shock. The other species were glycerol, protein and the associated mRNA measured in wild type cell following 0.5 M. NaCl treatment. Those species' concentrations were also normalized by their corresponding maximum concentration throughout their time profiles. We used only the dynamic part of the time profiles which start after the osmotic shock. Some of the interior time points were missing in the original data so we interpolated between the existing data points

to achieve a full data matrix of 13 time points and 15 columns. We needed a full data matrix because calculating the prediction residuals for the comparison of the two approaches is a very essential step in our analysis and for this purpose, we need to know the real values of the concentration values at the data points that we leave out as test sets. Therefore we imputed the missing values prior to SPCA & ODE modeling by interpolation. In total, more than half of the time points were calculated by interpolation for the Hog1 dependent proteins (mainly Gpd1p) and the glycerol. We questioned two different models as in the case of the synthetic data. The simplified model lacked the post-translational modification of glycerol production by the Hog1p.

We used forecast analysis in which we left out the last 3 time points from each column of the data matrix in each run. SPCA on this data matrix with 9 PC's and $\lambda = 8.10^6$ results in a very good representation of the dataset. Forecasting prediction error obtained from SPCA equals 0.6% of the sum of squares of the whole data matrix. This value is below the residuals obtained by the kinetic modeling using the full model and the simplified model, being 0.9% and 1.5% of the sum of squares of the whole data. Those predictions can be seen in Figure 2.12.

The results showed us that the model in question did not prove to be sufficient to explain the real data from [128] and[103] that we have used in our study. However, here we used only some part of the data that was available. Furthermore we had to impute many missing values prior to our calculations as mentioned above in this section. The reason for this is that we preferred to use the minimum amount of data that would suffice for the parameterization of the ODE model. Therefore the results we highlight here should be regarded as a more realistic demonstration of our approach rather than means of arriving at strict biological conclusions.

2.4. Conclusions

We introduced the use of two resampling methods, namely cross validation and forecast analysis for the analysis of kinetic systems biology models. Cross validation and forecast analysis allowed us to use a part of the available time series metabolite concentration data to infer the proposed model's kinetic parameters and the remaining part of the same dataset to assess the predictive power of the model. This way, we have showed that resampling strategies eliminated the need for additional datasets for the assessment of predictive capabilities of models. We used those two approaches within a Smooth Principal Components Analysis (SPCA)-based comparative approach for the invalidation of models.

Our approach depends on the assumption that correct kinetic model descriptions

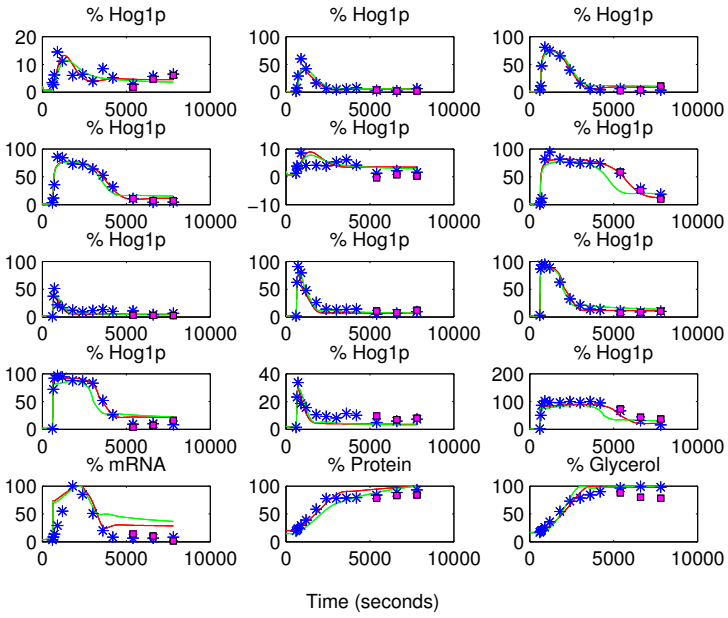

Figure 2.12: Predictions using real data on HOG signaling pathway. In this figure, blue stars denote the synthetic data whereas magenta squares denote the SPCA predictions when the associated data points were left out as test set points. The red and the green solid lines show the time profiles predicted by the full and the simplified models respectively. The upper 6 subplots belong to the phosphorylated Hog1p in Sln1 deletion mutant following 0.1, 0.2, 0.4, 0.6, 0.07 and 0.8 M. NaCl shock, respectively. The next 6 subplots belong to the same species in Sho1 deletion mutant after the same osmotic shocks. The last 3 subplots show the normalized mRNA, protein and glycerol concentration values.

can predict the test data better than unsupervised data analysis methods which do not make use of any biochemical knowledge. Therefore, deficiency of a kinetic model in prediction compared to prediction by unsupervised data analysis methods tells us that the model cannot describe the data sufficiently well. A solid measure of this level of 'sufficiency' is needed by the biochemical modeling community because most of the time, we aim at the simplest model which is still competent in explaining the data as was also given as a guideline in [61]. On the other hand, it is very important to emphasize that this kind of comparison to unsupervised methods is only needed for the assessment of kinetic models' validity. We do not intend to underestimate the role of kinetic modeling by showing that there can be cases where unsupervised data analysis methods are superior to some kinetic models. Every kinetic model in systems biology is valuable and deserves attention just because they

aim at providing mechanistic explanations which the unsupervised data analysis methods in statistics lack. That independence from kinetic model structure is also exactly the reason why we used the predictive power of unsupervised data analysis methods as a reference point in this study. We used Smooth Principal Components Analysis for this purpose. SPCA offers better predictive capabilities than normal PCA since it can make use of also the underlying time profile and hence is more suitable for time series data. SPCA is also very robust against small changes in the smoothing parameter λ, proving to be a stable reference point.

With our simulations study using synthetic data generated by a toy model, we showed that until high amount of experimental noise in the data, cross validation SPCA prediction error can work as a threshold to invalidate a too simple kinetic model with high specificity and sensitivity. It is however important to note that for an accurate comparison of predictive power, the inferred parameters of the kinetic model have to be optimal. Although proven to be not an easy task, there are many methods proposed in the literature to overcome the local minima problems encountered [110, 112, 150] during parameter inference.

Forecast analysis requires higher penalties for smoothing of the scores in SPCA and noise is more influential. Predictions by SPCA forecasting and kinetic modeling are more dependent on the noise realization in the data compared to cross validation with interior time points. Therefore, we need to be more aware of the estimated noise level in the data if we want to use SPCA forecasting prediction error as an invalidation measure.

Our SPCA-based invalidation approach can also be employed iteratively for model reduction. Analyses of model families derived from a master model has proved to be a popular approach in biochemical modeling [47, 61, 90, 128]. In this approach, a master model is allowed to be manipulated in certain directions, either by changing the interactions and the species involved or changing the kinetic laws of the model. By this way, a very high number of models with very different number of parameters are created and analyzed. Here, selection of the best model within the large family of models is a critical task. Our invalidation approach can be very useful in that stage. The most complex models within the model family can be questioned first for their validity. Later, they can be subject to step-wise simplification by removal of interactions or simplification of reaction kinetics. At a certain stage, the models would be invalidated by our approach meaning that they fail to explain the data sufficiently well. This would mean that the models are in their simplest acceptable form one step before the invalidation decision. However, at that step there would still be a number of models with different characteristics which could not be invalidated. Therefore, the problem of model invalidation turns to a

problem of model selection between a number of models with similar complexities. Therefore, at that point, we can make use of model selection criteria like AIC or BIC complementary to our invalidation approach for the ultimate selection of the best model.

Acknowledgements

This project was financed by the Netherlands Metabolomics Centre (NMC) which is a part of the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research. The authors thank Maikel Verouden for the m-files performing SPCA.

3

Cross validation of kinetic models across different experimental conditions

Most ordinary differential equation (ODE) based modeling studies in systems biology involve a hold-out validation step for model validation. In this framework a pre-determined part of the data is used as validation data and, therefore it is not used for estimating the parameters of the model. The model is assumed to be validated if the model predictions on the validation dataset show good agreement with the data. Model selection between alternative model structures can also be performed in the same setting, based on the predictive power of the model structures on the validation dataset. However, drawbacks associated with this approach are usually under-estimated. We have carried out simulations by using a recently published High Osmolarity Glycerol (HOG) pathway from S.cerevisiae to demonstrate these drawbacks. We have shown that it is very important how the data is partitioned and which part of the data is used for validation purposes. The hold-out validation strategy leads to biased conclusions, since it can lead to different validation and selection decisions when different partitioning schemes are used. Furthermore, finding sensible partitioning schemes that would lead to reliable decisions are heavily dependent on the biology and unknown model parameters which turns the problem into a paradox. This brings the need for alternative validation approaches that offer

flexible partitioning of the data. For this purpose, we have introduced a stratified random cross validation (SRCV) approach that successfully overcomes these limitations. SRCV leads to more stable decisions for both validation and selection which are not biased by underlying biological phenomena. Furthermore, it is less dependent on the specific noise realization in the data. Therefore, it proves to be a promising alternative to the standard hold-out validation strategy. [1]

3

3.1. Background

Ordinary differential equation (ODE) based kinetic models are able to capture all of the available kinetic information regarding a biological system. Therefore, they are used extensively in systems biology especially for the purpose of predicting time dependent profiles and steady state levels of biochemical species in conditions where experimental data is not available. Examples from the literature show that there is a common path taken by the modeling community for the construction and the analysis of ODE based systems biology models. The first step is to define the model structure and the associated kinetics. Due to serious concerns about the validity of model structures and kinetics, many studies include the parallel development and analysis of multiple alternative model structures [100, 105, 106]. The second step is the estimation of the unknown model parameters by fitting the model to the data using global and local minimization algorithms. Data here are usually *in vivo* time series concentration data of the observable biochemical species included in the model. At this step, uncertainty in the estimated values of the model parameters can also be quantified by constructing confidence intervals [76, 82, 129]. Last but not least, models are assessed for the quality of their fit to the data and for their predictive power on independent data. Independent data are datasets that were not used for parameter estimation. Selection between alternative model structures can also be performed at this step. A complete modeling cycle includes all these steps to achieve sufficiently good models [37, 38].

A good model has to be sufficient both in explaining the data on which it was built and in predicting independent data [83]. The first is taken into account mostly by likelihood ratio tests which can be used to reject models based on the quality of fit to the data [29, 74, 115, 169]. The second aspect has been considered in conceptually two different ways. The first approach uses a penalized likelihood based metric such as Akaike's (AIC) [2] or Bayesian Information Criterion (BIC) [133]. This metric

[1] This chapter is based on:
D. Hasdemir, Huub CJ Hoefsloot, Age K Smilde. Validation and selection of ODE based systems biology models: how to arrive at more reliable decisions. *submitted for publication.*

is calculated using the whole dataset for parameter estimation but provides an expected value of the prediction error on an independent dataset. Therefore, it makes selecting the true complexity of a model possible because unnecessarily complex models are poor in predicting independent datasets. However, it is an 'in-sample' measure which means that the expected prediction error is valid only for the exact same experimental conditions as of the parameter estimation dataset [29]. Predicting the kinetics of the biological system under different experimental conditions is the very purpose of kinetic models, though. Therefore, modelers would like to show that the newly built model is good in qualitative or quantitative prediction of experimental data that was collected at different experimental conditions. This strategy which uses data at different experimental conditions as validation data constitutes the second approach to assess the prediction error [38, 78].

Different experimental conditions are usually based on the following scenarios:

- Inhibition of enzymes.
- Reduction of protein levels by RNAi mediated suppression.
- Gene deletions.
- Over-expression of genes in gene networks.
- Dose-response experiments in which different doses of triggering chemicals are used to stimulate the system.

These validation scenarios are popularly applied since the common goal of the modelers is to demonstrate the models' competency under challenging conditions. Estimating the parameters of a model in certain experimental conditions and showing their competency in other conditions within these scenarios requires multiple datasets under different conditions and, therefore, it is an example to the hold-out validation strategy. That is, a pre-determined set of conditions are held out of the training data and used as validation data instead. However, rules about the application of this strategy are not straightforward.

There are potential pitfalls associated with the application of hold-out validation strategies in the validation and selection of kinetic systems biology models. These arise due to the lack of a satisfying answer to the question: *which part of the dataset should be held out of the parameter estimation and instead should be used as the validation dataset?* We carried out simulations to demonstrate the phenomena that hinder us from giving satisfactory answers to this question which can also be referred to as the problem of selecting an appropriate hold-out partitioning scheme for the data. The problem arises due to incomplete biological knowledge of the

system and unknown true values of the model parameters. This makes the problem a paradoxical one since our knowledge about the system will never be complete and the true values of the model parameters are themselves what we are looking for. However, statistics literature offers an established method which is independent of this knowledge, namely cross validation.

Cross validation (CV) is a resampling method traditionally used for model selection, determining the optimal complexity of a model or assessment of its generalizability in statistics [43, 143]. It is based on the partitioning of the data in training and test sets. The training set is used to build the model and the predictions of the model on the test set are used for model assessment. Since the test set is completely independent of the parameter estimation process, selection will not be biased towards more complicated models. The efficiency of cross validation and its difference from hold-out validation strategy lies in the fact that the partitioning is made not in a pre-determined but in a random way and the procedure is repeated multiple times so that each partition can be used as test set at least once.

CV has been applied in different ways in the ODE based modeling framework. Partitions can consist of different experiments (such as different cell types, experimental conditions or cultures), data belonging to different biochemical species in the same experiment or different data points within the time profile of the same biochemical species. In [60], the authors present an example for the latter. In this work, prediction errors on test sets obtained by using an ODE based model are compared to the residuals from an unsupervised data analysis method which does not make any use of biochemical knowledge. Better predictions found by using the unsupervised principal components analysis (PCA) method give hints on the low informative level of the ODE model leading to a rejection of the proposed ODE model structure. CV by using different species from the same or different cultures with different experimental conditions was considered by [90]. In that study, prediction errors were used to select between two families of models each constituting from models of slightly changed topologies. Both approaches use a k-fold stratified partitioning scheme in which time points or species were approximately equally distributed between k different partitions. The prediction errors from different test sets are averaged for the final measure of the predictive power.

Existence of only very few examples like we mentioned above show that CV has been highly neglected in the field. Also, the risks associated with the hold-out validation strategy have been underestimated. The conceptual differences between the two methods and the difference between their outcomes have not been presented in detail. Therefore, with this study we aim to present a detailed comparison of the hold-out and cross validation methods by using simulations and emphasize the

advantages of CV over hold-out partitioning schemes. More details on our implementation of CV are given in the Methods section.

The reason for choosing simulations for our demonstrative purposes is that simulations and synthetic data allows us to know the ground truth, in this case the true model parameters and the true model structure. Therefore, we can analyze the results we achieved in different partitioning schemes in a comparative manner. We mainly look at the effect of different partitioning schemes in the outcome of model validation and selection. However, we report also results related to its effect on parameter estimation which is very influential on validation and selection in order to present a complete explanation.

3.2. Methods

3.2.1. Simulated data

We used the high osmolarity glycerol pathway model in *S.cerevisiae* taken from [128] (see Figure 3.1) to generate synthetic data. The pathway can be triggered by using an NaCl shock and is activated via two parallel upstream signaling routes. The activity of the upstream routes is encoded by a binary input parameter which indicates that the route is either active or not. The level of the NaCl shock is also an input parameter which can be manipulated. Therefore, the model can be used also for deletion mutants where only one of the signaling routes is active, following different doses of NaCl shock, by changing only those two input parameters. It includes additional 20 free parameters which can be estimated from data.

We mimicked the real experimental conditions used in [128] when generating the data. These include different cell types and different NaCl doses. The different cell types were deletion mutants in which only the signaling branch through Sln1 activation or Sho1 activation was active and the wild type cell in which both branches were active (Figure 3.1). The different NaCl shock levels ranged between 0.07 and 0.8 M (Figure 3.2). The data consisted mainly of the ratio of the active phosphorylated Hog1 protein to the maximum Hog1 protein level observed in the wild type cell which was expressed as a percentage. The Hog1 protein phosphorylation percentage data (Hog1PP data) from 3 cell types and 6 doses formed 18 different subsets of Hog1PP data. We used different subsets for parameter estimation and model validation/selection each time within different data partitioning schemes which we explain in detail in the following section. Concentration data of other species in the model were essential for the estimation of the parameters downstream from the Hog1 protein. For this reason, measurements of mRNA, protein and glycerol levels at 0.5 M. NaCl shock were always a part of the training dataset. Therefore, the

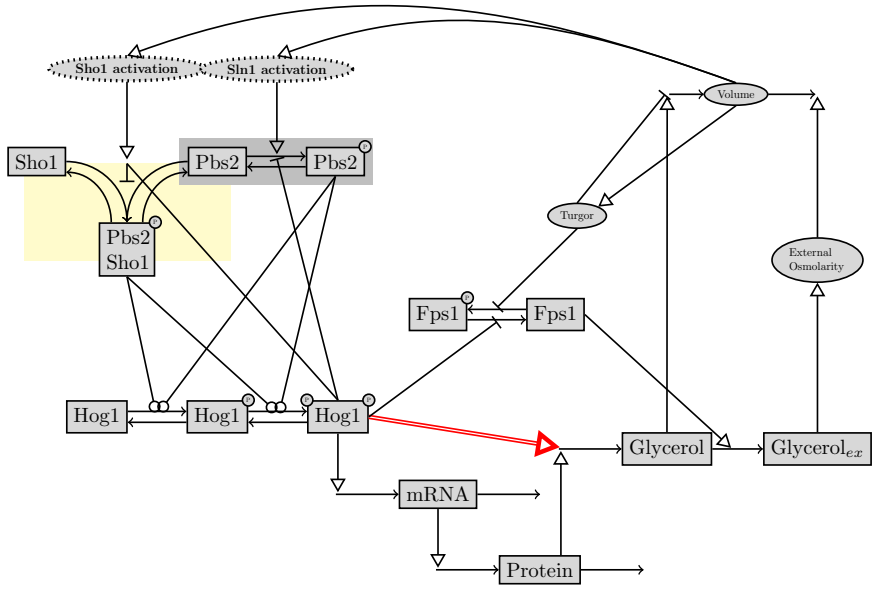

Figure 3.1: **The pathway topology proposed in [128]**. We used this model as our true model and generated data based upon it. The black lines with small arrow tips depict the transition between different species in the model like production, degradation or complex formation. The black lines with open circle tips depict the phosphorylation process by kinases. The lines with open triangle tips show activating regulatory interactions whereas lines with blunt ends show deactivating regulatory interactions. The red colored double arrow denotes the post translational regulation of glycerol production by the active phosphorylated Hog1 protein. We did not consider this regulatory interaction in our simplified model. The dotted ellipses in the upper left hand corner indicate the two different upstream activation routes important in our study. Parts of the pathway whose parameters were affected by the choice of the partitioning scheme were highlighted yellow and gray. We explained the changes in the parameters of those regions in our results section. (Figure adopted from [60].)

terms 'validation data' and 'training data' refer only to Hog1PP data, throughout the text.

We generated 100 different realizations of synthetic data by adding error to the time profiles obtained by the model. We added heterogeneous noise where the noise term for each concentration value was drawn from a normal distribution with a standard deviation equal to 10% of the concentration value itself which reflects realistic noise levels and structure for these type of experiments. The time series data contained 15 time points during a course of 160 minutes.

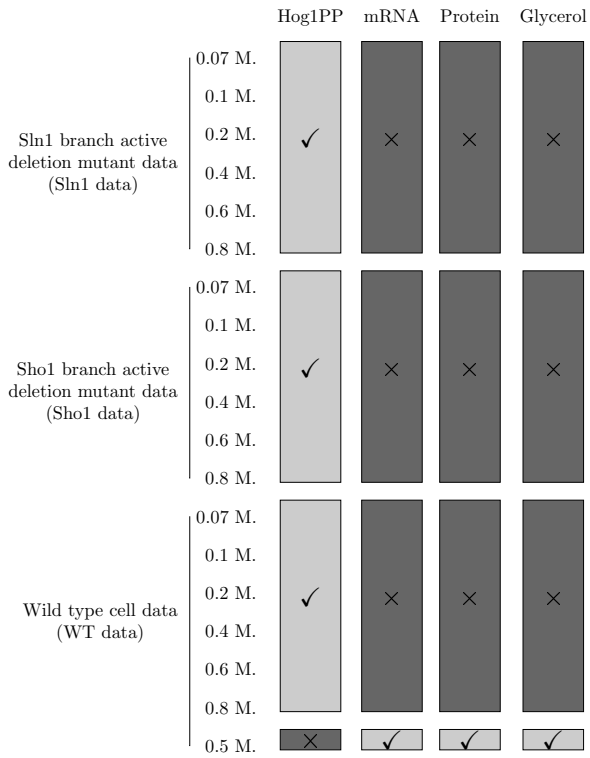

Figure 3.2: **Experimental conditions under which the data was generated.** Check marks indicate the measurements that were performed. Each row shows a different dose in a different cell type whereas columns are for different biochemical species measured. Hog1PP data consists of 18 subsets (6 different doses and 3 different cell types) and is the main subject of variability between different partitioning schemes that we evaluated.

3.2.2. Data partitioning schemes

Hold-out partitioning schemes

In this work, we evaluate the performance of the hold-out validation partitioning schemes based on two most popularly applied challenge scenarios: gene deletions and dose-response experiments. The first scenario in our study mimics a gene deletion challenge. In each scheme of this scenario (Figure 3.3), the training set is composed of all six doses of a single cell type. All six doses of the other two cell types can be used as validation data. The outcomes of model validation and selection are determined based on each of these twelve different subsets of validation data, separately. The schemes are named throughout the chapter as Sln1, Sho1 and WT schemes depending on the cell type used for training.

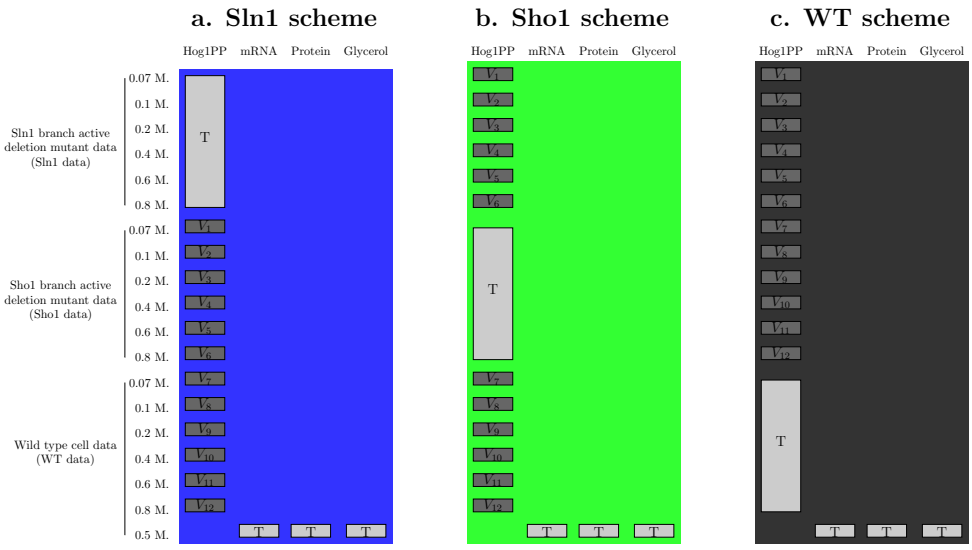

Figure 3.3: **Scenario 1 partitioning schemes.** Light gray colored boxes show parts of the data which we used as the training set (T) for parameter estimation. Dark gray colored boxes show parts which we used as validation sets (V). Different background colors represent different partitioning schemes and are consistent with the colors used in the graphs in the Results and Discussion section. Each partitioning scheme offered the use of six subsets of the Hog1PP data as the training set and the remaining twelve subsets of the Hog1PP data could be used for validation, separately.

Our second scenario mimics the dose-response strategy. In this scenario, the training set is composed of one dose from each cell type. In the lowest dose scheme, only the data following a 0.07 M. NaCl shock are used for training (Figure 3.4). In the highest dose scheme, only the data following a 0.8 M. NaCl shock are used for training. The remaining five doses from each cell type can be used as validation data. Similar to the first scenario, the outcomes of model validation and selection are determined based on each of these fifteen subsets of validation data, separately.

Lastly, we introduce variation in the training sets. We update our first scenario in such a way that in each partitioning scheme (Figure 3.5a-c), we use data from two cell types for training. All six doses from the remaining cell type can be used as validation data. We make consensus decisions on model validation and selection considering all the validation subsets. The schemes are named throughout the chapter as Sln1/Sho1, Sln1/WT and Sho1/WT schemes depending on the pair of training cell types. We update our second scenario in such a way that in each partitioning scheme (Figure 3.5d-e), we use either data from the four highest or four lowest doses from each cell type for training. The remaining two doses from each cell

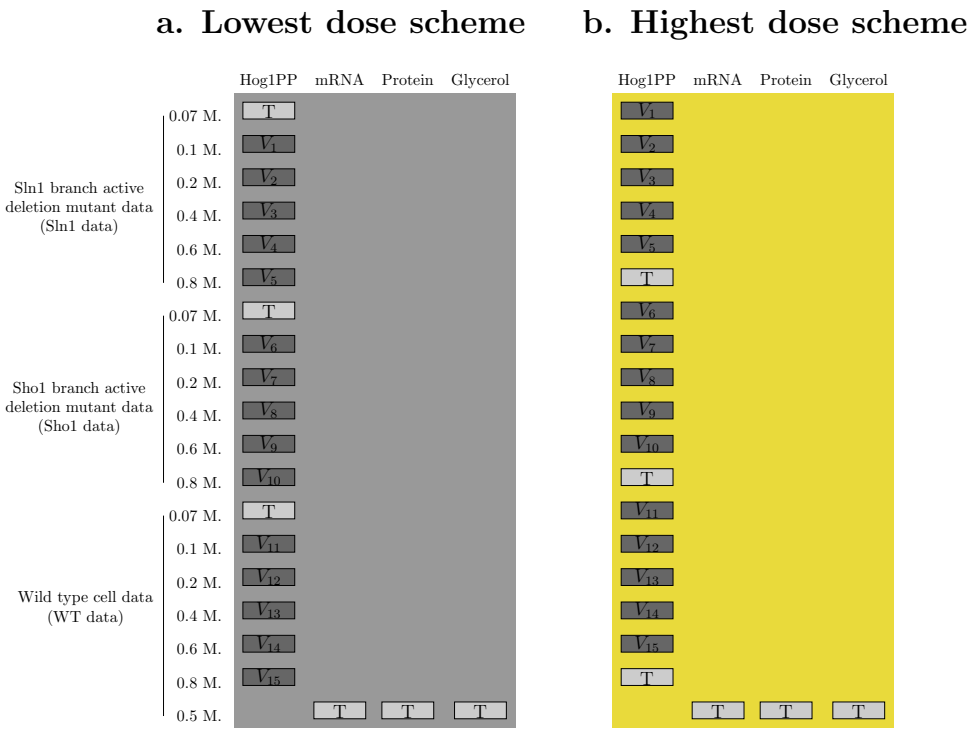

Figure 3.4: **Scenario 2 partitioning schemes.** Light gray colored boxes show parts of the data which we used as the training set (T) for parameter estimation. Dark gray colored boxes show parts which we used as validation sets (V). Different background colors represent different partitioning schemes and are consistent with the colors used in the graphs in the Results and Discussion section. Each partitioning scheme offered the use of three subsets of the Hog1PP data as the training set. These are the lowest dose subset of each cell type in the lowest dose scheme and the highest of each in the highest dose scheme. The remaining fifteen subsets of the Hog1PP data could be used for validation, separately.

type can be used as validation data. Similar to the first updated scenario we make consensus decisions using all validation subsets at once. The schemes are named as low doses and high doses schemes based on the doses used in the training set. This way, we can obtain five schemes (Figure 3.5) each of which uses twelve subsets of the phosphorylated Hog1 (Hog1PP) data for training and the remaining six subsets for validation. Therefore, these five schemes can be compared to the stratified cross validation scheme which also makes use of twelve subsets of Hog1PP data in each training set.

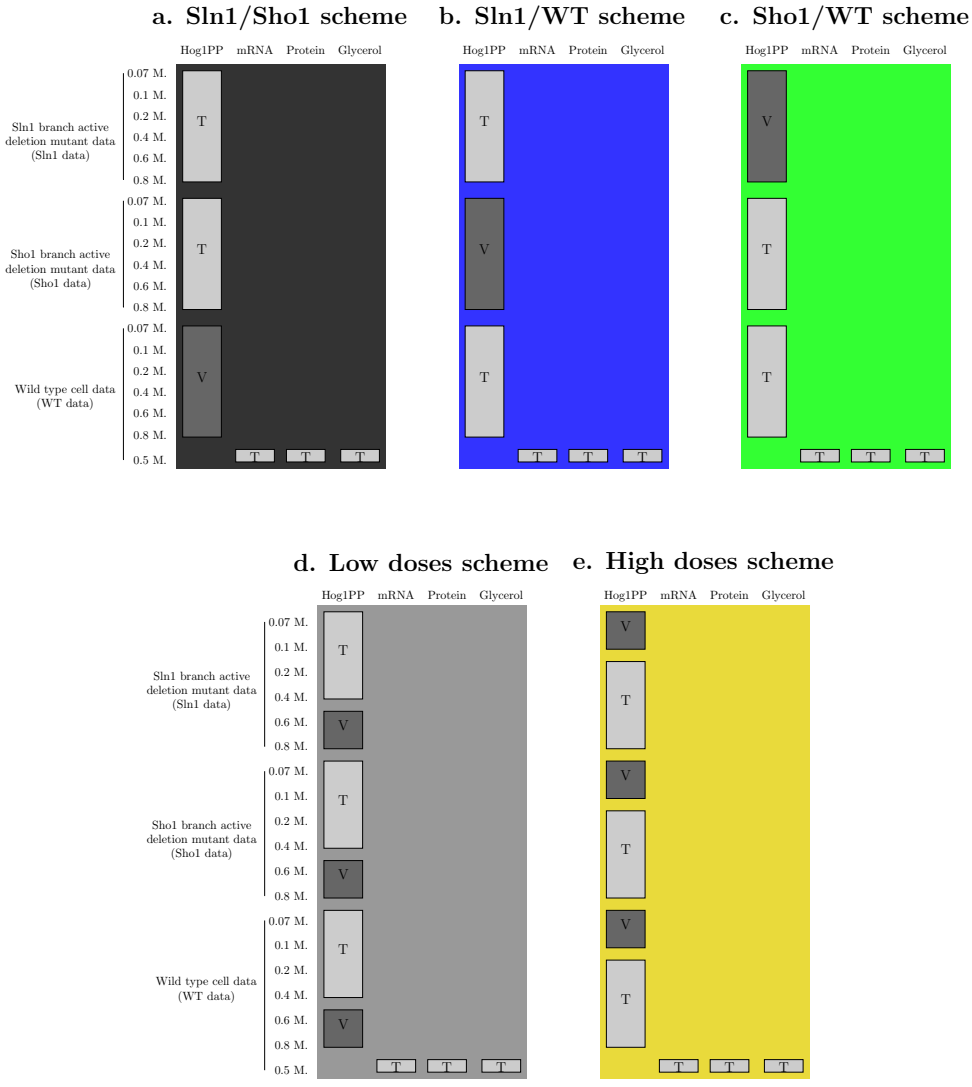

Figure 3.5: **Partitioning schemes used in the adapted scenarios.** Light gray colored boxes show parts of the data which we used as the training set (T) for parameter estimation. Dark gray colored boxes show parts which we used as the validation set (V). Different background colors represent different partitioning schemes and are consistent with the colors used in the graphs in the Results and Discussion section. Each partitioning scheme offers the use of twelve subsets of the Hog1PP data as the training set and the remaining six subsets as the validation set.

Stratified random cross validation scheme

In a random cross validation scheme, there are no pre-defined partitions, unlike the hold-out partitioning schemes. Here, we implement stratified random cross

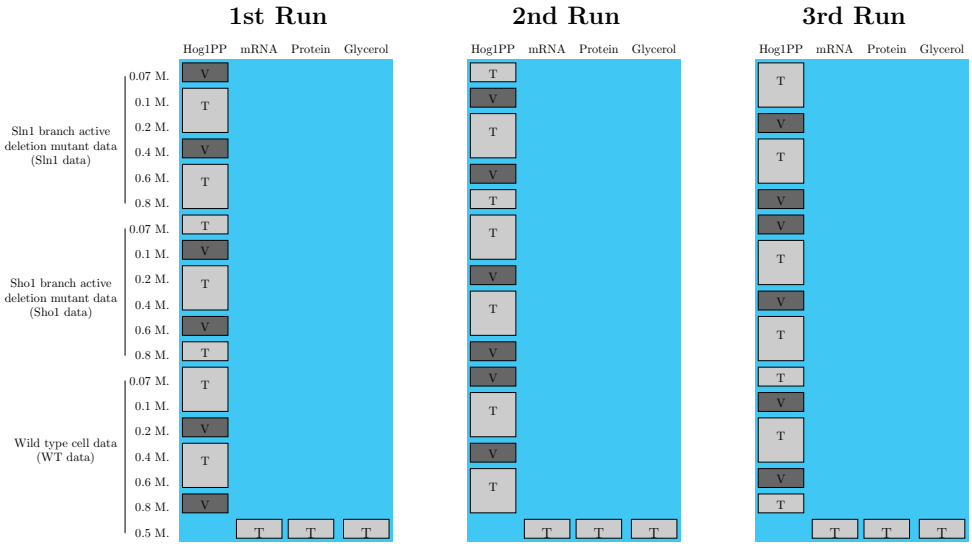

Figure 3.6: **Stratified random cross validation scheme (SRCV).** Light gray colored boxes show parts of the data which we used as training sets (T) for parameter estimation. Dark gray colored boxes show parts which we used as validation sets (V). In each of the three runs, the training and the validation sets change as indicated in these graphs.

validation which is a specific type of cross validation in which the training sets can be forced to follow a certain structure. We randomly partition the data into training and validation sets, in three different runs. In each run, we force the training sets to include the same amount of data from each cell type and dose level. We estimate the parameters in each run and also calculate the percentage prediction errors (PE) of the true and the simplified models and the model separation (ΔTS) on the validation data. Later, we make consensus decisions using the average PE and ΔTS from all three different validation datasets of different runs. The different partitioning schemes applied in each run can be seen in Figure 3.6.

3.2.3. Measures used for the analysis of the simulations

For each partitioning scheme, we estimated the parameters of both the true model (the model that we accepted as the ground truth and generated the data upon that) and the simplified model (the model that lacked one of the important regulatory interactions in the true model which can be seen in Figure 3.1). We repeated the parameter estimation process by using 100 different realizations of the data. The estimation of the parameters required the minimization of the difference between the data and model predictions. We carried out the minimization using the local

minimizer 'lsqnonlin' function of Matlab [31, 32]. We considered a local optimizer to be sufficient since we work with generated data and could use the true values of parameters as starting points. We analyzed four main features from the simulations, namely the amount of bias in the parameter estimates, the predictive power of the models on the validation datasets, the number of wrong decisions in which the simplified model structure was selected over the true model structure and the distance between the predicted profiles by the true and the simplified model structures (model separation).

We use normalized bias (nBi) as a measure of the bias in each estimated parameter (Equation 3.1). The median of its distribution across different noise realizations gives us the median amount of bias in each parameter estimated in a certain scheme.

$$nBi_j^i = \frac{|p_j^i - p_{true}^i|}{p_{true}^i} \times 100$$
$$i\text{=1:20 index for parameters in the model}$$
$$j\text{=1:100 index for noise realizations} \tag{3.1}$$

We quantify the lack of good predictive power of models by using percentage errors. Percentage error is the percentage of the sum of squares of the prediction error to the sum of squares of validation data (Equation 3.2). Model selection between the two models gives wrong results when $PE_T > PE_S$, meaning that the simplified model gives lower prediction error than the true model structure.

$$PE = \frac{\sum_i^I (\frac{\sum_j^{15} (x_{ijk} - \hat{x}_{ijk})^2}{\sum_j^{15} x_{ijk}^2} \times 100)}{I}$$
$$i\text{=1:I index for validation subsets of Hog1PP data}$$
$$j\text{=1:15 index for time points}$$
$$I = \text{total number of Hog1PP subsets used for validation} \tag{3.2}$$

The difference between the true and the simplified model predictions (ΔTS) can be calculated by using the trapezoidal rule as in Equation 3.3. With this method, the area between two curves can be approximated as a series of trapezoids (see Figure 3.7). The sum of the areas of the trapezoids provide a good approximation of the area between the curves when the number of trapezoids are sufficiently high. Here, the two curves are the profiles of the Hog1PP predicted by the true and the simplified model structures. We normalize the calculated area with respect to the maximum of the Hog1PP data in the corresponding validation subset. Large areas between the

two curves mean that the separation of the two model structures is easier. Therefore, when correct model selection decisions are given, model separation (ΔTS) can be used as an additional criteria of enhanced model selection.

$$\Delta TS_i = \frac{\sum_k^{K-1} \frac{|T(t_{k+1})-S(t_{k+1})|+|T(t_k)-S(t_k)|}{2}.(t_{k+1}-t_k)}{max(x_{ij})}$$

$$\Delta TS = \frac{\sum_i^I \Delta TS_i}{I}$$

T: numerical values of the Hog1PP predictions by the true model structure

S: numerical values of the Hog1PP predictions by the simplified model structure

k=1:K-1 index for trapezoids

i=1:I index for validation subsets of Hog1PP data

j=1:15 index for time points

I = total number of validation subsets (3.3)

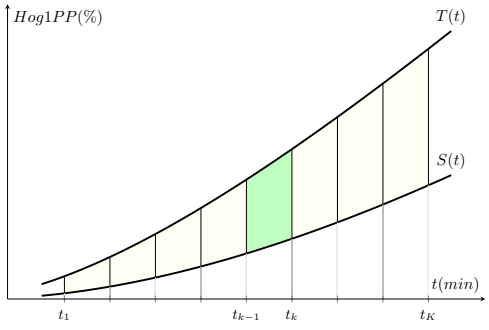

Figure 3.7: **Trapezoidal Rule.** The figure explains the trapezoidal rule visually. The green shaded area refers to the area of the $(k-1)^{th}$ trapezoid. The total area of the trapezoids is equal to ΔTS_i in Equation 3.3.

3.3. Results and discussion

3.3.1. Scenario 1: partitioning of data from different cell types

Firstly, we would like to stress that in all of our simulations, we observed very good fit of the true model structure to the data. Additionally, our emphasis in this work is on model validation and selection using validation datasets which were excluded from the training set. Therefore, we do not present detailed analysis of the quality

of model fits. Only in Figures 3.8 and S2, we present the model fits together with the predictions in two examples. We should also mention that the term 'prediction' always refers to predictions on validation datasets, throughout the text. Finally, we present our results on both percentage error (PE) and model separation (ΔTS) using a box plot representation. With this representation, each box plot shows the distribution of the associated measure across the 100 different noise realizations. For example, in the case of percentage error, the median of this distribution gives an idea on how high the prediction errors are in general. In addition, the box plots show also the outliers with relatively high prediction errors by the red points outside the boxes.

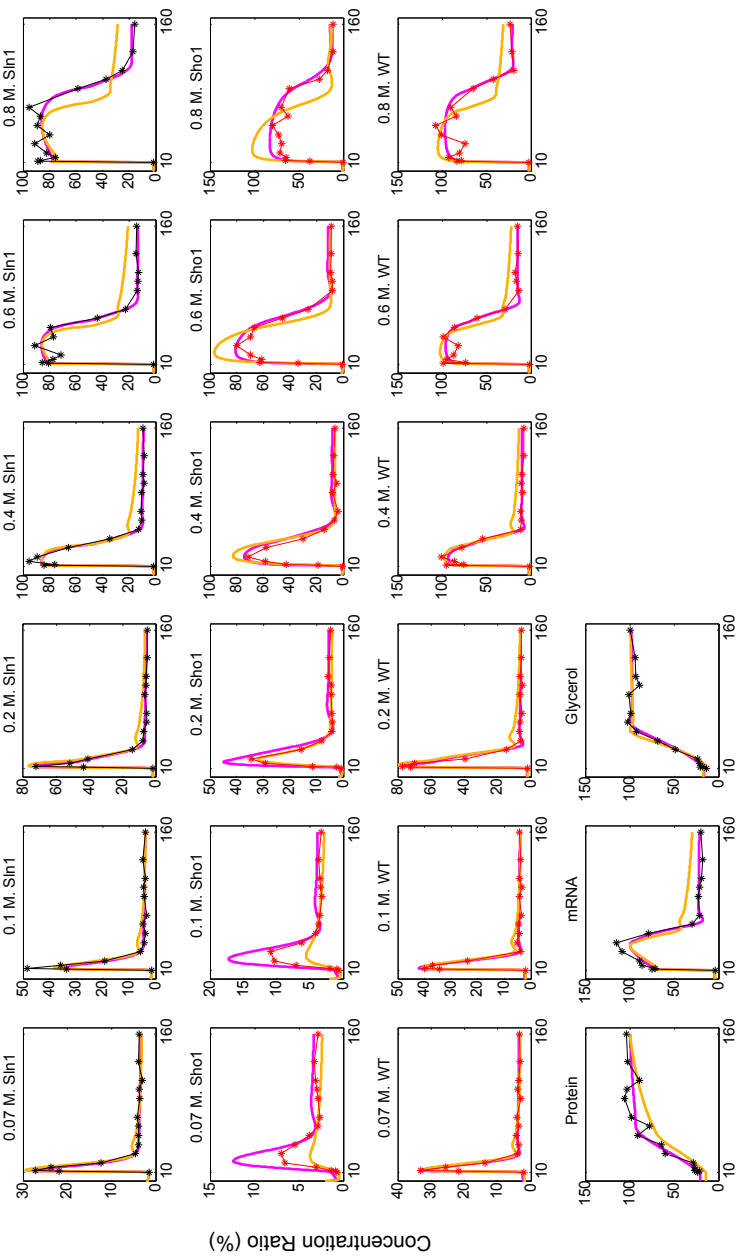

Figure 3.8: **Fit and predictions obtained on a single realization of data in the Sln1 scheme.** Black and red points (connected by lines of the same color) refer to data points which were used for parameter estimation and validation, respectively. In this example, all doses of Sln1 data and the data on the downstream species (protein, mRNA and internal glycerol) were used for parameter estimation. The magenta lines show the profiles obtained (both fit and prediction) by using the true model for the parameter estimation. The orange lines belong to the profiles obtained by the simplified model structure. All concentrations are given in percentages. The top three rows are for the Hog1PP data. The titles for each graph show the dose and the cell type related to the experiment in which the Hog1PP data was collected. The last row of graphs give the concentration ratios for the downstream species. The associated data was collected in a single experiment with WT cells following a 0.5 M. NaCl shock.

When only data from the Sln1 branch active deletion mutant (Sln1 data) is used for parameter estimation, validation using the Sho1 branch active deletion mutant data (Sho1 data) can be very misleading. This is because models trained by using Sln1 data results in bad predictions on the Sho1 data. On the other hand, the same models can achieve reasonable predictions on the WT data (See Figure 3.8 for an example). This can be seen from the distribution of the percentage prediction errors represented by box plots for each validation set in Figures 3.9a and 3.9b.

Existence of realizations with very high prediction errors in box plots with low medians shows that extremely bad predictions can occur even when the median prediction error is not very high. Examples of this can be observed also in the Sho1 scheme shown in Figures 3.9c and 3.9d. Indeed, the models trained by using only the Sho1 data can lead to extremely high prediction errors both on Sln1 and WT data. This can be seen from the existence of realizations with a percentage prediction error above 30% and 15%, respectively. On the other hand, models trained by using only the WT data perform well in predicting the Sln1 data but not the Sho1 data (maximum of medians = 1.11% vs 4.20% in Figures 3.9e and 3.9f). However, they are still better than those obtained by the models trained by the Sln1 data (maximum of medians = 4.20% vs 8.85% in Figures 3.9f and 3.9a).

As a summary of the observations on the predictive power, we can say two things. Firstly, models trained by using only the data from one of the deletion mutants is poor in predicting the data from the other. Secondly, models trained by using the data from the WT cell can predict the data from one of the deletion mutants better than the other one. The poor predictions might easily lead to misleading decisions on model validation. True model structures might fail to be validated due to weak predictive power of some partitioning schemes. To study the reasons leading to weak predictive power we investigated the parameter estimation quality.

We measured the parameter estimation quality by using the normalized bias of each parameter. The median of this measure across all noise realizations shows how well the parameter was estimated in general in a certain scheme. In Figure 3.10, we see that the parameters related to the complex formation of Sho1 and Pbs2 proteins and this complex' phosphorylation, p8 and p9, were predicted with very high bias in the Sln1 scheme. (see the yellow region in Figure 3.1). This means that when the Sln1 data is used for model training, we estimate the Sho1 branch parameters with a very high uncertainty with a median bias of 31% and 33%, respectively. The same reasoning is valid also for the estimation of two of the parameters related to the phosphorylation of the Pbs2 protein, p4 and p5. The median bias for these parameters (see the gray region in Figure 3.1) were found to be 17% and 14%, respectively. There is an interesting difference between the estimation quality of the

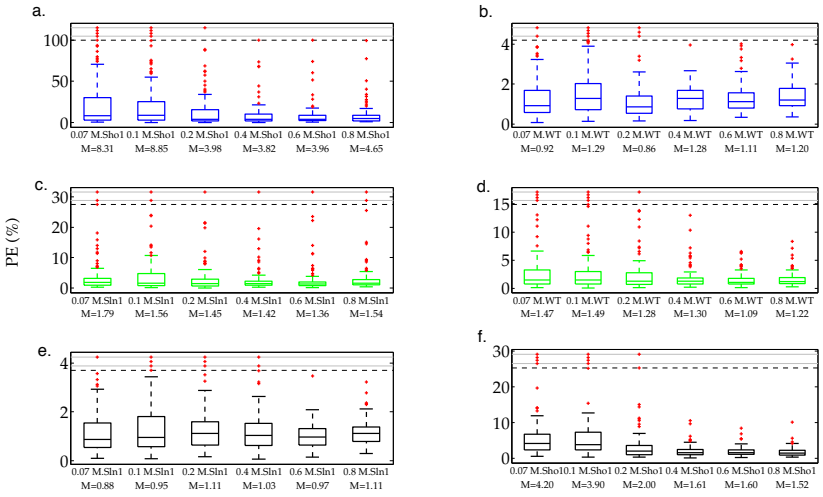

Figure 3.9: **Percentage prediction errors (PE) of the true model structure in scenario 1.** Each box plot shows the distribution of PE over 100 different realizations of the data. The red dots indicate the outliers which lie outside approximately 99.3% coverage if the data is normally distributed. They indicate realizations with relatively higher PE. Blue, green and black boxes refer to Sln1, Sho1, and WT schemes. Each row in the figure corresponds to a single scheme. The labels on the x-axis show the specific dose and the cell type of the data on which the validation was performed. The labels indicate also the medians of the PE distribution summarized visually by the box plots. In each graph, the ten realizations with the highest PE are located above the black dashed line. The region above this line is compressed for visual ease. **a.** PE obtained on Sho1 validation subsets in the Sln1 scheme. **b.** PE obtained on WT validation subsets in the Sln1 scheme. **c.** PE obtained on Sln1 validation subsets in the Sho1 scheme. **d.** PE obtained on WT validation subsets in the Sho1 scheme. **e.** PE obtained on Sln1 validation subsets in the WT scheme. **f.** PE obtained on Sho1 validation subsets in the WT scheme.

parameters in the two different branches, though. We could decrease the bias of the Sln1 branch parameters considerably when we used the WT data for training the model. However, the level of bias in the Sho1 branch parameters was still relatively high in the WT scheme compared to the Sho1 scheme. Therefore, the Sln1 data could be predicted well in the WT scheme whereas the prediction of the Sho1 data was still problematic. As a further investigation on the system dynamics, we tuned one of the branch parameters each time within a range limited by the minimum and maximum of their estimated values. This allowed us to confirm the deteriorating effect of biased branch parameters on the predictions (data not shown).

The asymmetrical behavior of the predictive power (that is, the WT and the Sln1

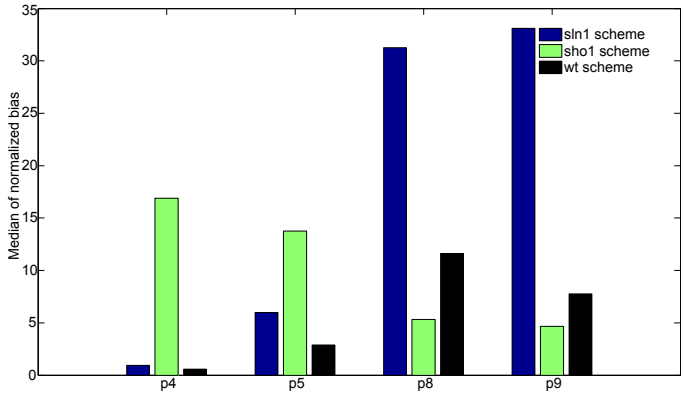

Figure 3.10: **Normalized bias of branch parameters.** Bar graphs show the median of the normalized bias of parameters across all noise realizations. Only some of the branch parameters are shown in the figure. Parameters p4-p5 play a role in the Sln1 branch and parameters p8-p9 are in the Sho1 branch. Blue, green and black refers to the Sln1, Sho1 and WT schemes respectively.

schemes were good in predicting Sln1 and WT validation data, respectively but none of them could achieve good predictions on the Sho1 validation dataset) stems from an underlying biological property which is the inequality of the two phosphorylation branches in the model. Although the two branches (Figure 3.1) act redundantly for the ultimate goal of Hog1 protein phosphorylation, the fluxes in each branch are not equal. As also mentioned in [128], the Sho1 branch active deletion mutant produces less output in terms of phosphorylated Hog1 protein. This biological fact manifests itself also in the data. The WT data is characterized more by the activity in the Sln1 activation branch rather than the Sho1 branch. In other words, the Hog1PP levels in the WT cell are affected more by the changes in the Sln1 branch parameters than by the changes in the Sho1 branch parameters. Therefore, the WT data can substitute for the Sln1 data for training the models. However, the cost of excluding the Sho1 data from the training set is higher due to the asymmetry we mentioned above. The Sho1 branch parameters are weakly identifiable when the Sho1 data is not used for parameter estimation. This asymmetry in the information content of the data is clearly the output of the pathway machinery. This machinery is summarized into a model with a model structure and parameter values. Therefore, the decisions of model validation using a hold-out strategy is dependent on the underlying biological properties (the asymmetrical branch structure in this particular example) and reflections of these properties in the model parameters (parameter values that allow less flux in one of the branches). The data partitioning task, hence, proves to be a difficult one since the prior knowledge about the underlying biology would never be

complete.

Another important observation is the NaCl dose dependency of the predictive power using the Sho1 data. The predictive power using Sho1 data was especially lower in the lowest two doses compared to the higher doses as can be seen in Figures 3.9a and 3.9f (maximum of the medians 8.85% vs. 4.65% in the Sln1 scheme and 4.20% vs. 1.52% in the WT scheme).

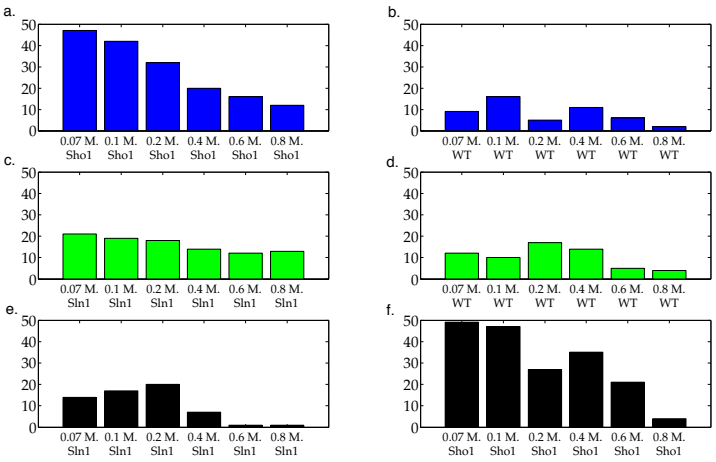

Figure 3.11: **Number of wrong decisions in scenario 1.** Bars show the number of realizations in which the simplified model gave lower residuals than the true model structure and therefore, was wrongly selected over the true model structure. Blue, green and black bars refer to Sln1, Sho1, and WT schemes. Each row in the figure corresponds to a single scheme. The labels on the x-axis show the specific dose and the cell type of the data on which the validation was performed. **a.** Number of wrong decisions using Sho1 validation subsets in the Sln1 scheme. **b.** Number of wrong decisions using WT validation subsets in the Sln1 scheme. **c.** Number of wrong decisions using Sln1 validation subsets in the Sho1 scheme. **d.** Number of wrong decisions using WT validation subsets in the Sho1 scheme. **e.** Number of wrong decisions using Sln1 validation subsets in the WT scheme. **f.** Number of wrong decisions using Sho1 validation subsets in the WT scheme.

The asymmetry in the contribution of the Sln1 and the Sho1 branches to the phosphorylation of the Hog1 protein also has consequences for model selection. Figure 3.11 shows the number of wrong decisions given on each validation subset in each of these three partitioning schemes. We see that in a high number of realizations, the simplified model structure was selected over the true model structure when the Sho1 data was used for validation (Figures 3.11a and 3.11f). On the other hand, using only the Sho1 data for training also resulted in an increased number of wrong decisions on the Sln1 data compared to the WT scheme (minimum number of wrong

decisions 12 vs. 1 in Figures 3.11c and 3.11e).

In this section, we focused on partitioning schemes which use the data from only one cell type for parameter estimation. Our results show the importance of having a variety of different validation sets. This is because decisions of model validation and selection vary considerably depending on the experimental conditions of different validation sets due to unknown values of the underlying parameters.

3.3.2. Scenario 2: partitioning of data in different doses

In the second scenario, where we use different doses as training sets, we see a change of predictive power on Sho1 validation data (see Figures 3.12b and 3.12e) . When the lowest dose data from all three cell types are used for training, the predictive power on the Sho1 data decreases with increasing dose (median 1.11% vs. 3.16% on 0.1 M. and 0.8 M. dose levels, shown in Figure 3.12b). Also, when the highest dose scheme is used, the predictive power increases with increasing dose (median 2.59% vs. 1.11% on 0.07 M. and 0.6 M. dose levels, shown in Figure 3.12e).

These results showed us that predictive power becomes lower with increasing distances between the training and the validation sets, where the distance is measured in terms of the dose of the triggering chemical . This means that the risk of invalidating the true model structure increases when the validation set is too distant from the training set . However, the limits between which the model parameters stay applicable depend very much also on the cell type as we have observed. The predictive power on the Sho1 data deteriorated more rapidly compared to the other cell types. These observations helped us to identify a serious pitfall of dose-response strategy: as long as we do not have realistic prior information on the limits for which we expect the estimated values of the model parameters to be applicable, we face the risk of invalidating a true model structure by over-challenging the model. Unfortunately, determination of the limits is not possible beforehand since it depends on the underlying biological properties which will never be completely known.

When it comes to model selection, we face a different challenge. Figure 3.13 shows the number of realizations in which the simplified model structure was selected over the true model structure. For example, the lowest dose scheme results in 22 wrong decisions whereas the highest dose scheme results in only 2 wrong decisions when the 0.1 M. Sln1 dataset is used as validation dataset as can be seen in the upper left hand side corner of Figure 3.13. Here, only the results on validation sets that can be used in both schemes are shown because our focus is on comparing the performance of two different schemes on shared validation sets. The most important observation from the figure is that the number of wrong decisions by the lowest scheme is higher on the 0.1 M. - 0.2 M. Sln1 and WT data compared to the

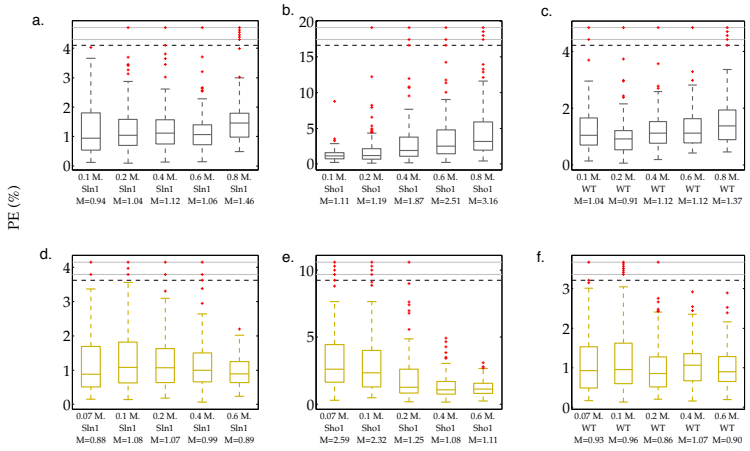

Figure 3.12: **Percentage prediction errors (PE) of the true model structure in scenario 2.** Each box plot shows the distribution of PE over 100 different realizations of the data. The red dots indicate the outliers which lie outside approximately 99.3% coverage if the data is normally distributed. Gray and yellow boxes refer to the lowest and the highest dose schemes, respectively. Each row in the figure corresponds to a single scheme. The labels on the x-axis show the specific dose and the cell type of the data on which the validation was performed. The labels indicate also the medians of the PE distribution summarized visually by the box plots. In each graph, the ten realizations with the highest PE are located above the black dashed line. The region above this line is compressed for visual ease. **a.** PE obtained on Sln1 validation subsets in the lowest dose scheme. **b.** PE obtained on Sho1 validation subsets in the lowest dose scheme. **c.** PE obtained on WT validation subsets in the lowest dose scheme. **d.** PE obtained on Sln1 validation subsets in the highest dose scheme. **e.** PE obtained on Sho1 validation subsets in the highest dose scheme. **f.** PE obtained on WT validation subsets in the highest dose scheme.

highest dose scheme. The number of wrong decisions by the lowest scheme is very high (22 and 30 on the Sln1 and WT validation data, respectively) especially on the 0.1 M. dose which is very close to the 0.07 M. dose where the models were trained. In addition, we see that the highest scheme gives a slightly higher number of wrong decisions compared to the lowest dose scheme on the 0.6 M. Sln1 and WT data. These observations suggest that model selection is problematic when the training and validation sets are too close to each other. We looked further at the model separation between the true and the simplified models (see Figure 3.14) to investigate the separation between the two model structures in higher resolution.

In cases where the differences between the number of wrong decisions is too low for a meaningful comparison, the model separation, ΔTS is more informative. Figure 3.14 shows the percentage of the realizations in which one specific scheme resulted

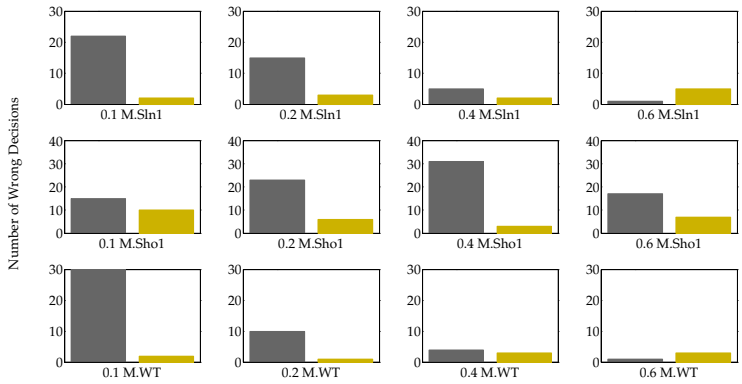

Figure 3.13: **Number of wrong decisions in scenario 2.** Bars show the number of realizations in which the simplified model gave lower residuals than the true model structure and therefore, was wrongly selected over the true model structure. Gray and yellow bars refer to the lowest and the highest dose schemes. The labels on the x-axis show the specific dose and the cell type of the data on which the validation was performed. Here, only the twelve validation subsets which could be used in both the lowest and the highest schemes are shown.

in better model separation than the other scheme. The percentages are based on the number of realizations in which a correct decision was made by using both the lowest and the highest dose schemes. For example, we know that both schemes result in a correct decision in 77 realizations of the Sln1 data at 0.1 M. NaCl shock (data not shown). The first pie chart in Figure 3.14 shows that in 99% of these 77 realizations, the highest dose scheme resulted in better separation between the two model structures than the lowest dose scheme. As can be seen from this figure, model separation obtained by the highest dose scheme is higher than that obtained by the lowest dose scheme in almost all realizations of 0.1 M. - 0.2 M. Sln1 and WT data. At 0.6 M. dose, the situation is reverse and the lowest scheme provides a better separation of the two model structures, in most of the realizations of all three cell types. These findings support the observation we made from the number of wrong decisions: model selection becomes problematic with too close training and validation sets. This is mainly because the simplified model might also predict well in the close proximity of the training dose (See Figure S1). However, it will perform worse than the true model structure as the training and validation sets become more distant from each other. However, too much distance can also pose a problem for model selection due to increased uncertainty in the predictive power. Uncertainty in the predictions shows that different noise realizations can either give very good or

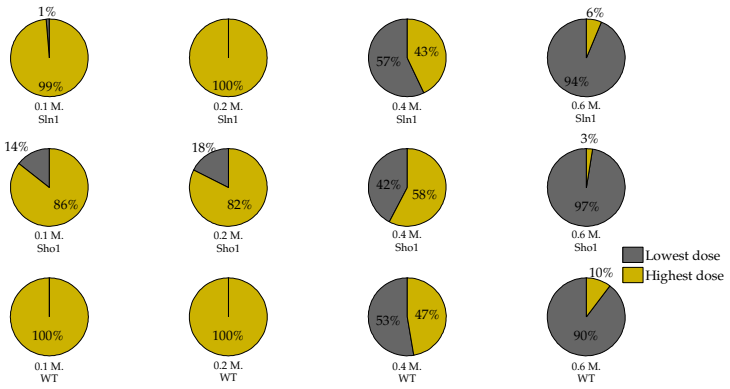

Figure 3.14: **Comparison of model separation by the two schemes.** Each pie chart shows the percentage of correct decisions, where the model separation achieved by a certain scheme is better than the other scheme. The labels on the x-axis show the specific dose and the cell type of the data on which the validation was performed. Gray and yellow colors in the charts refer to the lowest and the highest dose schemes. Here, only the twelve validation subsets which could be used in both the lowest and the highest schemes are shown. For example, when the 0.1 M. Sln1 data was used for validation, in 99% of the realizations in which a correct decision was given by both schemes, higher distance between the predictions of the true and the simplified model structures was achieved in the highest dose scheme than the lowest dose scheme.

very bad predictions. High levels of uncertainty reveals itself in the wide box plots of especially Sho1 validation data in Figures 3.12b, 3.12e and S1, showing a wide dispersion of predictive power across different noise realizations. In these regions where the uncertainty is high, it becomes more difficult to anticipate the predictive powers of the true and the simplified model structures on a single noise realization. This hampers also model selection. Using Sho1 validation data results in such a situation where uncertainty is very high at certain doses. This is why the trends in model selection that we have presented in this section cannot be observed on the Sho1 validation data as sharply as on the other cell types.

We can understand the risks associated with high uncertainty in a hold-out strategy, if we remember that in a single real experiment we have only one realization of noise. The outcomes of both model validation and selection depend highly on the specific noise realization in the data but we have only one realization available. This means that it is highly probable that we end up in wrong decisions just due to experimental noise. Therefore, we need partitioning schemes that minimize the effect of idiosyncratic noise realizations and lead to similar decisions for all of them. The stratified random cross validation scheme is promising in this sense as we will

explain in the following section.

3.3.3. Introducing variation in the training and the test data

In the previous sections, we showed the pitfalls that we might come across if we use single doses or single cell types as validation data. Therefore, we stress the importance of consensus results obtained from a collection of different validation sets. In this section, we take it one step further and introduce variation of experimental conditions also in the training data. We do it in three different ways as described by the adapted scenarios and the stratified cross validation scheme in the Methods section. First, we include two different cell types in the training data, namely in the Sln1/Sho1, Sln1/WT and Sho1/WT schemes. Second, we include four different doses from each cell type in the training data, namely in the low doses and the high doses schemes. These are examples of hold-out validation strategies just like the previous two scenarios. However, unlike those, the training and the validation sets include a variety of different cell types or doses. The third way is not an example of a hold-out strategy. It is the stratified random cross validation (SRCV) scheme about which we have given the details in the methods section. With this approach we can introduce variation in the training and validation sets in terms of both cell types and doses at the same time.

Firstly, we compare the schemes in which the training set includes different cell types. The most important observation regarding these three schemes is the low predictive power in the Sln1/WT scheme as can be seen in Figure 3.15a. This shows that when the models are trained without using the Sho1 data, validating them on Sho1 data is risky. On the contrary, when the Sln1 data is missing in the training set, we do not observe such low predictive power. The reasons for this can be traced back to the asymmetrical branch structure that we explained in detail in the Scenario 1 section. Therefore, we do not discuss those here again.

In addition to the risks associated with model validation, the Sln1/WT scheme performs poorly also in model selection with 16 wrong decisions. Therefore, we conclude that the Sln1/WT scheme is not a good scheme for model validation and selection whereas the Sln1/Sho1 and the Sho1/WT are sensible partitioning schemes. The SRCV scheme results in prediction errors that are comparable with the sensible partitioning schemes (Figure 3.15a). Furthermore, it results in no wrong decisions and it gives the highest model separation compared to the Sln1/Sho1 and Sho1/WT schemes which also result in all correct decisions (Figure 3.15b). In addition to this, the predictive power of such a scheme is less dependent on the noise realization compared to the other schemes as can be seen from the smaller box plots in Figure 3.15a. This indicates the low amount of uncertainty in the predictions.

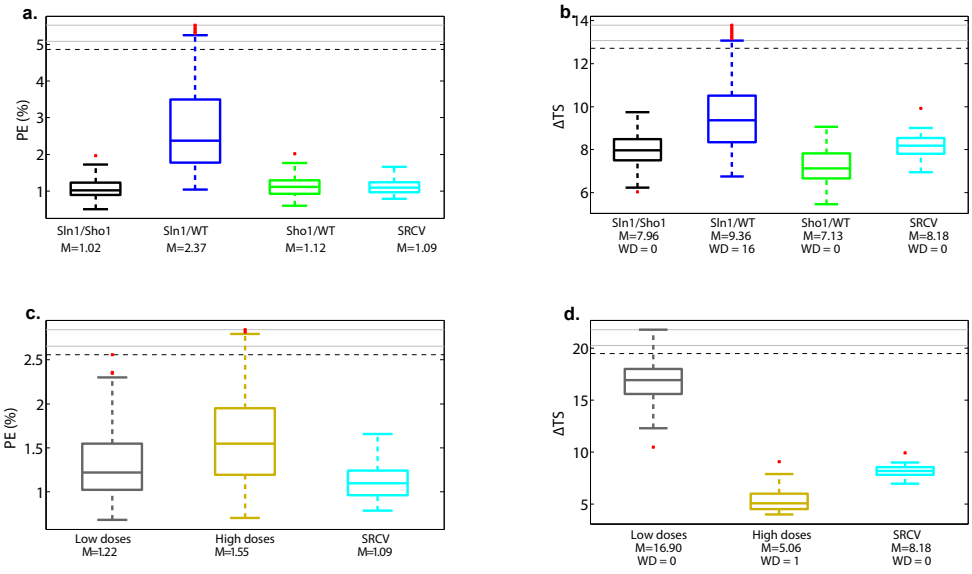

Figure 3.15: **Percentage prediction errors (PE) and model separation (ΔTS) in the adapted scenarios.** Each box plot shows the distribution of PE or ΔTS over 100 different realizations of the data obtained in a single scheme. The red dots indicate the outliers which lie outside approximately 99.3% coverage if the data is normally distributed. Black, blue and green boxes in the first row of graphs refer to the Sln1/Sho1, Sln1/WT and Sho1/WT schemes. Cyan boxes refer to the stratified cross validation (SRCV) schemes. Gray and yellow boxes in the second row refer to the low doses and high doses schemes. The labels on the x-axis indicate the medians of the PE or ΔTS distribution summarized visually by the box plots. The axis labels in the ΔTS graphs show also the number of wrong decisions given in each scheme. In each graph, ten realizations with the highest PE or ΔTS are located above the black dashed line. The region above this line is compressed for visual ease. **a.** PE obtained in adapted cell type scenario. **b.** ΔTS obtained in adapted cell type scenario. **c.** PE obtained in adapted dose scenario. **d.** ΔTS obtained in adapted dose scenario.

When only doses were allowed to vary in the training set as in the case of the low doses and the high doses scheme, there was no significant difference in the predictive powers of the two schemes (Figure 3.15c). This revealed that none of the schemes posed more risk of invalidating the true model structure compared to the other scheme. However, there was a large difference in the model separation achieved by the two schemes (median = 16.9 vs. 5.06 in the low and high doses schemes respectively, shown in Figure 3.15d). This shows that the high doses scheme is unsuitable for model selection. A simulation showing weak model separation according to the highest dose scheme can be seen in Figure S2. The SRCV scheme performed better than the unsuitable hold-out partitioning scheme for model selection (median ΔTS

= 8.18) and the predictive power was in the range of the two hold-out partitioning schemes (Figure 3.15c).

These results indicated that a stratified CV scheme is favorable for both model validation and selection. In most of the comparisons, it achieves predictive power and model separation as high as the optimal hold-out partitioning scheme. In addition, it leads to lower uncertainty which means that the outcomes of model validation and selection depend less on the specific noise realization. More importantly, it never performs worse than unsuitable hold-out partitioning schemes. The importance of this last statement lies in the fact that finding a sensible hold-out partitioning scheme can never be guaranteed. It depends highly on the biology and therefore, on the model structure and the model parameters most of which are typically unknown prior to modeling. Therefore, there are no rules that can be set beforehand to make the finding of sensible partitioning schemes certain. Those factors might hinder us from opting for a sensible scheme. However, SRCV offers a judicious and reliable partitioning scheme for which no biological knowledge is required. Its good performance relies on two properties. Firstly, it is iterative which means that it allows each piece of data to contribute as both training and validation datasets in an iterative manner and summarizes the results as the average of different iterations. Secondly it offers random stratified partitioning, so it allows fair partitioning of the data while it prevents from certain cell types or doses dominating the training data. Therefore, issues like parameter estimation and model validation/selection are not biased in a certain direction as an artifact of an underlying biological property of the system, in contrast to the hold-out validation schemes we have extensively investigated with this study.

3.4. Conclusions

Our results showed that the final decisions on model validation and selection can differ significantly when different hold-out partitioning schemes are employed. The selection of a sensible hold-out partitioning scheme that will help us to make reliable decisions depends on the biology. A good biological knowledge on the system and, hence, prior information on the structure and the true parameter values of the model are essential. Unfortunately, this is not possible in many instances. This turns the problem of finding a sensible partitioning scheme for model validation and selection into a Catch 22 problem. When the determination of a sensible partitioning scheme fails, we face the risk of invalidating true model structures or of failing to select the true model structure over the other alternatives. Examples of the first situation are very difficult to find in the literature, though, because, only successful validation

examples are usually presented, leading to a 'verification bias'. Furthermore, partitioning schemes that are sensible for model selection are not necessarily suitable for model validation. Datasets from very similar experimental conditions have only weak model selection capability whereas datasets from very diverse experimental conditions are not appropriate for model selection either due to high uncertainty in the predictions. However, using a proper cross validation approach such as stratified random cross validation can help us to overcome these problems while being independent of any prior biological knowledge.

With the SRCV approach, we can partition the data randomly into training and validation sets iteratively and arrive at consensus decisions by averaging over all different validation datasets. SRCV performs at least as well as sensible hold-out partitioning schemes for both model validation and selection. On top of that, this comes without the risk of opting for an incorrect partitioning scheme which would lead us to biased conclusions. Furthermore, the decisions given within a SRCV scheme are less affected by the specific realization of the experimental noise. Due to all these reasons that we mention, SRCV proves to be a judicious, unbiased and promising alternative to the hold-out validation strategy for the validation and selection of ODE based models.

Acknowledgements

This project was financed by the Netherlands Metabolomics Centre (NMC) which is a part of the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research. The authors thank Gooitzen Zwanenburg for reading the manuscript and Jacky L. Snoep for fruitful discussions.

Appendix

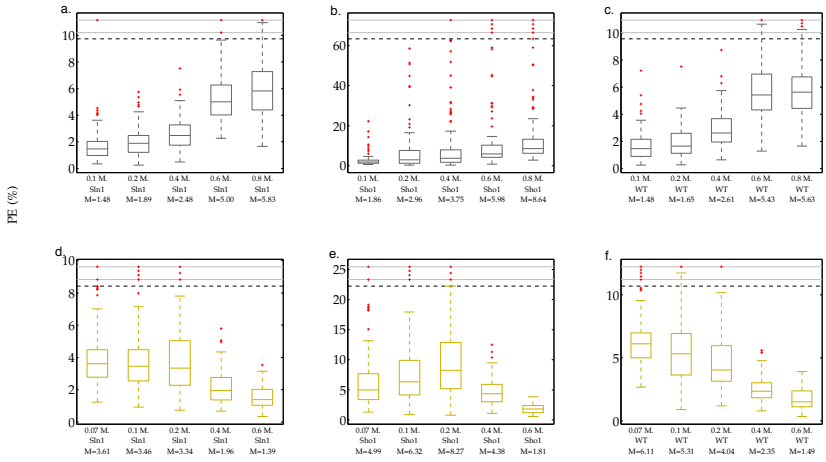

Figure S1. Percentage prediction errors (PE) of the simplified model structure in scenario 2. Each box plot shows the distribution of PE obtained the simplified model structure over 100 different realizations of the data. The red dots show the outliers which lie outside approximately 99.3% coverage if the data is normally distributed. They indicate realizations with relatively higher PE. Gray and yellow boxes refer to the lowest and the highest dose schemes. Each row in the figure corresponds to a single scheme. The labels on the x-axis show the specific dose and the cell type of the data on which the validation was performed. The labels indicate also the medians of the PE distribution summarized visually by the box plots. In each graph, the ten realizations with the highest PE are located above the black dashed line. The region above this line is compressed for visual ease. **a.** PE obtained on Sln1 validation subsets in the lowest dose scheme. **b.** PE obtained on Sho1 validation subsets in the lowest dose scheme. **c.** PE obtained on WT validation subsets in the lowest dose scheme. **d.** PE obtained on Sln1 validation subsets in the highest dose scheme. **e.** PE obtained on Sho1 validation subsets in the highest dose scheme. **f.** PE obtained on WT validation subsets in the highest dose scheme.

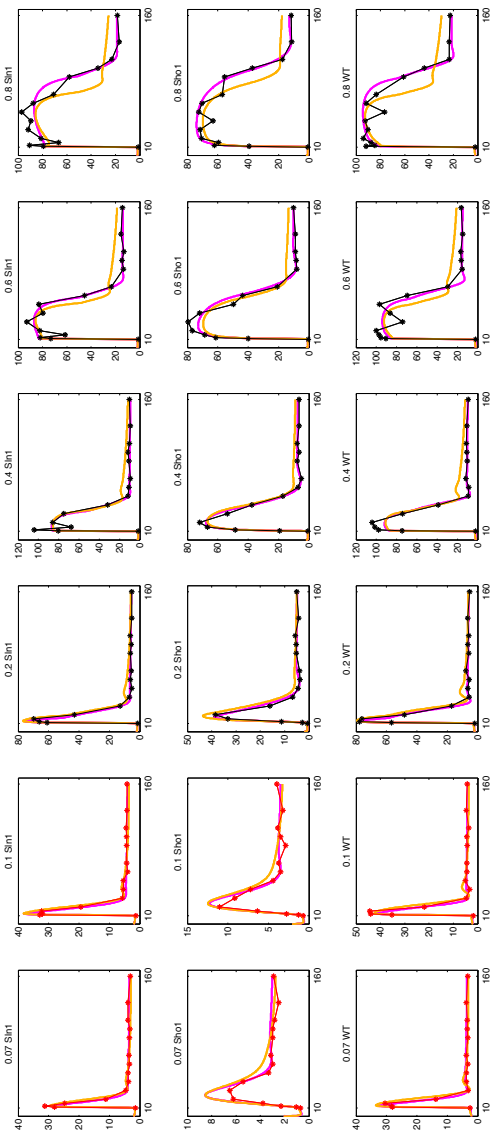

Figure S2. Fit and predictions obtained on a single realization of data in the high doses scheme. Black and red points refer to data points which were used for parameter estimation and validation, respectively. In this example simulation, data from doses between 0.2 M.-0.8 M. in each cell type and the data on the downstream species (protein, mRNA and internal glycerol) were used for parameter estimation. The magenta lines show the profiles obtained (both fit and prediction) by using the true model for the parameter estimation. The orange lines belong to the profiles obtained by the simplified model structure. All concentrations are given in percentages. Top three rows are for the Hog1PP data. The titles for each graph show the dose and the cell type related to the experiment in which the Hog1PP data was collected. On the other hand, the last row of graphs are for the downstream species and the associated data was collected in a single experiment with WT cell following 0.5 M. NaCl shock.

4

Selection and improvement of transcriptional regulatory networks

With the increasing amount and complexity of data generated in biological experiments it is becoming necessary to enhance the performance and applicability of existing statistical data analysis methods. This enhancement is needed for the hidden biological information to be better resolved and better interpreted. Towards that aim, systematic incorporation of prior information in biological data analysis has been a challenging problem for systems biology. Several methods have been proposed to integrate data from different levels of information most notably from metabolomics, transcriptomics and proteomics and thus enhance biological interpretation. However, in order not to be misled by the dominance of incorrect prior information in the analysis, being able to discriminate between competing prior information is required. In this study, we show that discrimination between topological information in competing transcriptional regulatory network models is possible solely based on experimental data. We use network topology dependent decomposition of synthetic gene expression data to introduce both local and global discriminating measures. The measures indicate how well the gene expression data can be explained under the constraints of the model network topology and how much each regulatory connection in the model refuses to be constrained. Application of the method to the cell cycle

regulatory network of Saccharomyces cerevisiae leads to the prediction of novel regulatory interactions, improving the information content of the hypothesized network model. [1]

4.1. Introduction

In recent years, multiplex and high-throughput technologies provided biologists with the opportunity to increase the amount of data generated on various biological systems. Analysis of these data allows to gain comprehensive information on the system on various levels such as transcriptome, proteome, metabolome and interactome. However, there are two major challenges which are directed by the systems biology perspective. The first challenge is to integrate all the information from these different levels. Statistical approaches for this aim have mostly stemmed from the need of integrating transcriptome data with other omics data sets. A noticeable example in this field has focused on mapping gene expression data on protein-protein and protein-DNA interactome data to reveal the active sub-networks in the course of perturbation experiments [68].

The second challenge is to interpret the massive information collected from experiments in a biologically meaningful manner. Data analysis can be directed towards knowledge already available on the investigated system to facilitate its biological interpretation. This approach is referred to as incorporation of prior information in data analysis.

Systematic incorporation of prior information in data analysis has been an important topic in statistics mostly due to Bayesian approaches. With the increasing demand of statistical approaches in biology, methods have been proposed also in this particular area. Several studies have focused on exploiting different kinds of prior information in biological systems. One important approach is based on Factor Analysis directed by prior information. Network Component Analysis (NCA) [99] set the framework for the decomposition of microarray data based on the transcriptional regulatory network topology provided as prior information. This decomposition leads to the reconstruction of both the connection strengths between gene - transcription factor pairs and the transcription factor activities over a range of different conditions. The NCA approach has been the subject of several followup studies which aimed either at increasing the applicability range of the method [50, 151], the stability of the solutions [151] or finding more efficient ways of carrying out the

[1] This chapter is based on:
D. Hasdemir, Gertien J. Smits, Johan A. Westerhuis, Age K. Smilde. Topology of transcriptional regulatory networks: testing and improving. *PLoS ONE*, 7(7): e40082, 2012.

decomposition involved [30].

In some studies, the prior information is exploited in a controlled manner where the analyst can set the limit for the intervention level of the prior information [127, 152, 166, 173]. In these studies, penalized or stepwise regression methods and Bayesian approaches are utilized. By some of these approaches, the prior information is also changed in accordance with the data at hand [152, 173]. In [173] this is accomplished by updating the prior information back and forth between different prior information with different reliabilities whereas in [152] forward stepwise regression is used to determine the true positive interactions in the prior information.

However, there is one point which must always be kept in mind while incorporating any type of prior information in data analysis. It is very likely to lead to incorrect results if incorrect prior information is allowed to dominate the data analysis process. Therefore, being able to discriminate between competing sources of prior information has always been an important issue. With this study, we propose measures to make this discrimination available on both global and local levels based on the assumption that correct models must behave consistent with experimental observations. To explain it more clearly; if there are two different hypotheses for a certain type of prior information, these measures will guide us to identify which hypothesis (on global level) or which parts of each hypothesis (on local level) are supported more by the experimental data and thus are closer to the underlying biological reality. Our focus is on topological prior information in transcriptional regulatory networks. However, such an approach can also be used for discriminating between competing prior information at other levels such as metabolomics and proteomics when appropriately adapted.

In this chapter, we show how two different regulatory networks can be distinguished on a global level using an NCA type decomposition framework. New regulatory interactions between genes and transcription factors can also be proposed by using our method. This feature represents our method's local performance. Furthermore, we show how application of our method to cell cycle transcriptional regulatory network of *Saccharomyces cerevisiae* led to the improvement of the regulatory interactions in the network.

4.2. Methods

4.2.1. Guideline for data decomposition

A Factor Analysis model for gene expression data can be written as in Equation 4.1. This type of model relates the gene expression data to the underlying hidden factors, namely the activity of the transcription factors. In this decomposition scheme, **X**

contains gene expression profiles of the I genes in the J conditions in terms of log_2 ratios. The score matrix **T** contains the binding association information between the I genes and the K transcription factors. The **P** matrix contains the activities of the K transcription factors in the J conditions in its columns. The matrix **E** contains the residual of the model, namely the part of the data that could not be modeled. Network Component Analysis (NCA) puts restrictions on the decomposition. In an NCA model, the score matrix **T** must be an element of Z, a special set of matrices. These matrices have a predefined structure based on the imposed topological pattern of the network. Binding of a transcription factor on the promoter region of a gene is represented with a nonzero value -the connection strength between the genes and the transcription factors- and lack of binding is represented with a 0. The decomposition in Equation 4.1 was proven to be unique up to scaling under certain criteria for the *identifiability* of the system [99]. The estimation of **T** and **P** under the imposed topological constraints gives us both the connection strengths between gene - transcription factor pairs as well as the transcription factor activities over a range of different conditions.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{4.1}$$

In our approach, the decomposition in Equation 4.1 is carried out by Alternating Least Squares with two types of constraints; topological constraints on **T** as NCA puts and unit column length constraint on **P**. In other words, as demanded by the first constraint, **T** has to stay a member of the set Z, the set of all the matrices which obeys the imposed topological pattern. The imposed topological pattern is represented by fixed places of zeros in **T**. With the second constraint, the length of all the columns in the estimated **P** (the activity profiles of all the transcription factors in the system) are fixed to unit length. This makes the comparison between different estimates of the **T** possible in different simulations as will be explained later.

The **first step** of Alternating Least Squares is the initialization step. In this step, $\mathbf{T_{initial}}$, an initial educated guess for **T** is given. For obtaining this initial guess, a PCA decomposition is carried out on the data matrix **X**. The resulting score matrix, $\mathbf{T_{PCA}}$ is a good initialization for **T** itself. However, the PCA score matrix can be rotated further towards the imposed topological pattern with the requirement of staying within the PCA space. So, the elements which are restricted to 0 based on the topology would be as close to 0 as possible and thus the nonzero elements would be adjusted accordingly. This is achieved by multiplying the PCA score matrix, $\mathbf{T_{PCA}}$ by a nonsingular rotating matrix, **R**. The minimization function for the estimation of **R** is given in Equation 4.2. The target minimization in Equation 4.2

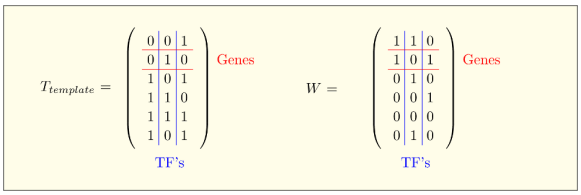

Figure 4.1: **Example network with 6 genes and 3 transcription factors (TF's).** The imposed network structure is encoded in $T_{template}$. Existing connections are depicted as 1's. W has 1's at the positions where the difference between the initial estimate, $t_{initial}(i,k)$ and the template, $t_{template}(i,k)$ is subject to minimization. These positions correspond to the 0's in $T_{template}$ where no connections exist.

is carried out only on the restricted elements as imposed by the binary matrix **W** and the use of the Hadamard Product (Figure 4.1). The $\mathbf{T_{template}}$ is the matrix in which the imposed topology is encoded with 1's showing the interaction and 0's showing lack of interaction between genes and transcription factors.

$$\begin{aligned} &\min_{R} || \ \mathbf{W} \circ (\mathbf{T_{template}} - \mathbf{T_{initial}}) \ ||_2 \\ &where \quad \mathbf{T_{template}} \in \mathrm{Z} \quad \text{and} \quad \mathbf{T_{initial}} = \mathbf{T_{PCA}}.\mathbf{R} \end{aligned} \tag{4.2}$$

In the **second step**, estimation of **P** is achieved by an Iterative Restricted Least Squares approach (Personal communication with Henk Kiers, University of Groningen) using the educated guess $\mathbf{T_{initial}}$. In the **third step**, Ordinary Least Squares is used for estimating **T** based on $\widehat{\mathbf{P}}$ which was estimated in Step 2. In this step only the nonzero values in **T** are subject to change. The elements which are restricted to 0 as imposed by the network topology are always kept as 0. The computation in this step follows the guideline which was defined within the NCA framework [99]. The constraints and objectives of the overall optimization scheme are summarized in Equation 4.3 where the estimated variables are shown with a hat ($\widehat{\ }$) on them. The **fourth step** is the termination step where the alternating least squares algorithm is terminated when the relative change in the residuals is below a previously determined threshold.

How this type of supervised decomposition of gene expression data is used to provide us with a guideline to discriminate between competing network information

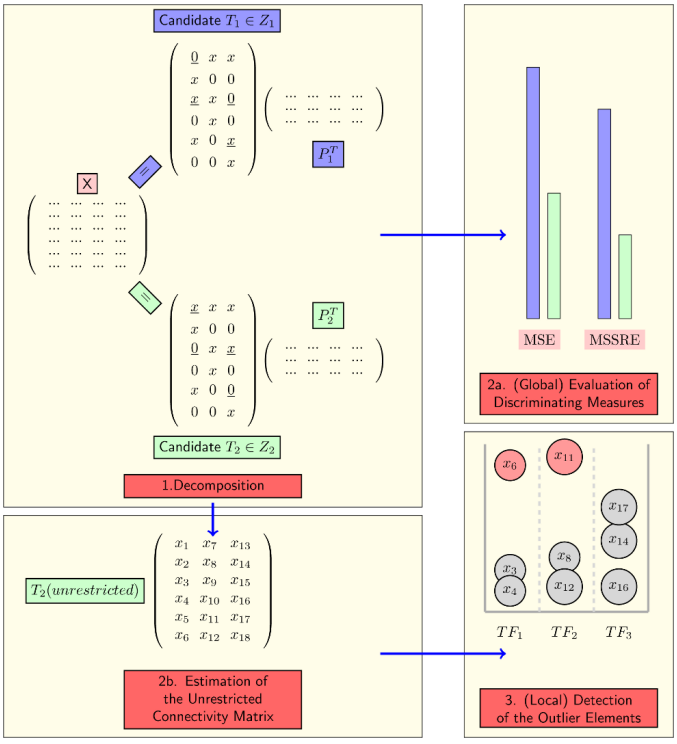

Figure 4.2: **General view of the approach.** The individual steps in the figure are explained in detail in the Methods Section.

is depicted in Figure 4.2.

$$\begin{aligned} &\underset{\widehat{\mathbf{T}},\widehat{\mathbf{P}}}{min} \; || \, \mathbf{X} - \widehat{\mathbf{T}}\widehat{\mathbf{P}}^{\mathbf{T}} \, ||_2 \\ &s.t \quad \widehat{\mathbf{T}} \in \mathrm{Z} \\ &s.t \quad \widehat{\mathbf{p_k}}^{\mathbf{T}}\widehat{\mathbf{p_k}} = 1 \quad \forall \; \mathrm{k} \in 1{,}2{,}...{,}\mathrm{K} \\ &where \quad \mathrm{K} = \#\text{ (Transcription Factors)} \end{aligned} \tag{4.3}$$

4.2.2. Discriminating measures

MSE: Mean sum of squared residuals in the model

The first proposed measure is the model fit. This simple but yet very important measure is strongly dependent on the prior information regarding the network structure. The concept of model fit as a discriminating measure is based on the idea that the model data matrix, $\hat{\mathbf{X}}$ will be closer to the measured data matrix, $\mathbf{X}$ when a

network model which is closer to the real network structure is used as prior information. This is due to the strictness of the constraints imposed by the topology of the network.

MSE is calculated via Equation 4.4 as the mean sum of squared residuals in the model.

$$\widehat{\mathbf{X}} = \widehat{\mathbf{T}}\widehat{\mathbf{P}}^{\mathbf{T}}$$
$$MSE = (\sum_{i}^{I}\sum_{j}^{J}(x_{ij} - \widehat{x}_{ij})^2)/(I \times J) \tag{4.4}$$
$$where \quad \text{I} = \#\ \text{(Genes) and J} = \#\ \text{(Conditions)}$$

MSSRE: Mean squared sum of restricted elements

The second proposed measure uses $\hat{\mathbf{T}}_{\mathbf{un}}$ which is the connectivity matrix estimated after the relaxation of the topological constraints. MSSRE summarizes the distance of specific elements from 0 in $\hat{\mathbf{T}}_{\mathbf{un}}$. These specific elements are the ones which were restricted to 0 in the imposed network structure.

Once the alternating least squares is terminated and the $\hat{\mathbf{P}}$ matrix is estimated, we can obtain the unrestricted connectivity matrix, $\hat{\mathbf{T}}_{\mathbf{un}}$ by solving Equation 4.5. This would be the ordinary least squares solution to the problem under no topological constraints. So, there are now two predicted connectivity matrices; restricted ($\hat{\mathbf{T}}$) and unrestricted ($\hat{\mathbf{T}}_{\mathbf{un}}$). The unrestricted connectivity matrix has been relaxed from any type of topological constraints. In principle, if the network model that was supplied as prior information is indeed close to the real network structure, the elements which were previously restricted to 0 should not deviate far from 0 in the unrestricted connectivity matrix. This information can be accessed via the mean squared sum of these elements, referred to as MSSRE (Mean Squared Sum of Restricted Elements). A similar approach was also used in introducing the Core Consistency Diagnostic (Corcondia) in 3-way analysis [22].

The idea of MSSRE intrinsically assumes that $\hat{\mathbf{P}}$ has been properly estimated. This measure would not be appropriate if $\hat{\mathbf{P}}$ cannot be estimated accurately. A major reason for probable inaccuracy in the estimation of $\hat{\mathbf{P}}$ within the NCA framework and how it was challenged will be discussed in more detail later while showing the

application of the method on a real biological system.

$$\begin{aligned}
&\widehat{\mathbf{T}}_{\mathbf{un}} = \mathbf{X}\widehat{\mathbf{P}}(\widehat{\mathbf{P}}^{\mathbf{T}}\widehat{\mathbf{P}})^{-\mathbf{1}} \\
&MSSRE = (\sum_{i}^{I}\sum_{k}^{K}\widehat{t}_{un}(i,k)^2)/n_r \quad \forall \text{ i,k} \qquad \text{for which} \quad \widehat{t}_{i,k} = 0 \\
&where \quad \widehat{\mathbf{T}} \in \mathrm{Z}, \quad \widehat{\mathbf{T}}_{\mathbf{un}} \notin \mathrm{Z},
\end{aligned} \tag{4.5}$$

I = # (Genes), K = # (Transcription Factors) and n_r = # elements restricted to 0 in $\widehat{\mathbf{T}}$

4

The unrestricted connectivity matrix gives the opportunity to evaluate the proposed network model from a local point of view as well. Investigation of the individual elements in the unrestricted matrix makes it possible to see which connections in the network model are supported by the gene expression data. Some of the elements which were restricted to 0 in the originally imposed network structure would tend to deviate far from 0 in the unrestricted connectivity matrix more than the others. These elements would indicate additional potential connections in the network. Furthermore, the idea of unrestricted connectivity matrix can be extended to include a new set of genes whose expression profiles were not used for the estimation of $\hat{\mathbf{P}}$.

In addition to the original model set with the I genes in the analysis, the estimated $\hat{\mathbf{P}}$ can be used to estimate the connection strengths of the transcription factors with a new the set of L genes which were not previously included in the analysis, $\hat{\mathbf{T}}_{\mathbf{un}}(\mathbf{new})$ as in Equation 4.6. For this purpose, the gene expression data of these new genes in the J conditions ($\mathbf{X}_{\mathbf{new}}$) is used. This extension of the approach assumes that $\hat{\mathbf{P}}$ could be properly estimated by using only the expression profiles of the model set genes.

$$\hat{\mathbf{T}}_{\mathbf{un}}(\mathbf{new}) = \mathbf{X}_{\mathbf{new}}\hat{\mathbf{P}}(\hat{\mathbf{P}}^{\mathbf{T}}\hat{\mathbf{P}})^{-\mathbf{1}} \tag{4.6}$$

4.2.3. Simulations setup

The main goal of the simulations study was to model the simulated data by embedding different types of prior information during modeling and to elaborate on the measures that made it possible to discriminate between these different cases. In this sense, synthetic data gives us the opportunity to know exactly which network model is closer to the real network structure, $\mathbf{T}_{\mathbf{true}}$ which was used to generate the data. The simulated dataset consisted of 240 genes, 20 transcription factors, 40 different

conditions and was constructed based on Equation 4.7.

$$\begin{aligned} \mathbf{X_{true}} &= \mathbf{T_{true}}\mathbf{P_{true}^{T}} \\ \mathbf{X} &= \mathbf{X_{true}} + \mathbf{N} \end{aligned} \tag{4.7}$$

The values of the elements in the $\mathbf{T_{true}}$, $\mathbf{P_{true}}$ and the measurement noise term, $\mathbf{N}$ were randomly drawn from standard normal distribution. Then some of the elements in $\mathbf{T_{true}}$ were randomly set to 0 representing the imposed topological pattern. This pattern was very sparse where one gene is regulated by at most 6 transcription factors in order to mimic the sparsity of real biological transcriptional regulatory networks. The level of the added measurement noise was either 0% , 5% or 20% . The noise level was calculated based on the sum of squares of the true expression data, $\mathbf{X_{true}}$ as shown in Equation 4.8.

$$noise = \frac{\sum_i^I \sum_j^J n(i,j)^2}{\sum_i^I \sum_j^J x_{true}(i,j)^2} \tag{4.8}$$

Table 4.1: **Prior Information Properties**

Case Label	Type of Prior Information	Remarks
CN	Correct Network	-
MN3%	3% Randomly Misconnected Network	Misconnection level 1 in Figure 4.3
MN5%	5% Misconnected Network	Misconnection level 2 in Figure 4.3 -Extra misconnections were added on top of the corresponding readily miswired network structures of case MN3%.-
RN	Random Network	Certain graph properties were kept the same with case CN.

The simulations were carried out by embedding both correct and incorrect prior information at different noise levels in various transcriptional regulatory network structures. Here, correct prior information (Correct Network (CN) in Table 4.1) refers to the true regulatory pattern used to create the synthetic network structure as explained earlier. For the incorrect prior information (MN3% and MN5% in Table 4.1), the connections of randomly selected transcription factors were changed by adding a connection not present and removing another connection which was present in the correct prior information. The properties of the various prior information used in the simulations are shown in Table 4.1. As shown in Table 4.1, in a 3% randomly misconnected network structure, 3% of the total number of the

connections differ with respect to the correct prior information. We derived 5% randomly misconnected network structures from the 3% misconnected structures by adding extra random misconnections. The misconnection levels were kept low in order to test the discrimination capability of our method even at very low levels. Furthermore, random network structures were generated completely unrelated to the correct one except the size, connectivity and total number of connections in the network. We used random networks to see the changes in the discriminating measures when the prior information is completely incorrect and thus to test whether the discriminating measures relied on chance.

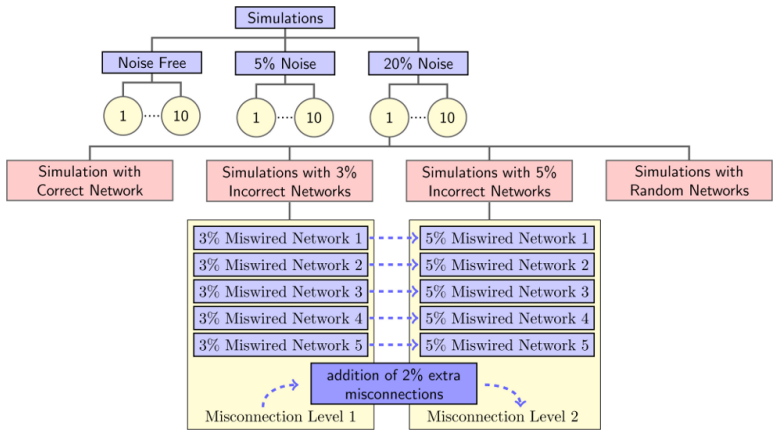

Figure 4.3: **Structure of the Simulations.** At each noise level, 10 different regulatory networks were simulated as shown with circles in the figure. Later, each of these were manipulated to obtain different prior information as depicted in pink rectangles.

The structure of the simulations setup is depicted in Figure 4.3. For each specific regulatory network structure, there existed simulations with four types of prior information (see Table 4.1 for details). The incorrect prior information cases consisted of 5 different sets of misconnections at each misconnection level as shown in Figure 4.3 with the yellow boxes. On the other hand, simulations with random prior information were carried out 100 times for each corresponding CN. Most important of all, each of the nodes in this simulation scheme was repeated with 100 different noise realizations. This means that the simulation experiments were repeated 100 times at each case of prior information. This allowed statistical comparison between different cases.

4.3. Results and discussion

4.3.1. MSE: A sensitive global measure for discrimination

In Figure 4.4, the medians and median absolute deviations of MSE are shown for 12 different simulation cases in each of the three example network structures. As can be seen from the figure, the MSE discriminates between different types of prior information steadily well even with 20% noise in the data which was the maximum level of noise in the simulations. We have chosen this noise level based on the expected level of reproducibility in different types of microarray data. In a comprehensive study where they calculated the coefficient of variation of gene expression in replicate experiments, the median of this variation coefficient across all genes changed between 5% and 23% [135]. This indicated that the noise to signal ratio never exceeded 23% in over 80 experiments that they have performed with 6 different platforms. Our limit of 20% thus seems realistic for microarray data in general. The results of simulations with less noise were more apparent so they were not discussed here.

In Figure 4.4, each shape in blue (MN3%) has an MSE distribution with a higher median than the correct network (CN) model and has an MSE distribution with a lower median than the corresponding same shape in green (MN5%). Besides that, all of these MSE distributions depicted in the main parts of the graphs locate separately from the MSE distributions of totally random network structures (RN) shown in the upper right corners. This shows that when misconnections are introduced in the prior information of network structure, the model fit gets worse. One sided t-tests between these cases with different prior information also indicate that all of these distributions can be separated statistically well from each other at a 5% significance level. This means that the mean of the MSE distribution of the 100 experiments with a MN5% is greater than the mean of both a CN and a MN3% .

What is important here is that the solutions in 100 different noise realizations are very close to each other in each case with a certain prior information. The superiority of using the alternating least squares approach not with random guesses but with an educated initial guess for $\mathbf{T}$ is important, in this sense because the ordinary least squares optimization with random initial guesses leads to local minima in a considerable number of cases. However, by using an educated initial guess, the local minima problem was encountered only in 0.5% of all the simulations that have been carried out in total. This finding eliminated the need for additional runs with different starting points.

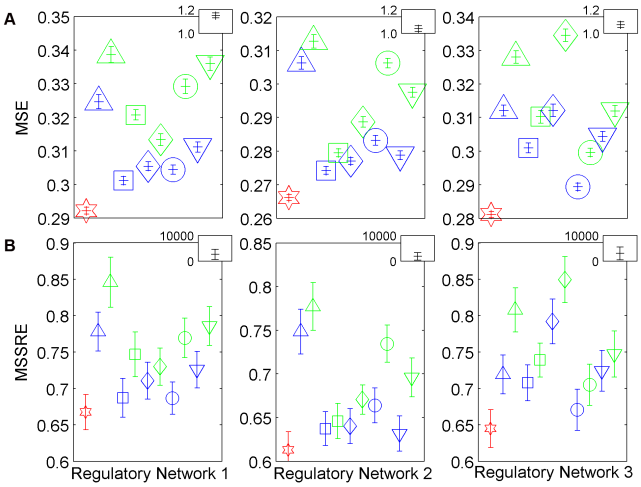

Figure 4.4: **MSE and MSSRE in the simulations with 20% noise.** Each point corresponds to the median of the MSE values (**Panel A**) and MSSRE values (**Panel B**) in 100 simulations with different noise realizations. The error bars represent the median absolute deviation. The results of the simulations with the correct network (CN) are plotted in red whereas the blue and green colors represent the solutions with 3% and 5% misconnected network structures, respectively (MN3% and MN5%).The networks which are represented by the green shapes share the same misconnections with the corresponding shapes in blue and have extra misconnections on top of these. In the upper right hand side corner of each plot, the results of the simulations with totally random networks (RN) are depicted. In **Panel A**, each error bar is surrounded with different shapes, whereas in **Panel B**, the medians are denoted by the corresponding shapes for better readability of both graphs. (See Table 4.1 for the details of the prior information used).

4.3.2. MSSRE and the local investigation of the unrestricted network structure

In accordance with the observations in MSE, the relaxation of topological constraints for calculating the MSSRE led to relatively higher MSSRE in simulations with incorrect prior information (Figure 4.4). MSSRE acted in a consistent manner with the previously discussed MSE measure. The results of non-parametric tests between these different prior information cases (Figure 4.4) indicated that the distributions of MSSRE could be separated in 95% of all the comparisons at a 5% significance level. The results suggested that the variability of MSSRE was higher than the MSE. Even when the MSE of different models were small, the values that the originally restricted elements took in $\hat{\mathbf{T}}_{\mathbf{un}}$ could vary to a higher degree. However, the unrestricted connectivity matrix $\hat{\mathbf{T}}_{\mathbf{un}}$ offered more. Inspection of the individual elements in the unrestricted matrix gave indications on the locations of the misconnections. This

generates the opportunity for a local evaluation and possible improvement of the proposed network structure.

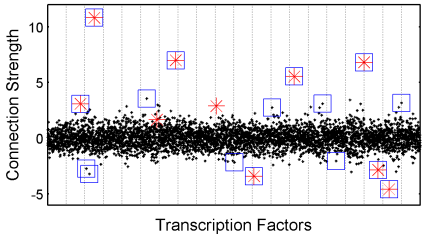

Figure 4.5: **Values of Connection Strengths in $\hat{T}_{un}$.** In this plot, each black point represents the connection strength of an element in the unrestricted connectivity matrix which was restricted to 0 previously in the imposed network structure. Each column represents one transcription factor. Red stars are the connections which were missing in the imposed network structure whereas in CN these elements have nonzero values indicating existing connections instead. The outlier elements for each transcription factor identified at a whisker length of 2 are surrounded with additional squares in blue.

In Figure 4.5, the values of specific elements in the unrestricted connectivity matrix, $\hat{\mathbf{T}}_{\mathbf{un}}$ are shown. These specific elements are the ones which were originally restricted to 0 in the proposed network structure. This example figure comes from one of the simulations with MN5% with 20 % noise. The connection strengths of these originally restricted elements estimated after the relaxation of the topological constraints are shown with black dots. In some of the elements, the deviation from 0 is very strong and easily distinguishable from the others. These are the outlier connections framed in blue squares.

The outlier connections are the ones that were estimated to be uncommonly strong compared to other connections. Their values were estimated either higher than $q_3 + w * (q_3 - q_1)$ or smaller than $q_1 - w * (q_3 - q_1)$, where q_1 and q_3 are the 25^{th} and 75^{th} percentiles of the distribution, respectively and w is the whisker length as suggested by [49]. This distribution refers to the distribution of the elements which were originally restricted to 0, in one column of $\hat{\mathbf{T}}_{\mathbf{un}}$ (corresponds to one transcription factor at each time). As a result of their extra-ordinary locations in the very edge of the distributions, they have the potential to point to the existing connections which were missing in the imposed network structure. Therefore, we expect the outlier connections and these missing connections (Figure 4.5) to overlap at a considerable degree. The non-parametric definition of outliers is very beneficial for our case where the underlying distributions of the connection strengths per transcription factor might not be normal. The sensitivity and false discovery rate

Table 4.2: **Discrimination Performance**

Whisker Length	Sensitivity	False Discovery Rate
	# Missing connections identified in the outliers/ # All missing connections	# False positives in the outliers / # All outliers
1.20	0.7228	0.9487
1.40	0.6897	0.9080
1.60	0.6593	0.8342
1.80	0.6274	0.7114
2.00	0.5980	0.5409
2.20	0.5672	0.3642
2.25	0.5598	0.3231
2.30	0.5528	0.2846
2.40	0.5382	0.2182
2.45	0.5309	0.1885
2.50	0.5237	0.1641
2.60	0.5099	0.1218
2.70	0.4965	0.0909
2.80	0.4830	0.0690

for the identification of the missing connections depend on the chosen whisker length (w) as summarized in Table 4.2. In Table 4.2, whisker length has been varied from the very loose value, 1.2 up to the extremely strict value of 2.8. When the number of outliers were kept high at a whisker length of 1.2, 72% of the missing connections (denoted by red stars in Figure 4.5) were identified in the outliers. This indicated the sensitivity of the method. As the whisker length increased, the number of the missing connections which could be identified by the method (true positives) decreased as a result of the decreasing number of outliers detected. This decreased the sensitivity from 0.72 to 0.48. On the other hand, the false discovery rate decreased at a faster rate from 0.95 to even 0.07. When whisker length was set to the most extreme value, only 7% of the outlier connections were false positives. Although it depends on the analyst to decide which one to sacrifice, in most of the cases, we think that the number of false positives should be reduced as much as possible. This performance summary depicted in Figure 4.2 proves to be a very useful tool when the whisker length has to be optimized for real biological data. The FDR value calculated for simulated data can give good indication of the expected FDR in real data.

4.3.3. Overall results of the simulations study

For further investigation of the discriminating capacity of the measures, we checked the magnitude of the connection strengths differing between the competing networks.

Indeed, the answer to the question whether these networks are easily distinguishable heavily depended on the magnitude of the difference between the competing networks. If the connections which the imposed network (either MN3% or MN5%) lacked were indeed strong connections in the correct network (CN), the differences in both measures with respect to the simulations with CN were shown to be larger. This relation is more clear when Figure 4.6 is investigated. In these figures, the relative values of the two measures in MN3% and MN5% simulations with respect to their values in CN simulations were shown with respect to the magnitude of the misconnections. The magnitude of the misconnections is formulated as the sum of squares (SS) of all the connection strengths that were existent but later had been restricted to 0 to create the misconnected networks. The differences increased as the sum of squares of the misconnections increased and thus made the discrimination as clear as it deserves. This conclusion is based on the idea that strong connections deserve more to be identified than the weaker interactions. However, Figure 4.6 suggests that this dependence on magnitude is weaker in MSSRE. In some cases, the difference in MSSRE was smaller when SS of the misconnections was larger. This can be explained by the dependence of the MSSRE on the topology. Discrimination by MSSRE might be more difficult when certain topologies are involved.

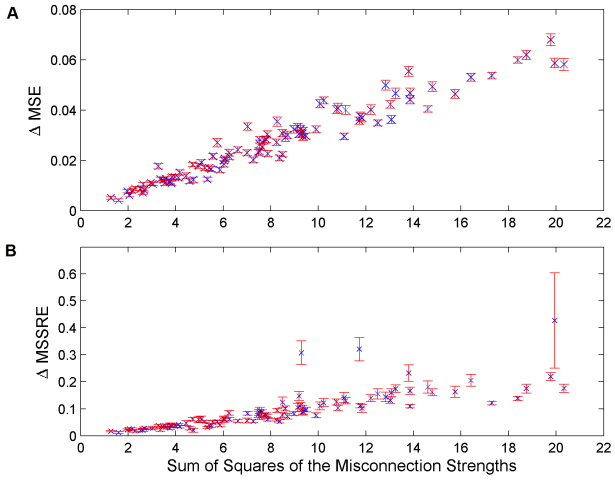

Figure 4.6: **Behavior of MSE and MSSRE with respect to changes in the connection strengths differing between the competing networks.** The change in MSE (**Panel A**) and MSSRE (**Panel B**) in simulations with MN3% and MN5% relative to the simulations with CN are shown in the y-axis. In the x-axis, the total magnitude of the missing connections which existed in CN but had been ignored in the imposed network structure are plotted.

Based on the results we obtained from the simulations study, it can be concluded that MSE and MSSRE both can be used efficiently to discriminate between two competing network structures. The discrimination was possible in 95% of cases with MSSRE and 100% of cases with MSE even when the network structures shared 97% of the connections in common. However, the discrimination capabilities of both measures increased with the strength of the connections differing between the two networks. In this sense, MSE was more successful than MSSRE in discriminating weaker differences. MSSRE seemed to be more dependent on the specific topologies of the networks that were questioned. On the other hand, the unrestricted network structure was worth to be inspected in more detail as a tool for local evaluation rather than global evaluation. When the connection strengths in this unrestricted connectivity matrix were investigated, the unexpectedly outlying elements which deviate far from 0 proved to be mostly the connections which exist in reality but had been ignored in the previously imposed network structure. However this last conclusion must always be carried out with care since the setting of the whisker length for the definition of an outlier heavily affects the sensitivity and false discovery rate in the discrimination process. The values reported for FDR at different whisker lengths in the simulations study can work as a very useful reference for application of the method to real biological systems.

An important point of discussion regarding our method might be the questioning of the applicability range of our method's local improvement feature. How misconnected can a network be at most to still allow this method to indicate potential connections in the network, and what is the maximum noise level allowing reliable analysis? The actual misconnection level in TF-binding data is thought to be between 10% and 50% [104]. Therefore, we constructed even more misconnected networks (25% misconnection level) to test this particular feature with 30%-50% measurement noise. The sensitivity values were affected by both noise in the gene expression measurements and misconnection level of the network. The sensitivity calculated at the whisker length of 1.2 decreased to 0.54 in the most extreme case with 50% noise and 25% misconnection level. This decrease in the sensitivity at extreme cases indicated that identification of the misconnections became difficult. However, the FDR values were not affected at all by the increasing noise or the misconnection level. The stability of the FDR values calculated at the whisker length of 2.8 (below 0.07) showed that even at these extreme cases, the candidate interactions identified by our approach were very unlikely to include false positives. Keeping the FDR low in such a discovery scheme makes more sense from a biological point of view than achieving high sensitivity. Hence, these results indicated that our local approach is reliable even at these high experimental uncertainties. When the global

discrimination was considered, MSE could discriminate between these extreme cases in 100% of the comparisons. The performance of MSSRE in discriminating networks with small differences decreased with increasing noise. However, it could still discriminate in 80% of all the comparisons made between highly similar networks with 95% connections in common. Another important point we observed was that an increase in the misconnection level of the network resulted in an increase in the number of local minima encountered in the simulations. This indicates the need for an optimization scheme with multiple starting points when prior information on the approximate extent of the misconnection levels of the networks is not available. The results of simulations at high experimental uncertainties can be found in Table S1.

4.3.4. Application to the cell cycle trancriptional regulatory network of yeast

We applied our discrimination algorithm to a real biological transcriptional regulatory network structure. For this purpose, we chose to work on the transcriptional regulatory system controlling the well studied cell cycle in *S. cerevisiae*. The system chosen included 11 cell cycle transcription factors: Ace2p, Fkh1p, Fkh2p, Mbp1p, Mcm1p, Skn7p, Ndd1p, Stb1p, Swi4p, Swi5p and Swi6p. For the identifiability of the network, Ndd1p had to be removed [99]. The network structure regarding these transcription factors was adapted from the study of Harbison *et al.* [59], named the Harbison network throughout the text. In this benchmark study, the genes that are likely to be targets for transcriptional regulators have been identified by consensus of information from genome-wide location data, phylogenetically conserved sequences and prior information. For our purposes, we used the most reliable transcription factor-gene interactions with a binding p-value smaller than 10^{-3} that have been conserved in at least two yeast species.

Cell cycle microarray data from Spellman *et al.* was used for the analysis [142]. The data analyzed consisted of time series gene expression data from four synchronization experiments, leading to a total of 77 sampled conditions.

Problem of degeneracy

A degeneracy problem arose when the expression data was modeled with 10 underlying factors. This led to an extremely high condition number of the estimated activity matrix which indicated that the activities of different transcription factors were linearly dependent on each other. However, for proper discrimination between networks independent profiles are required, because otherwise the activity profiles and thus the connections of the transcription factors cannot be distinguished. It is very important to notice that, in such situations, one of the criteria for identifiability

[99] is severely disturbed. This loss in the rank of the activity matrix might be easily missed due to the compensation by the noise in the system. In other words, noise in the data may hide the elevated levels of linear dependence between the activity profiles of the transcription factors. In the end of the analysis, estimated profiles of the transcription factors might be extremely correlated although there is no loss in the calculated rank of the activity profile matrix.

When more factors are extracted from a data set than can be supported by it, this kind of degeneracy occurs [174]. The SVD decomposition of the data also indicated clearly that the data should be modeled with fewer underlying factors. A scree plot of the singular values revealed an optimal number of 7 independent factors. This high dependency between the activity profiles can be explained by the partial redundancy and serial regulation structure regarding different cell cycle regulators. Indeed, in earlier studies it was shown that cell cycle regulators had important roles in controlling each other's expression profiles[138]. A high degree of overlap between the target genes of cell cycle regulators was also mentioned in the same study. These regulators were not only homologues or partners in regulating complexes, but they could also be regulators that were not known to be related at all in terms of their specific role in regulation. These findings support our estimated dependency between the factors.

To find out the best combination of 7 transcription factors, a trade-off approach was used. Out of all the possible combinations, a set of factors was selected that were most independent but yet resulted in low residuals and thus good models of the data. For this aim, decompositions were carried out several times with all possible combinations of 7 transcription factors. Among the models with the best fit, we looked for the smallest condition number of $\hat{\mathbf{P}}$. This set of transcription factors both described the data well, indicated by the low residuals of the model, and were independent of each other, indicated by the small condition number of $\hat{\mathbf{P}}$. The resulting high confidence network contained the interactions between 342 genes and the 7 transcription factors: Ace2p, Fkh1p, Mbp1p, Mcm1p, Skn7p, Swi5p and Swi6p.

Another general solution to this degeneracy problem would be increasing the number of experimental conditions. At these new data points, the connectivity structure of the network should be the same but the biological interdependency of transcription factors should be lower. That would allow independent activity profiles in $\hat{\mathbf{P}}$ but it also necessarily requires design of new experiments and specific selection of data points. Indeed in [151], the authors have followed a specific application of such an approach where they incorporated microarray data from transcription factor deletion mutants. They achieved this incorporation by putting constraints on

$\hat{\mathbf{P}}$ such as zeros for certain elements. As a conclusion, carrying out new microarray experiments might solve the problem of degeneracy while a purely computational solution remains as a challenge.

Two competing networks for cell cycle regulation

We compared the Harbison network [59] to another network that has been constructed by reanalyzing the same ChIP data [104] by MacIsaac *et al.*. This alternative network has been reported as an improved map of the regulatory network in *S. cerevisiae*. We only included the genes and the transcription factors that have reported interactions in both networks. When the number of transcription factors was further reduced to overcome the non-identifiability and degeneracy problems, both networks included 308 genes and 7 transcription factors. The MSE values for the two networks did not differ significantly (0.1312 and 0.1313, respectively). This suggests that the cell cycle related part of the MacIsaac regulatory network used for this study does not show significant improvement in comparison to the Harbison network. It is important to note here that the size of the networks that can be compared is severely limited by the limitations of the NCA approach. First of all, the networks both must be identifiable, as has been discussed by Liao *et al.* in [99]. Secondly, the transcription factors involved in the study must have independent activity profiles as we have already discussed. Due to these restrictions, we could only compare certain parts of the cell cycle regulatory network. Still, we showed that this part represents the whole cell cycle regulatory mechanism well. The details regarding this latter assessment was already discussed in the previous section where we discussed about the best combination of the transcription factors. Another reason behind the insignificant improvement might be due to the connection strengths of the connections differing between the two networks. As we have already discussed in the results section for the simulations, networks with strong connections differing are more easily discriminated than the ones with weaker connections differing. It might be the case that the regulatory interaction map achived in [104] is indeed more realistic but these interactions that exist in the part of the network we tested are not strong enough to be identified. However, local investigation of the Harbison network showed possible points of improvement in the network as will be discussed in the next section.

Emerging interactions

The unrestricted Harbison cell cycle transcriptional regulatory network, $\hat{\mathbf{T}}_{\mathbf{un}}$ was calculated according to Equation 4.9 with an extension to a new set of 54 genes. The idea behind was introduced in Equation 4.6. The expression values of this new

set of genes, $\mathbf{X_{NewSet}}$ were not taken into account for estimating $\hat{\mathbf{P}}$ but were used for estimating their connection strengths with the transcription factors in the study as described before for simulated data. The genes in the new set were known to be cell cycle regulated [142] but had not been included in our network. The reason was that these genes were not connected to any of the transcription factors based on the interaction data adapted from [59] and thus were previously excluded from the analysis. $\hat{\mathbf{T}}_{\mathbf{un(NewSet)}}$ stored the unrestricted connection strengths of the 54 new set genes whereas $\hat{\mathbf{T}}_{\mathbf{un(ModelSet)}}$ stored the unrestricted connection strengths of the 342 model set genes whose expression values were used to estimate $\hat{\mathbf{P}}$.

Equation 4.9 also shows clearly how the degeneracy problem would lead to difficulties associated with the identification of the new interactions in both the model set and the new set. High condition number of $\hat{\mathbf{P}}$ would make it nearly rank deficient. This would introduce errors in the generalized inverse of $\hat{\mathbf{P}}$ used in Equation 4.9 and thus also in the $\hat{\mathbf{T}}_{\mathbf{un(NewSet)}}$ and $\hat{\mathbf{T}}_{\mathbf{un(ModelSet)}}$.

$$\begin{bmatrix} \hat{\mathbf{T}}_{\mathbf{un(ModelSet)}} \\ \hat{\mathbf{T}}_{\mathbf{un(NewSet)}} \end{bmatrix} = \begin{bmatrix} \mathbf{X_{ModelSet}} \\ \mathbf{X_{NewSet}} \end{bmatrix} \hat{\mathbf{P}}(\hat{\mathbf{P}}^{\mathbf{T}}\hat{\mathbf{P}})^{-\mathbf{1}} \tag{4.9}$$

Figure 4.7 shows the connection strengths estimated in $\hat{\mathbf{T}}_{\mathbf{un}}$. For each transcription factor, outlier connections were identified as described earlier for simulated data. The elements surrounded with squares are of high importance because they potentially point to connections which indeed do exist but had not been included in the network structure used. In the same figure, the boxes in each segmentation show the gene names for these potential interactions of each transcription factor.

When the whisker length was kept at 2.8, only 6 outliers were detected. Based on the performance evaluation in Table 4.2 we expect nearly 0 false positives in this set. It must not be forgotten that the real biological data will show differences in terms of sensitivity and false discovery rate compared to the simulated data. However, the performance measures for simulated data can still give an approximate idea of the discrimination performance in real data. This was also supported by the findings in real data when the outlier elements were further investigated. There is strong biological evidence for 5 out of the 6 outliers pointing to existing regulatory interactions between genes and transcription factors (Table 4.3). In such a case, the remaining predicted interaction (Skn7p with the *YGP1* gene) is worth being investigated further both through literature and experimentation.

The whisker length can be reduced to let more outliers show up in the analysis. This will identify weaker interactions at the cost of a higher incidence of false positives. We set the whisker length to a relatively loose value of 1.6, to let the number of outliers increase to 38. Out of these 38 potential interactions, nearly half come

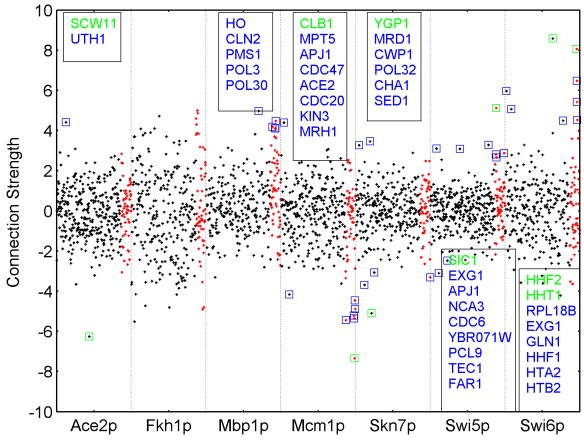

Figure 4.7: **Local Investigation of the Harbison Cell Cycle Transcriptional Regulatory Network.** The black dots correspond to the elements in $\hat{T}_{un}$ which were restricted to 0 in the imposed Harbison network structure, and the red denotes the elements belonging to the new set genes. Blue squares show the outliers defined at a whisker length of 1.6 and green squares denote the outliers defined at a whisker length of 2.8. The gene names in these outlier connections are shown in their respective color, as well. In the y-axis, the values of the connection strengths are shown.

from the new set of genes, as expected. The new genes were curated from literature as cell cycle regulated genes but had no interactions according to the imposed Harbison network structure. Therefore their regulation pattern was non-existent in the network used as prior information, and they immediately showed their regulation pattern in the unrestricted connectivity matrix.

Out of the model set connections, evidence for 10 of them was found in other sources of experimental data (Table 4.3). When the new set was considered the number of connections supported with biological evidence increased to 17 (Table 4.3).

Lastly, there are a considerable number of outliers that can be hypothesized regarding the Skn7p. Apparently, the activity profile of this transcription factor is essentially needed to explain the expression profile of these genes. However, we know from literature that Skn7p's role in cell cycle regulation is through its association with the Mbp1p [19], but there is little information about this role in the literature which is why we choose not to interpret these interactions in more detail here.

Table 4.3: **Predicted Interactions in Cell Cycle Regulatory Network of Yeast**

Whisker Length	TF	Gene	Biological Evidence
2.8	Ace2p	*SCW11*	ChIP [138, 170]
	Mcm1p	***CLB1***	ChIP [153, 170], Gene is known to be regulated in the G2/M phase of the cell cycle [53], Mcm1p is an important transcriptional regulator of this phase [138].
	Swi5p	***SIC1***	Regulation of *SIC1* gene by the Swi5p was already known[85, 138].
	Swi6p	*HHF2*, ***HHT1***	These histone genes (together with their homologues *HHF1* and *HHT2*) were shown to be both MBF (Mbp1p-Swi6p complex) and SBF (Swi4p-Swi6p complex) targets[71].
1.6	Ace2p	*UTH1*	ChIP [170]
	Mbp1p	*RAD27*	ChIP [138, 170]
	Mbp1p	*CWP1*	Comparative Microarray [16]
	Swi5p	*EXG1*, *CLN3*	ChIP [138, 170]
	Swi5p	*CDC6*	Gene's transcription at the end of mitosis is induced by Swi5p [122].
	Swi6p	*RPL18B*, *EXG1*	ChIP [138, 170]
	Mbp1p	***POL3***, ***POL30***, ***PMS1***	Identified as late G1 phase genes regulated by MBF complex [15, 88].
	Swi5	***PCL9***	It is known that the expression of the gene is regulated by Swi5p [1].
	Swi5	***TEC1***	ChIP [138]
	Mcm1p	***ACE2***, ***CDC20***, ***KIN3***, ***MRH1***, ***CDC47***	ChIP [153], Genes are regulated mainly in G2/M phase [36, 53, 132, 136, 142] or M/G1 boundary (*CDC47*) [107] and Mcm1p is known to be involved in the transcriptional regulation of these phases either by direct binding or together with Fkh1p, Fkh2p and Ndd1p [138].
	Swi6p	***HTA2***, ***HTB2***, ***HHF1***	These histone genes are targets for both MBF (Mbp1p-Swi6p complex) and SBF (Swi4p-Swi6p complex) [71, 138].

Only the ones which were supported by biological evidence from literature are shown here. **Bold font** genes belong to the new set of genes whereas the normal font genes are the model set genes.

4.3.5. Conclusions

We present this study as a contribution to both model discrimination and model improvement in the rapidly evolving world of network based approaches. In terms of model discrimination, we have presented measures to discriminate between competing regulatory network structures. Looking at the MSE and MSSRE in decompositions with two different network structures allows to comment on the consistency between the data and the network structure. This indicates which network structure is the most realistic. However, the magnitudes of differences in these measures

between two networks depend on the total strength of the connections differing between them. It is therefore easier to discriminate between networks with strong connections differing between one another. This conclusion makes sense through biological interpretation: strong connections are more easily identifiable and they should be so. This finding is more consistent in MSE whereas MSSRE is more dependent on the specific topologies that are questioned. Therefore, we suggest MSE as a sensitive global measure that discriminates between two different networks.

In terms of model improvement, the relaxation of the topological constraints for the estimation of an unrestricted connectivity matrix allows us to investigate the connections individually. Through this local approach, the unexpectedly strong connections in the unrestricted connectivity matrix can be identified as outliers. These outliers point to existing connections that were lacking in the hypothesized network as has been shown on simulated data. We also showed how the application of the method to the cell cycle regulatory network of *S. cerevisiae* led to the prediction of novel regulatory interactions, improving the information content of the hypothesized network model.

Acknowledgments

Huub C.J. Hoefsloot (NMC, University of Amsterdam) is acknowledged for his helpful comments on the problem of degeneracy.

Appendix

Table S1. Summary of the simulations with high experimental uncertainties. The sensitivity and False Discovery Rate (FDR) values regarding the local part of our approach are given at two different whisker lengths. P_{MSE} and P_{MSSRE} stand for the percentage of the simulations where global measures MSE and MSSRE could discriminate between different networks, respectively. The misconnection levels of the networks that have been compared are stated in the second column.

		Whisker Length = 1.2		Whisker Length = 2.8			
Noise Level	Misconnection Level	FDR	Sensitivity	FDR	Sensitivity	P_{MSE}	P_{MSSRE}
30 %	3% & 5%	0.96	0.66	0.06	0.40	100 %	72 %
	10% & 20%	0.85	0.66	0.06	0.38	100 %	100 %
	20% & 25%	0.79	0.63	0.07	0.34	100 %	82 %
40 %	3% & 5%	0.96	0.61	0.06	0.33	100 %	62 %
	10% & 20%	0.86	0.61	0.04	0.31	100 %	100 %
	20% & 25%	0.80	0.58	0.06	0.28	100 %	82 %
50 %	3% & 5%	0.95	0.57	0.06	0.27	100 %	56 %
	10% & 20%	0.86	0.57	0.04	0.26	100 %	88 %
	20% & 25%	0.81	0.54	0.05	0.23	100 %	80 %

6

Conclusions, Reflections and Perspective

6.1. Analysis of nonlinear kinetic models

6.1.1. Model validation

ODE based kinetic models usually include a large amount of parameters whose values are unknown. A consensus strategy is estimating the unknown parameters from data. The data used for this purpose is typically time series concentration values of biochemical species involved in the model. When a validation step also has to be considered, a pre-determined part of the available data is excluded from the parameter estimation step. The predictive power of the model on the left-out part of the data is used for model validation or selection purposes in the case of multiple competing model structures. In the light of our findings from Chapter 3, we conclude that the partitioning of the data is not a trivial task. Decisions regarding the validity or superiority of a model structure compared to others are affected by the choice of the partitioning scheme. This is because some parts of the data are essential for parameter inference due to reasons originating from the underlying biological properties of the system in question. However, determination of these parts beforehand is not possible since it depends on the parameters of the model which might be unknown. This fundamental drawback associated with the hold-out approach has been underestimated in the field. Stratified random cross validation overcomes the limitations of the hold-out approach by making use of

each part of the data in an iterative and structured manner.

6.1.2. Standards in ODE modeling

Model parameters that have been identified *in vitro* can be accessed via established enzyme databases such as Brenda [131]. However, most kinetic models deposited in model databases such as Biomodels Database [94] contain a substantial number of parameters estimated using *in vivo* concentration data. MIRIAM (Minimum information required in the annotation of models [118]) protocol is an initiative to standardize the curation of quantitative biological models. It sets certain rules primarily regarding the correspondence information of the model and the format in which the model is deposited. One of the requirements entails the usage of a machine readable format such as SBML (Systems Biology Markup Language [67]) which provides a standardized representation of the model in terms of the model constituents and the governing biochemical rules. The optimal parameter values estimated in the associated studies are embedded in the deposited model files and are available to be used by other modelers. However, information about the characteristics of the data such as the experimental conditions under which the data were collected is not provided in a standardized format and is available only in the reference article describing the model. Keeping in mind that independent data are needed for model validation, it is crucial to have detailed information on the data from which the parameters of the model were inferred. It is not difficult to realize that standardizing the presentation of such information is a fairly high target to achieve. However, it is ultimately needed for paving the way to more systematic construction and validation of quantitative models.

6.1.3. Optimal experimental design

There is substantial literature presenting guidelines for optimal experimental design for the analysis of ODE based models [27, 46, 89, 137, 140]. The optimal design is dependent on the ultimate goal of the modeler. The optimal experimental conditions, measurement points and the needed levels of input parameters (such as disturbances to the system) change according to whether the primary goal is better estimation of the parameters or better discrimination of competing model structures [89]. In other words, the experimental conditions that have to be used for parameter inference with less uncertainty are not necessarily the same with those that have to be used for improved selection between alternative model structures.

Our findings in Chapter 3 can be considered as complementary to this fact. In Chapter 3, we have shown that the experimental conditions that are appropriate to

validate a model structure are not necessarily appropriate for model selection. This is because, other competing models might also perform well under these experimental conditions. The concept of optimal experiment design and our cross validation concept agree on the fact that the details regarding the experimental conditions are of vital importance. The application of the optimal targeted experimental design concept is however, almost not existent in the field. Speculating on the reasons; this is partly due to the fact that the optimal conditions for experiments are assumed to be known to biologists through prior biological information on the system. However, efficient construction of reliable quantitative models definitely rely on well designed experiments.

6.1.4. Assessment of predictive power

Models are abstractions of the reality and should be as simple as possible but yet capable of explaining the observed data and of predicting the unobserved data. In Chapter 2, we have shown that cross validation provides a means of calculating the predictive power of kinetic models when applied in a stratified way, leaving out as test set in each turn, sets of data points which are homogeneous in time points and biochemical species. We have also shown that the final assessment of a model's predictive power's sufficiency can be made by comparing it to the predictive power of an unsupervised approach which does not rely on any biochemical knowledge. Smooth principal components analysis (SPCA) works well as such an empirical threshold, since it relies not only on the correlation between the species but also on their time dependent smooth behavior.

It is important to re-emphasize that we assess the predictive power of a model by this approach. The task accomplished here is, hence, not serving the same purposes as residual tests for assessing the quality of model fit [29]. In our approach, the data on which the prediction residuals are obtained have to be excluded from the data that are used for parameter inference. For this purpose, we either leave out interior time points or a set of consecutive time points at the end of the complete time profiles of biochemical species. Upon reflecting on our work, we believe that leaving the interior time points out as test set points can be slightly controversial. This is because, data at a certain time point are not fully independent of the neighboring time points. Therefore, one can argue that when a time point is left out of the training set, the neighboring time points are enough to represent it in the training set. However, we see clear distinctions between the predicted time profiles when different time points are left out of the training set. This proves that removing interior time points leads to the loss of information that they carry even though the neighboring points are still in the training set.

Excluding consecutive time points at the end of the profiles from the training set can be considered as a better approach regarding the same dependency problem. Since forecast analysis is about making predictions regarding unobserved outcomes in the future, it does not suffer from the problem of dependency between the time points. It is also more appealing from a biological point of view. However, this approach has to be applied with care since test sets should not be composed of only the time points that are in the steady-state region. Otherwise, the technique would lead to trivial test sets.

In this respect, leaving out consecutive interior time points (windows of data) as test sets can be more satisfactory [102, 134] for proper application of cross validation on time series data. In this way, when windows of data are left out, the remaining time points would be less representative of the left out time points.

An exceptional case where our approach might not be directly applicable would be a quantitative model consisting of biochemical species with a significantly lagged correlation structure. If the time needed for them to arrive at steady state differs significantly, problems can occur regarding the application of SPCA which depends on the correlation. Therefore, simulations on toy models of such a structure would help to test the applicability of our approach in such cases.

Another possible point of improvement is due to the typical unidentifiability problem of ODE based models. Due to scarcity of the data and high amount of unknown parameters, many of the parameters in such models can only be estimated with very large confidence intervals [125, 129]. The average predictive power of the model across the whole set of parameters in the confidence interval can be, therefore, claimed as a better indicator of the specific predictive power of a model structure instead of the power calculated only at the optimal parameter setting. Such an approach requires a good definition of the confidence intervals of parameter estimates obtained by e.g. bootstrapping approaches [76, 82] or posterior distributions of parameters obtained by a Bayesian perspective [155].

The stability of our approach can be improved by using bootstrapping. Our results from Chapter 2 indicate that model invalidation decisions are more affected by the idiosyncratic noise realization in the data when the experimental noise is high. Complementing our approach with bootstrapping would help to decrease the effect of noise on the decisions when high levels of experimental noise in the data are suspected. In this approach, prediction errors have to be calculated as an average of all bootstrap samples.

As a final remark, we should stress once more that our approach assesses the predictive power of a biochemical model structure based on a given dataset. Therefore, the decisions can change depending upon the dataset. For example, a previously

validated model structure might be invalidated when a dataset with more measurement points in time becomes available. However, this is valid for any validation approach that exploits the consistency between the model predictions and the data.

6.2. Transcriptional regulatory networks

In Chapter 4, we have exploited network component analysis to detect inconsistencies between gene expression data and fixed topology of a transcriptional regulatory network. Discrimination of competing static network models is possible by looking at the measures of consistency. Local improvements of existing topology can also be made through the application of our approach by identifying missing connections. The last feature is valid even when the experimental noise and the number of misconnections and missing connections in the data are fairly high as would be in reality. Furthermore, the local improvement feature of our approach can be extended to sets of genes which were not initially in the proposed network structure. These make it possible to apply our approach iteratively in a modeling-experiment cycle. Potential points of improvement in the networks can be detected by computational means and can be confirmed by targeted experiments.

Our approach, however, might suffer from the limitations of the network component analysis [99] approach upon which it is built. A primary criterion for the applicability of network component analysis is that the transcription factor activity matrix has to be full rank. We conclude that this can be disturbed when the activities of transcription factors are heavily dependent on each other. This problem can be solved by exploiting data from a larger number of experimental conditions. However, this can also result in instability of the transcription factor - gene connection strengths. The experimental conditions used should not be very distinct from each other to keep the connection strength values constant across these conditions. Due to this reason, the selection of the experimental conditions used in such an analysis is a nontrivial task.

6.3. Clustering of large scale biological data

6.3.1. Assessment of validity

Every clustering algorithm gives its own view of the data. Therefore, validating the results from a specific clustering algorithm is necessary. Stability of the resulting clusters is an important feature of the cluster analysis. High stability against small perturbations in the data which are due to the specific realization of the experimental noise shows the reproducibility of the clusters. Therefore, stability provides a means

of validating the analysis. However, it is important to investigate the behavior of the stability measures on the original data relative to a reference value. Analytical calculation of such a reference value would be highly demanding in terms of the assumptions needed for describing the distribution of the data. For this purpose, we have employed a rather different approach in Chapter 5. Observing the stability level obtained from the clustering of a densely distributed, single cluster synthetic dataset helps us to have an understanding of how stable clustering on data without a nontrivial cluster structure would be. We have applied this comparative technique in its very crude form, though.

The stability of the single cluster dataset might be affected by factors such as the size of the dataset, amount and structure of the experimental noise, and the level of covariance between the variables. At relatively high noise levels, the specific realization of the noise can be influential by itself. Therefore, the issue of setting a well defined reference point deserves a detailed study. Simulations by using datasets differing at the levels of the aforementioned factors and considering their average stability level would be beneficial for this purpose.

6.3.2. Parameter optimization in cluster analysis

A major task in a cluster analysis is the optimization of the parameters of a clustering algorithm. In many of the algorithms, the number of clusters imposed is a parameter that can be optimized by using the data. Stability measures employing cross validation [147] and quality measures such as the distance based metric Silhouette Width [126] have been proposed for this task. However, when the clusters in the data are not well separable, the optimal number of clusters obtained by these measures are very low. This leads to very coarse clustering of the data which is too rough for biological interpretation. Optimality might have to be traded off to gain more biological interpretability. Therefore, we need measures that regulate this trade-off. The metric introduced in [52] and the biological homogeneity index of [34] both score the biological relevance of the cluster analysis based on the functional categories to which the genes belong. The biological stability index of [24, 34] takes it one step further and helps in the stability assessment of the co-clustering of functionally related genes. However, it does not specifically regulate the trade-off we have mentioned.

6.3.3. Dealing with vague structures

The clusters in biological data are often not well separable. It is not hard to imagine this phenomenon since we are dealing with overlapping groups of genes that belong

to multiple functional categories. In this manner, fuzzy clustering algorithms provide a suitable framework. However, they require more complex algorithms compared to their traditional counterparts. Consequently, their validation and the optimization of their parameters still deserve active research [164].

Application of different clustering algorithms in parallel is also promising to deal with vaguely structured datasets. Detecting the commonly emerging patterns in the data by multiple approaches is important in the sense that it paves the way to consensus clusters that are free of the individual bias of the algorithms. However, patterns that can only be detected by particular algorithms might also be very important since not every algorithm can reveal different subtle structures in the data. Therefore, not only the comparison but also the reliable integration of different clustering algorithms should attract attention in the field of clustering analysis.

6.3.4. Incorporation of validation and quality measures in cluster analysis software

Software packages are available to perform external validation and to calculate internal quality measures, in most common languages for statistical analyses such as Matlab [162] and R [24]. However, rather dedicated, standalone software for cluster analysis are extensively used by biologists. Validation of clustering has to be conceptually more integrated in the cluster analysis and thus validity measures have to be incorporated also in standalone cluster analysis software.

6.4. Incorporating resampling approaches

6.4.1. Error models for bootstrapping

Applications of bootstrapping proves to be beneficial in the analysis of uncertainty of models and detecting model stability in a variety of systems biology applications including clustering and nonlinear kinetic models of biochemical systems. However, reliable conclusions can be achieved only by employing reliable bootstrap datasets. Reliability of a bootstrap sample depends on how well repeated sampling of the data is mimicked by the bootstrap samples. There are certain assumptions in constructing bootstrap samples and hence, the reliability depends on how realistic these assumptions are. These assumptions are related to the error structure in the original data such as the exchangeability of the experimental error term between the time points or between the biochemical species in the model. Their fulfillment depends on the experimental methods used to collect the original data. Therefore, resampling approaches can prevail more in systems biology if studies focus on modeling

the error structures associated with individual types of experimental methods.

Summary

The paradigm shift from qualitative to quantitative analysis of biological systems brought a substantial number of modeling approaches to the stage of molecular biology research. These include but certainly are not limited to nonlinear kinetic models, static network models and models obtained by the analysis of large scale datasets such as clusters in gene expression data or principal component analysis models. However, the concept of 'model validation' is not encountered as often as the introduction of new models in the field. This leaves many of the proposed models untested and therefore, creates a gap between the number and the reliability of the models. With this thesis, we present computational approaches for model validation and provide guidelines for reliable model validation and selection with examples on real biological data.

The second and third chapters of this thesis focus on nonlinear kinetic models by which we can model the dynamics that underlie processes within a cell. They are usually formulated by using sets of ordinary differential equations (ODE) with many unknown parameters. Most of the time, there are competing hypotheses on the biochemical species, the regulatory relations that these models contain and the governing biochemical rules. This uncertainty brings the need for careful model selection and model validation.

In the **second** chapter, we present a comparative approach for model invalidation employing cross-validation which is a widely used resampling technique in statistics. Our approach is based on assessing the predictive power of a model compared to an unsupervised data analysis method, namely smooth principal components analysis. Low levels of relative predictive power indicate low informative levels of biochemical model structures that are under study. We also present results from the application of our approach on an eicosanoid production model in human and a high osmolarity glycerol pathway model in *Saccharomyces cerevisiae*.

In the **third** chapter, we extend the concept of using cross validation in the analysis of ODE based models and apply it across multiple experimental conditions. The commonly applied method for validating such models is a hold-out validation approach in which a pre-determined part of the data is used for estimating the parameters of the model and the predictions by the model on the remaining part are used to test the model structure and to select the best model structure between

competing hypotheses. However, this strategy is prone to substantial risks. The most important of them is being biased by the underlying biological facts. As an alternative, we introduce a cross validation scheme across multiple experimental conditions for more reliable model selection and validation.

The fourth and the fifth chapters focus on the validation of two conceptually different types of models, both regarding transcriptional regulation. In the first type, a regulatory transcriptional network model is used to summarize the physical association between genes and transcription factors. In the second type, clusters obtained from a cluster analysis summarize the similarity of the expression profiles of genes across various arrays.

In the **fourth** chapter, we present local and global measures to detect inconsistencies between fixed transcriptional regulatory network topologies and gene expression data. The measures we present are based on the supervised decomposition using network component analysis of gene expression data under the topological constraints defined by competing models. Competing transcriptional regulatory networks can be discriminated by using the global measure whereas unknown regulatory interactions can be identified using the local measure. Besides the simulations study through which we show the applicability of the measures, we also present the application of our method on the network model of cell cycle regulation in *Saccharomyces cerevisiae* and introduce potential points of improvement for the network model.

In the **fifth** chapter, we present an assessment of the applicability of cross validation and bootstrapping based stability measures for validation of clusters obtained from k-means clustering on large scale gene expression datasets. Additionally, we apply these approaches for the validation of clusters obtained from *Synechocystis* gene expression data and present biological results regarding the transcriptional control mechanisms regulating the day and night rhythm of the particular organism.

The final chapter of this thesis includes concluding remarks and an overview of the future perspective regarding the construction of more reliable systems biology models.

Samenvatting

De paradigmaverschuiving van kwalitatieve naar kwantitatieve analyse van biologische systemen heeft een aanzienlijk aantal benaderingen voor het modelleren van deze systemen ten tonele gevoerd in moleculair biologisch onderzoek. Deze benaderingen zijn onder andere niet-lineaire kinetische modellen, statistische netwerk modellen en modellen die verkregen zijn door de analyse van grote dataverzamelingen zoals clusters in gen-expressiedata of hoofdcomponentenanalyse modellen. Modelvalidatie heeft de introductie van nieuwe modellen echter niet kunnen bijhouden waardoor veel van de voorgestelde modellen ongetest blijven en er een kloof is ontstaan tussen het aantal modellen en de betrouwbaarheid van die modellen. In dit proefschrift worden computationele benaderingen voor modelvalidatie gepresenteerd en worden richtlijnen gegeven voor betrouwbare modelvalidatie.

In hoofdstukken 2 en 3 van dit proefschrift ligt de nadruk op niet-lineare kinetische modellen waarmee de dynamica die ten grondslag ligt aan processen in de cel gemodelleerd kan worden. Gewoonlijk worden deze processen geformuleerd in termen van gewone differentiaalvergelijkingen met een groot aantal onbekende parameters. Bovendien worden er veelal verschillende hypothesen gebruikt ten aanzien van de betrokken biochemische actoren en hun onderlinge regulatie alsook met betrekking tot de toepassing van biochemische regels.

In het **tweede** hoofdstuk wordt de in de statistiek veelgebruikte 'resampling' techniek van de kruisvalidatie gebruikt voor een vergelijkende methode om modellen te valideren. Deze methode is gebaseerd op het beoordelen van de voorspellende waarde van een model en deze te vergelijken met een data-analyse techniek zonder supervisie namelijk de gladde hoofdcomponentenanalyse. Een lage relatieve voorspellende waarde geeft aan dat er weinig informatie zit in de beoordeelde modellen. De resultaten van de toepassing van deze methode op een model voor de productie van eicosanoïde in de mens en op een model voor het hoge osmolarieteit glycerol reactiepad in *Saccharomyces cerevisiae* worden ook in dit hoofdstuk gepresenteerd.

In het **derde** hoofdstuk wordt het gebruik van kruisvalidatie in de analyse van modellen gebaseerd op gewone differentiaalvergelijkingen verder uitgebreid en wordt het toegepast voor verschillende experimentele condities. De gebruikelijke methode voor de validatie van dergelijke modellen is een 'hold-out' validatie waarbij een vooraf bepaald deel van de data wordt gebruikt voor het schatten van de modelpa-

rameters en de overgebleven data worden gebruikt voor het testen van de modelstructuur en het selecteren van de beste modelstructuur uit verschillende hypothesen. Deze methode brengt echter aanzienlijke risico's met zich mee. De belangrijkste daarvan is bias door de onderliggende biologische gegevens. Als alternatief wordt een kruisvalidatieschema geïntroduceeerd dat meerdere experimentele condities omvat waardoor modelselectie en validatie betrouwbaarder worden.

In het vierde en vijfde hoofdstuk ligt de nadruk op de validatie van twee conceptueel verschillende types van modellen die beide betrekking hebben op regulatie van transcriptie. Het eerste type model maakt gebruik van een transcriptie regulatie netwerkmodel om de fysieke associatie tussen genen en transcriptiefactoren samen te vatten. In het tweede type model worden clusters die verkregen worden via een clusteranalyse gebruikt om overeenkomsten in het expressieprofiel van genen in verschillende arrays samen te vatten.

In het **vierde** hoofdstuk worden lokale en globale maten gepresenteerd voor de detectie van inconsistenties tussen vaste transcriptie regulatie netwerktopologieën en genexpressiedata. Deze maten zijn gebaseerd op de 'supervised' decompositie van genexpressiedata met topologische randvoorwaarden die gedefinieerd zijn door coccurerende modellen, namelijk de netwerk componentenanalyse modellen. Mogelijke transcriptie regulatie netwerken kunnen worden onderscheiden door gebruik te maken van de globale maat terwijl onbekende regulatie patronen met de lokale maat geïdentificeerd kunnen worden. Naast de simulatiestudie om de toepasbaarheid van de maten te demonstreren, wordt de methode ook toegepast op een netwerkmodel voor de celcyclus regulatie in *Saccharomyces cerevisiae*. Tevens worden punten voor mogelijke verbeteringen van dit netwerkmodel aangedragen.

In het **vijfde** hoofdstuk wordt de toepasbaarheid beoordeeld van op kruisvalidatie en bootstrapping gebaseerde maten voor stabiliteit voor de validatie van clusters die zijn verkregen door k-gemiddelden clustering op grote genexpressie dataverzamelingen. Bovendien worden deze methoden gebruikt voor de validatie van clusters van *Synechocystis* genexpressie data en worden biologische resultaten met betrekking tot transcriptie controle mechanismen voor de regulatie van het dag- en nachtrithme van dit organisme gepresenteerd.

Het laatste hoofdstuk van dit proefschrift omvat afsluitende opmerkingen en een toekomstperspectief ten aanzien van de constructie van meer betrouwbare modellen in de systeembiologie.

Acknowledgements

This book is the end product of almost five year long commitment to a PhD project, a working routine, a team of colleagues and multiple corridors of the second floor of the faculty building where I usually walked a lot while thinking. Sometimes, I was mentally in other places, though. Places nearby a calm, blue sea and mountains covered with dried-out bushes, melting at 35 °C, i.e., (under conditions that actually does not let you think of something serious for more than a couple of seconds)... Anyway! Here I am at the very end, happy and ready to tell a couple of nice words who had a direct or indirect role in the realization of this thesis. First of all come my promotor Age Smilde, my co-promotor Huub Hoefsloot and my other supervisor Johan Westerhuis.

Dear Age, thanks for appreciating my work, for the inspiring discussions which motivated me to look from different angles and for the short, spontaneous courses you usually gave in our meetings. I was always impressed by your ability to clarify things very rapidly with your 'graphical thinking'. The last time you did it, the subject was not even scientific at all. The graphs you drew on the board were basically explaining three versions of my thesis! Also, the linear algebra course with you was of particular joy for me. Lastly, after those years I have worked with you, I can predict better when you will say "Now, I am lost".

Dear Huub, thanks for being such a friendly supervisor. You were indeed a daily supervisor to me, you were ready to help whenever I knocked on your door although sometimes that was only to say "stop and just write!". There was always a solution for my mathematical problems when I walked out of the door of your office, sometimes written by you on a paper smaller than half of an A4. Sorry for some rare cases in which my unanswered questions followed you when you tried to sleep at night! I still keep my promise that I don't do that to you at least on fridays.

Dear Johan, thanks for giving me the opportunity to do a PhD with you and being my supervisor in the first half of my PhD period. It is sad that it had to finish unplanned due to your busy schedule. You were always very kind and that helped me a lot to adapt to a new working environment in my early times. I also really enjoyed helping you with the BDA course.

I am also grateful to Gertien Smits and Andreas Angermayr with whom I had fruitful collaborations and to the members of my PhD committee for accepting to

do this time-demanding task. I would like to mention my special gratitude to Betül Kırdar who has been my supervisor in my bachelor and master projects. Dear Betül Hanım, I did not know that I would go all the way to a PhD degree when I started my first project with you 9 years ago. Those were the times that I did not think about the future, anyway. Back then, how you could derive biologically meaningful conclusions from a bunch of metabolite levels that I was measuring was magic for me. Your caring guidance brought me to a point where I could dare doing a PhD in the field.

Many thanks also to people who have contributed to my nice memories about the working environment in the BDA group. Dear everyday coffee-mate Gooitzen, thanks especially for the lovely pattypan squashes from your garden, for reading my paper manuscript in the most careful manner I have ever seen and translating my thesis summary to Dutch, among many other things. Dear Joe, Polina and Mateusz, my international partners of occasional 'PhD blues'! It was very nice to know you right by my side whenever I had the urge of a relaxing, after-work beer. Those evenings will always be remembered. Joe, thanks for the 'eisbein' in Berlin. I wish we could have visited more conferences together. Polina, I am impressed by your ability to manage the PhD and Veronica at the same time. I hope, it will be only Veronica, soon.

Also some people from my early times in the group deserve special thanks for their long-lasting friendliness (each in their own way). Dear Ewa, five years ago I would not easily believe that I could share so much with a woman *from somewhere called Poland*. Now I certainly do! One of the best things that living abroad teaches you... Dear Maikel, thanks for your original Dutch guidance in our trips and the technical tips about thesis writing. Dear Daniel (the 'dude'), thanks for being the cool, helpful senior with great tips about almost everything. Jeroen, thanks for making me realize that some Turkish songs I knew for all my life were actually pyschedelic! Many other great colleagues I worked together, it was a pleasure to meet each of you.

Also thanks to my previous colleagues at (KB440-441) for all the 'systems biology' mentoring I received when I was a master student. Special thanks to Duygu for teaching me a lot, to Tunahan for being one of the reasons why I ended up in Amsterdam, and to the female subset for the breakfasts sprinkled with a bit of academic communication. I miss those a lot.

My dear friends! I am aware that many of you don't have any idea on what I do in this thesis. I hear some of you shouting "Something about biology but not in the lab, right?" Yes, sort of... Actually, doesn't matter! I knew you always wanted the best for me. Ceylan, maybe you understood me the most during all ups and downs

in this period since you experienced similar things. Thanks for becoming a sister to me. Çilo, Melo, Ceyda; I have known you for half of my lifetime. Maybe a bit of a cliche but definitely true: I am very happy that the physical distance between us could not break us apart. All friends from Holland with whom I have shared very nice memories; in particular, Sinan, Erman and Poonam. You made it here much warmer for me. Dear Ezgi, thanks furthermore for your enthusiasm in helping me adapt to Holland. Friends from good old İstanbul! Seda (As.), Seda (Ak.), Umut, Hicran, Deniz, Sezo, İbo, Özgür, Mustafa, Mümin, Birim, Fidan; thanks for opening your doors for us, the two travelers who are longing for your friendship. Dear people of 'Laterna', listening to our recordings was always great joy to me when I felt the need of connecting with my past.

Dear Aydın abi, maybe my motivation for science did not start with the microscope you once had given me. But, your wisdom and never-ending teaching motivation as the only PhD in the family and the visits I made to your physics lab at Boğaziçi definitely shaped my thoughts about what I wanted to do in life.

(Şengül) anne, (İbrahim) baba; bana böylesine yakın, yeni bir aile olduğunuz için kendimi çok şanslı hissediyorum. İyi ki varsınız. Bundan sonra kiraz bahçesini daha çok ziyaret edebilmek umuduyla...

My dear brother! The small sister who wanted to follow you everywhere you went, is now finishing PhD. This means that you are pretty old now. Enjoy! When I look to the past, I see very clearly that you had great influence on me. You and your patient support play a role in this thesis. I am looking forward to our next adventure with you and Deniz abla.

Yunus, my dearest Yunlu(m), thanks for sharing every single moment with me for the last eight years, thanks for accepting the challenge and coming to Amsterdam with me, thanks for standing by me even at times that I was too difficult to deal with. You always motivated me to work more and also to be more aware of my successes and enjoy them. Both this work and my life are definitely better with you.

Son olarak, sevgili anne ve babacığım; küçüklüğümden beri bana koşulsuz olarak güvendiğiniz için teşekkürler. Sanırım, beni ben yapan en önemli şeylerden biri bu. Canım annem! Beni buraya gönderirken havaalanında el sallamıştın bana. Dün gibi aklımda. Şimdi o başladığım yolculuk bitiyor, keşke sen de görebilseydin, 'aferin benim akıllı kızıma' derdin. Seni çok seviyorum.

References

[1] B. L. Aerne, A. L. Johnson, J. H. Toyn, and L. H. Johnston. Swi5 controls a novel wave of cyclin synthesis in late mitosis. *Molecular Biology of the Cell*, 9(4):945–956, 1998.

[2] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[3] R. Albert. Scale-free networks in cell biology. *Journal of cell science*, 118(21):4947–4957, 2005.

[4] E. Alpaydin. *Introduction to machine learning*. MIT press, 2004.

[5] L. Ana and A. K. Jain. Robust data clustering. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–128. IEEE, 2003.

[6] J. Anderson and A. Papachristodoulou. On validation and invalidation of biological models. *BMC Bioinformatics*, 10(1):1–13, 2009.

[7] S. A. Angermayr, P. van Alphen, D. Hasdemir, G. Kramer, M. Iqbal, W. van Grondelle, H. C. Hoefsloot, Y. H. Choi, and K. J. Hellingwerf. Dynamics of the molecular composition and physiology of synechocystis sp. pcc 6803 subject to a circadian regime relevant for mass culturing. *in preparation*.

[8] K. Y. Arga, Z. İ. Önsan, B. Kırdar, K. Ö. Ülgen, and J. Nielsen. Understanding signaling in yeast: Insights from network analysis. *Biotechnology and bioengineering*, 97(5):1246–1258, 2007.

[9] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[10] M. Ashyraliyev, Y. Fomekong-Nanfack, J. A. Kaandorp, and J. G. Blom. Systems biology: parameter estimation for biochemical models. *FEBS journal*, 276(4):886–902, 2009.

[11] S. Asur, D. Ucar, and S. Parthasarathy. An ensemble framework for clustering protein–protein interaction networks. *Bioinformatics*, 23(13):i29–i40, 2007.

[12] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann. Structure and evolution of transcriptional regulatory networks. *Current opinion in structural biology*, 14(3):283–291, 2004.

[13] D. G. Bates and C. Cosentino. Validation and invalidation of systems biology models using robustness analysis. *IET systems biology*, 5(4):229–44, July 2011.

[14] N. Battchikova, M. Eisenhut, and E.-M. Aro. Cyanobacterial ndh-1 complexes: novel insights and remaining puzzles. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1807(8):935–944, 2011.

[15] G. A. Bauer and M. J. Burgers. Molecular cloning, structure and expression of the yeast proliferating cell nuclear antigen gene. *Nucleic Acids Research*, 18(2):261–265, 1990.

[16] J. M. Bean, E. D. Siggia, and F. R. Cross. High functional overlap between mlui cell-cycle box binding factor and swi4/6 cell-cycle box binding factor in the g1/s transcriptional program in saccharomyces cerevisiae. *Genetics*, 171:49–61, 2005.

[17] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Pacific symposium on biocomputing*, volume 7, pages 6–17, 2001.

[18] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536–540, 2000.

[19] N. Bouquin, A. L. Johnson, B. A. Morgan, and L. H. Johnston. Association of the cell cycle transcription factor mbp1 with the skn7 response regulator in budding yeast. *Molecular Biology of the Cell*, 10(10):3389–3400, 1999.

[20] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: forecasting and control*. Holden Day, 1976.

[21] J. N. Breckenridge. Replicating cluster analysis: Method, consistency, and validity. *Multivariate Behavioral Research*, 24(2):147–161, 1989.

[22] R. Bro and H. a. L. Kiers. A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics*, 17(5):274–286, June 2003.

[23] R. Bro, K. Kjeldahl, a. K. Smilde, and H. a. L. Kiers. Cross-validation of component models: a critical look at current methods. *Analytical and bioanalytical chemistry*, 390(5):1241–51, Mar. 2008.

[24] G. Brock, V. Pihur, S. Datta, and S. Datta. clvalid, an r package for cluster validation. *Journal of Statistical Software (Brock et al., March 2008)*, 2011.

[25] F. J. Bruggeman, J. J. Hornberg, F. C. Boogerd, and H. V. Westerhoff. Introduction to systems biology. In *Plant Systems Biology*, pages 1–19. Springer, 2007.

[26] M. W. Buczynski, D. S. Dumlao, and E. A. Dennis. Thematic review series:

Proteomics. an integrated omics analysis of eicosanoid biology. *Journal of Lipid Research*, 50(6):1015–1038, June 2009.

[27] A. G. Busetto, A. Hauser, G. Krummenacher, M. Sunnåker, S. Dimopoulos, C. S. Ong, J. Stelling, and J. M. Buhmann. Near-optimal experimental design for model selection in systems biology. *Bioinformatics*, 29(20):2625–2632, 2013.

[28] T. Cakır, M. M. Hendriks, J. A. Westerhuis, and A. K. Smilde. Metabolic network discovery through reverse engineering of metabolome data. *Metabolomics*, 5(3):318–329, 2009.

[29] G. Cedersund and J. Roll. Systems biology: model based evaluation and comparison of potential explanations for given biological data. *FEBS journal*, 276(4):903–922, 2009.

[30] C. Chang, Z. Ding, Y. S. Hung, and P. C. W. Fung. Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data. *Bioinformatics*, 24(11):1349–1358, Apr. 2008.

[31] T. F. Coleman and Y. Li. On the convergence of interior-reflective newton methods for nonlinear minimization subject to bounds. *Mathematical programming*, 67(1-3):189–224, 1994.

[32] T. F. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on optimization*, 6(2):418–445, 1996.

[33] M. Contreras, R. Zadoks, H. G. Allore, and Y. H. Schukken. Bootstrapping to obtain confidence intervals for parameters in ordinary differential equations - infectious disease models. *Unpublished manuscript*, 2000.

[34] S. Datta and S. Datta. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC bioinformatics*, 7(1):397, 2006.

[35] R. Díaz-Uriarte and S. A. De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.

[36] P. R. Dohrmann, G. Butler, K. Tamai, S. Dorland, J. R. Greene, D. J. Thiele, and D. J. Stillman. Parallel pathways of gene regulation: homologous regulators swi5 and ace2 differentially control transcription of ho and chitinase. *Genes & Development*, 6(1):93–104, 1992.

[37] F. B. du Preez, D. D. van Niekerk, B. Kooi, J. M. Rohwer, and J. L. Snoep. From steady-state to synchronized yeast glycolytic oscillations i: model construction. *FEBS Journal*, 279(16):2810–2822, Aug. 2012.

[38] F. B. du Preez, D. D. van Niekerk, and J. L. Snoep. From steady-state to synchronized yeast glycolytic oscillations ii: model validation. *The FEBS*

journal, 279(16):2823–36, Aug. 2012.

[39] D. C. Ducat, J. C. Way, and P. A. Silver. Engineering cyanobacteria to generate high-value products. *Trends in biotechnology*, 29(2):95–103, 2011.

[40] J. C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104, 1974.

[41] S. Durmuş Tekir, K. Yalçın Arga, and K. Ö. Ülgen. Drug targets for tumorigenesis: insights from structural analysis of egfr signaling network. *Journal of biomedical informatics*, 42(2):228–236, 2009.

[42] B. Efron. Bootstrap methods: Another look at the jackknife. *The annals of applied statistics*, 7(1):1–26, 1979.

[43] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*, volume 57. CRC press, 1994.

[44] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

[45] W. Filipowicz, S. N. Bhattacharyya, and N. Sonenberg. Mechanisms of post-transcriptional regulation by microrna: are the answers in sight? *Nature Reviews Genetics*, 9(2):102–114, 2008.

[46] R. J. Flassig and K. Sundmacher. Optimal design of stimulus experiments for robust discrimination of biochemical reaction networks. *Bioinformatics*, 28(23):3089–3096, 2012.

[47] M. Flotmann, J. Schaber, S. Hoops, E. Klipp, and P. Mendes. Modelmage: a tool for automatic model generation, selection and management. *Genome Informatics*, 20:52–63, 2008.

[48] M. Frank, M. H. Chehreghani, and J. M. Buhmann. *The minimum transfer cost principle for model-order selection*. Springer, 2011.

[49] M. Frigge, D. Hoaglin, and B. Iglewicz. Some implementations of the boxplot. *American Statistician*, pages 50–54, 1989.

[50] S. J. Galbraith, L. M. Tran, and J. C. Liao. Transcriptome network component analysis with limited microarray data. *Bioinformatics (Oxford, England)*, 22(15):1886–94, Aug. 2006.

[51] A. P. Gasch and M. B. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol*, 3(11):1–22, 2002.

[52] I. Gat-Viks, R. Sharan, and R. Shamir. Scoring clustering solutions by their biological relevance. *Bioinformatics*, 19(18):2381–2389, 2003.

[53] J. B. Ghiara, H. E. Richardson, K. Sugimoto, M. Henze, D. J. Lew, C. Wittenberg, and S. I. Reed. A cyclin b homolog in s. cerevisiae: Chronic activation of

the cdc28 protein kinase by cyclin prevents exit from mitosis. *Cell*, 65(1):163 – 174, 1991.

[54] S. Gupta, M. R. Maurya, D. L. Stephens, E. A. Dennis, and S. Subramaniam. An integrated model of eicosanoid metabolism and signaling based on lipidomics flux analysis. *Biophysical journal*, 96(11):4542–51, June 2009.

[55] M. C. Gustin, J. Albertyn, M. Alexander, and K. Davenport. Map kinase pathways in the yeastsaccharomyces cerevisiae. *Microbiology and Molecular biology reviews*, 62(4):1264–1300, 1998.

[56] M. Hafner, H. Koeppl, M. Hasler, and A. Wagner. 'glocal' robustness analysis and model discrimination for circadian oscillators. *PLoS Comput Biol*, 5(10):e1000534, Oct. 2009.

[57] P. Hall and S. R. Wilson. Two guidelines for bootstrap hypothesis testing. *Biometrics*, pages 757–762, 1991.

[58] J. Handl, J. Knowles, and D. B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005.

[59] C. Harbison, D. Gordon, T. Lee, N. Rinaldi, K. Macisaac, T. Danford, N. Hannett, J. Tagne, D. Reynolds, J. Yoo, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99, 2004.

[60] D. Hasdemir, H. C. Hoefsloot, J. A. Westerhuis, and A. K. Smilde. How informative is your kinetic model?: using resampling methods for model invalidation. *BMC Systems Biology*, 8(1):61, 2014.

[61] M. D. Haunschild, B. Freisleben, R. Takors, and W. Wiechert. Investigating the dynamic behavior of biochemical networks using model families. *Bioinformatics*, 21(8):1617–1625, 2005.

[62] D. M. Hendrickx, M. M. Hendriks, P. H. Eilers, A. K. Smilde, and H. C. Hoefsloot. Reverse engineering of metabolic networks, a critical assessment. *Molecular Biosystems*, 7(2):511–520, 2011.

[63] S. Hohmann. Osmotic stress signaling and osmoadaptation in yeasts. *Microbiology and Molecular Biology Reviews*, 66(2):300–372, 2002.

[64] G. Hongyi Li and Maddala. Bootstrapping time series models. *Econometric reviews*, 15(2):115–158, 1996.

[65] R. Horn. Statistical methods for model discrimination. applications to gating kinetics and permeation of the acetylcholine receptor channel. *Biophysical Journal*, 51(2):255–263, 1987.

[66] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[67] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, et al. The systems

biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[68] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl 1):S233–S240, 2002.

[69] S. Imoto, T. Higuchi, S. Kim, E. Jeong, and S. Miyano. Residual bootstrapping and median filtering for robust estimation of gene networks from microarray data. In *Computational Methods in Systems Biology*, pages 149–160. Springer, 2005.

[70] M. Ishiura, S. Kutsuna, S. Aoki, H. Iwasaki, C. R. Andersson, A. Tanabe, S. S. Golden, C. H. Johnson, and T. Kondo. Expression of a gene cluster kaiabc as a circadian feedback process in cyanobacteria. *Science*, 281(5382):1519–1523, 1998.

[71] V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Synder, and P. O. Brown. Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. *Nature*, 409:533–538, 2001.

[72] J. J. Jansen, H. C. J. Hoefsloot, H. F. M. Boelens, J. van der Greef, and A. K. Smilde. Analysis of longitudinal metabolomics data. *Bioinformatics (Oxford, England)*, 20(15):2438–46, Oct. 2004.

[73] M. Janssen, J. Tramper, L. R. Mur, and R. H. Wijffels. Enclosed outdoor photobioreactors: light regime, photosynthetic efficiency, scale-up, and future prospects. *Biotechnology and Bioengineering*, 81(2):193–210, 2003.

[74] R. Johansson, P. Strålfors, and G. Cedersund. Combining test statistics and models in bootstrapped model rejection: it is a balancing act. *BMC systems biology*, 8(1):46, 2014.

[75] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

[76] M. Joshi, A. Seidel-Morgenstern, and A. Kremling. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metabolic engineering*, 8(5):447–55, Sept. 2006.

[77] J. Josse and F. Husson. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153(2):79–99, 2012.

[78] K. L. Kadam, E. C. Rydholm, and J. D. McMillan. Development and validation of a kinetic model for enzymatic saccharification of lignocellulosic biomass. *Biotechnology progress*, 20(3):698–705, 2004.

[79] M. K. Kerr and G. A. Churchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences*, 98(16):8961–8965, 2001.

[80] H. a. L. Kiers. Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, 62(2):251–266, 1997.

[81] S. Kimura, Y. Shiraishi, and M. Okada. Inference of genetic networks using lpms: assessment of confidence values of regulations. *Journal of Bioinformatics and Computational Biology*, 8(04):661–677, 2010.

[82] P. D. W. Kirk and M. P. H. Stumpf. Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data. *Bioinformatics (Oxford, England)*, 25(10):1300–6, May 2009.

[83] E. Klipp, W. Liebermeister, C. Wierling, A. Kowald, H. Lehrach, and R. Herwig. Systems biology: A textbook, 2009.

[84] E. Klipp, W. Liebermeister, C. Wierling, A. Kowald, H. Lehrach, and R. Herwig. *Systems biology*. John Wiley & Sons, 2013.

[85] D. Knapp, L. Bhoite, D. Stillman, and K. Nasmyth. The transcription factor swi5 regulates expression of the cyclin kinase inhibitor p40sic1. *Mol. Cell. Biol.*, 16(10):5701–5707, 1996.

[86] O. Kotte, J. Zaugg, and M. Heinemann. Bacterial adaptation through distributed sensing of metabolic fluxes. *Mol Syst Biol*, 6:355, 2010.

[87] N. Krämer, J. Schäfer, and A.-L. Boulesteix. Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC bioinformatics*, 10(1):384, 2009.

[88] W. Kramer, B. Fartmann, and E. Ringbeck. Transcription of muts and mutl-homologous genes in saccharomyces cerevisiae during the cell cycle. *Molecular and General Genetics MGG*, 252:275–283, 1996. 10.1007/BF02173773.

[89] C. Kreutz and J. Timmer. Systems biology: experimental design. *FEBS journal*, 276(4):923–942, 2009.

[90] L. Kuepfer, M. Peter, U. Sauer, and J. Stelling. Ensemble modeling for analysis of cell signaling dynamics. *Nature biotechnology*, 25(9):1001–6, Sept. 2007.

[91] R. G. Labiosa, K. R. Arrigo, C. J. Tu, D. Bhaya, S. Bay, A. R. Grossman, and J. Shrager. Examination of diel changes in global transcript accumulation in synechocystis (cyanobacteria) 1. *Journal of phycology*, 42(3):622–636, 2006.

[92] T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann. Stability-based validation of clustering solutions. *Neural computation*, 16(6):1299–1323, 2004.

[93] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 16–22. ACM, 1999.

[94] N. Le Novère, B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, J. L. Snoep, and M. Hucka. BioModels Database: a free, centralized database of curated, published, quantitative

kinetic models of biochemical and cellular systems. *Nucleic Acids Research*, 34(Database issue):D689–D691, Jan 2006.

[95] D. J. Lea-Smith, N. Ross, M. Zori, D. S. Bendall, J. S. Dennis, S. A. Scott, A. G. Smith, and C. J. Howe. Thylakoid terminal oxidases are essential for the cyanobacterium synechocystis sp. pcc 6803 to survive rapidly changing light intensities. *Plant physiology*, 162(1):484–495, 2013.

[96] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, et al. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[97] R. Lehmann, R. Machné, J. Georg, M. Benary, I. M. Axmann, and R. Steuer. How cyanobacteria pose new problems to old methods: challenges in microarray time series analysis. *BMC bioinformatics*, 14(1):133, 2013.

[98] E. Levine and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Neural computation*, 13(11):2573–2593, 2001.

[99] J. C. Liao, R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury. Network component analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences*, 100(26):15522–15527, 2003.

[100] H. Link, K. Kochanowski, and U. Sauer. Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo. *Nat Biotech*, 31(4):357–361, Apr. 2013.

[101] S. Lloyd. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137, 1982.

[102] H. Lodhi and D. Gilbert. Bootstrapping parameter estimation in dynamic systems. In *Discovery Science*, pages 194–208. Springer, 2011.

[103] J. Macia, S. Regot, T. Peeters, N. Conde, R. Sole, and F. Posas. Dynamic signaling in the hog1 mapk pathway relies on high basal signal transduction. *Science signaling*, 2(63):ra13, 2009.

[104] K. D. MacIsaac, T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo, and E. Fraenkel. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, 7(1):113, Jan. 2006.

[105] T. Maiwald and J. Timmer. Dynamical modeling and multi-experiment fitting with potterswheel. *Bioinformatics*, 24(18):2037–2043, 2008.

[106] L. Marucci, S. Santini, M. di Bernardo, and D. di Bernardo. Derivation, identification and validation of a computational model of a novel synthetic regulatory network in yeast. *Journal of mathematical biology*, 62(5):685–706, 2011.

[107] C. J. McInerny, J. F. Partridge, G. E. Mikesell, D. P. Creemer, and L. L. Breeden. A novel mcm1-dependent element in the swi4, cln3, cdc6, and cdc47 promoters activates m/g1-specific transcription. *Genes & Development*, 11(10):1277–1288, 1997.

[108] M. Meilă. Comparing clusterings by the variation of information. In *Learning theory and kernel machines*, pages 173–187. Springer, 2003.

[109] P. Mendes, D. Camacho, and A. De La Fuente. Modelling and simulation for metabolomics data analysis. *Biochemical Society Transactions*, 33(6):1427–1429, 2005.

[110] P. Mendes and D. Kell. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, 14(10):869–883, 1998.

[111] A. Milias-Argeitis, R. Porreca, S. Summers, and J. Lygeros. Bayesian model selection for the yeast gata-factor network: A comparison of computational approaches. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 3379–3384, 2010.

[112] C. G. Moles, P. Mendes, and J. R. Banga. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome research*, 13(11):2467–74, Nov. 2003.

[113] A. M. Molinaro, R. Simon, and R. M. Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 2005.

[114] T. Mori, B. Binder, and C. H. Johnson. Circadian gating of cell division in cyanobacteria growing with average doubling times of less than 24 hours. *Proceedings of the National Academy of Sciences*, 93(19):10183–10188, 1996.

[115] T. Müller, D. Faller, J. Timmer, I. Swameye, O. Sandra, and U. Klingmüller. Tests for cycling in a signalling pathway. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(4):557–568, 2004.

[116] M. Nakao, S. Okamoto, M. Kohara, T. Fujishiro, T. Fujisawa, S. Sato, S. Tabata, T. Kaneko, and Y. Nakamura. Cyanobase: the cyanobacteria genome database update 2010. *Nucleic acids research*, page gkp915, 2009.

[117] L. Nedbal, M. Trtilek, J. Červeny̌, O. Komárek, and H. B. Pakrasi. A photobioreactor system for precision cultivation of photoautotrophic microorganisms and for high-content analysis of suspension dynamics. *Biotechnology and bioengineering*, 100(5):902–910, 2008.

[118] N. L. Novere, A. Finney, M. Hucka, U. S. Bhalla, F. Campagne, J. Collado-Vides, E. J. Crampin, M. Halstead, E. Klipp, P. Mendes, et al. Minimum information requested in the annotation of biochemical models (miriam). *Nature biotechnology*, 23(12):1509–1515, 2005.

[119] B. O. Palsson. *Systems biology*. Cambridge university press, 2006.

[120] K. R. Patil and J. Nielsen. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8):2685–2689, 2005.

[121] F. Pauly, K. E. Smedby, M. Jerkeman, H. Hjalgrim, M. Ohlsson, R. Rosenquist, C. A. Borrebaeck, and C. Wingren. Identification of b-cell lymphoma subsets by plasma protein profiling using recombinant antibody microarrays. *Leukemia research*, 38(6):682–690, 2014.

[122] S. Piatti, C. Lengauer, and K. Nasmyth. Cdc6 is an unstable protein whose de novo synthesis in g1 is important for the onset of s phase and for preventing a 'reductional' anaphase in the budding yeast saccharomyces cerevisiae. *The EMBO Journal*, 14(15):3788, 1995.

[123] F. Posas, S. M. Wurgler-Murphy, T. Maeda, E. A. Witten, T. C. Thai, and H. Saito. Yeast hog1 map kinase cascade is regulated by a multistep phosphorelay mechanism in the sln1–ypd1–ssk1 "two-component" osmosensor. *Cell*, 86(6):865–875, 1996.

[124] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[125] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[126] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[127] C. Sabatti and G. M. James. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22(6):739–746, 2006.

[128] J. Schaber, R. Baltanas, A. Bush, E. Klipp, and A. Colman-Lerner. Modelling reveals novel roles of two parallel signalling pathways and homeostatic feedbacks in yeast. *Molecular systems biology*, 8(622):622, Jan. 2012.

[129] J. Schaber and E. Klipp. Model-based inference of biochemical parameters and dynamic properties of microbial signal transduction networks. *Current Opinion in Biotechnology*, 22(1):109–116, 2011.

[130] R. Schenkendorf, M. Mangold, and A. Kremling. Optimal experimental design with the sigma point method. *IET systems biology*, 3(1):10–23, 2009.

[131] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg. Brenda, the enzyme database: updates and major new developments. *Nucleic acids research*, 32(suppl 1):D431–D433, 2004.

[132] T. Schuster, C. Price, W. Rossol, and B. Kovacech. New cell cycle regulated genes in the yeast saccharomyces cerevisiae. *Recent Results in Cancer Research*, 143:251–61, 1997.

[133] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[134] J. Shao and D. Tu. *The jackknife and bootstrap*. Springer, 1995.

[135] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. a. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. a. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. a. Cebula, J. J. Chen, J. Cheng, T.-M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X.-h. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. a. Hauser, S. Hester, H. Hong, P. Hurban, S. a. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q.-Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. a. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. a. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W. Slikker. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, 24(9):1151–61, Sept. 2006.

[136] M. Shirayama, W. Zachariae, R. Ciosk, and K. Nasmyth. The polo-like kinase cdc5p and the wd-repeat protein cdc20p/fizzy are regulators and substrates of the anaphase promoting complex in saccharomyces cerevisiae. *The EMBO journal*, 17(5):1336–1349, 1998.

[137] D. Silk, P. D. Kirk, C. P. Barnes, T. Toni, and M. P. Stumpf. Model selection in systems biology depends on experimental design. *PLoS computational biology*, 10(6):e1003650, 2014.

[138] I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106(6):697 – 708, 2001.

[139] R. M. Simon, J. Subramanian, M.-C. Li, and S. Menezes. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Briefings in bioinformatics*, 12(3):203–214, 2011.

[140] D. Skanda and D. Lebiedz. An optimal experimental design approach to model discrimination in dynamic biochemical systems. *Bioinformatics*, 26(7):939–945, 2010.

[141] S. Smit, M. J. van Breemen, H. C. Hoefsloot, A. K. Smilde, J. M. Aerts, and C. G. De Koster. Assessing the statistical validity of proteomics based biomarkers. *Analytica Chimica Acta*, 592(2):210–217, 2007.

[142] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.

[143] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–147, 1974.

[144] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(suppl 1):D561–D568, 2011.

[145] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature genetics*, 22(3):281–285, 1999.

[146] B. Teusink, J. Passarge, C. A. Reijenga, E. Esgalhado, C. C. van der Weijden, M. Schepper, M. C. Walsh, B. M. Bakker, K. van Dam, H. V. Westerhoff, and J. L. Snoep. Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? testing biochemistry. *European Journal of Biochemistry*, 267(17):5313–5329, 2000.

[147] R. Tibshirani and G. Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005.

[148] J. Timmer, T. G. M. Uller, I. Swameye, O. Sandra, and U. Klingmuller. Modeling the nonlinear dynamics of cellular signal transduction. *International Journal of Bifurcation and Chaos*, 14(6):2069–2079, 2004.

[149] T. Toni and M. P. H. Stumpf. Simulation-based model selection for dynamical

systems in systems and population biology. *Bioinformatics*, 26(1):104–110, Jan. 2010.

[150] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

[151] L. M. Tran, M. P. Brynildsen, K. C. Kao, J. K. Suen, and J. C. Liao. gNCA: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. *Metabolic engineering*, 7(2):128–41, Mar. 2005.

[152] L. M. Tran, D. R. Hyduke, and J. C. Liao. Trimming of mammalian transcriptional networks using network component analysis. *BMC bioinformatics*, 11(1):511, Jan. 2010.

[153] B. B. Tuch, D. J. Galgoczy, A. D. Hernday, H. Li, and A. D. Johnson. The evolution of combinatorial gene regulation in fungi. *PLoS Biol*, 6(2):e38, 02 2008.

[154] F. E. Turkheimer, R. Hinz, and V. J. Cunningham. On the undecidability among kinetic models: From model selection to model averaging. *J Cereb Blood Flow Metab*, 23(4):490–498, Apr. 2003.

[155] J. Vanlier, C. A. Tiemann, P. A. Hilbers, and N. A. van Riel. An integrated strategy for prediction uncertainty analysis. *Bioinformatics*, 28(8):1130–1135, 2012.

[156] L. J. van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.

[157] M. P. Verouden. *Fusing Prior Knowledge with microbial metabolomics*. PhD thesis, University of Amsterdam, 2012.

[158] E. A. von Wobeser, B. W. Ibelings, J. Bok, V. Krasikov, J. Huisman, and H. C. Matthijs. Concerted changes in gene expression and cell physiology of the cyanobacterium synechocystis sp. strain pcc 6803 during transitions between nitrogen and light-limited growth. *Plant physiology*, 155(3):1445–1457, 2011.

[159] V. Vyshemirsky and M. Girolami. Bayesian ranking of biochemical system models. *Bioinformatics (Oxford, England)*, 24(6):833–9, Mar. 2008.

[160] E.-J. Wagenmakers, R. Ratcliff, P. Gomez, and G. J. Iverson. Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, 48(1):28–50, 2004.

[161] S. Wagner and D. Wagner. *Comparing clusterings: an overview*. Universität

Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.

[162] K. Wang, B. Wang, and L. Peng. Cvap: Validation for cluster analyses. *Data Science Journal*, 8:88–93, 2009.

[163] R.-S. Wang, A. Saadatpour, and R. Albert. Boolean modeling in systems biology: an overview of methodology and applications. *Physical biology*, 9(5):055001, 2012.

[164] W. Wang and Y. Zhang. On fuzzy cluster validity indices. *Fuzzy sets and systems*, 158(19):2095–2117, 2007.

[165] R. Wehrens, H. Putter, and L. Buydens. The bootstrap: a tutorial. *Chemometrics and intelligent laboratory systems*, 54(1):35–52, 2000.

[166] J. A. Westerhuis, E. P. P. A. Derks, H. C. J. Hoefsloot, and A. K. Smilde. Grey component analysis. *Journal of Chemometrics*, 21(10-11):474–485, 2007.

[167] J. A. Westerhuis, H. C. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. van Velzen, J. P. van Duijnhoven, and F. A. van Dorsten. Assessment of plsda cross validation. *Metabolomics*, 4(1):81–89, 2008.

[168] R. H. Wijffels, O. Kruse, and K. J. Hellingwerf. Potential of industrial biotechnology with cyanobacteria and eukaryotic microalgae. *Current opinion in biotechnology*, 24(3):405–413, 2013.

[169] D. A. Williams. Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics*, pages 23–32, 1970.

[170] C. T. Workman, H. C. Mak, S. McCuine, J.-B. Tagne, M. Agarwal, O. Ozier, T. J. Begley, L. D. Samson, and T. Ideker. A systems approach to mapping dna damage response pathways. *Science*, 312(5776):1054–1059, 2006.

[171] K. Yang, W. Ma, H. Liang, Q. Ouyang, C. Tang, and L. Lai. Dynamic simulations on the arachidonic acid metabolic network. *PLoS computational biology*, 3(3):e55, Mar. 2007.

[172] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318, 2001.

[173] T. Yu and K.-C. Li. Inference of transcriptional regulatory network by two-stage constrained space factor analysis. *Bioinformatics*, 21(21):4033–4038, 2005.

[174] B. J. H. Zijlstra and H. A. L. Kiers. Degenerate solutions obtained from several variants of factor analysis. *Journal of Chemometrics*, 16(11):596–605, 2002.

About the author:

Dicle Hasdemir (Durmuş) was born on the 21st of October, 1984 in Diyarbakır, the city characterized by the river Tigris from which her name was borrowed. After finishing İstanbul High School of Science, she attended Department of Chemical Engineering at Boğaziçi University from where she received her B.Sc. degree with honor in 2006. She continued her master studies in the same department under supervision of Prof. dr. Betül Kırdar and Prof. dr. Z. İlsen Önsan with a focus on computational systems biology. During her master study, she also worked as a visiting researcher in the Center for Microbial Biotechnology at Denmark Technical University, under supervision of Prof. Jens Nielsen. In 2008, she completed her M.Sc. studies with high honor. Her thesis was focused on the network based integration of transcriptome and interactome data to reveal the glucose regulation pathways in yeast. In September 2010, she started her PhD research in Biosystems Data Analysis group at University of Amsterdam. She was supervised by Prof. dr. Age Smilde, Dr. Huub Hoefsloot and Dr. Johan Westerhuis. She currently works as a post-doctoral researcher at the Department of Clinical Epidemiology, Biostatistics and Bioinformatics (KEBB), Academic Medical Center, UvA in a joint project with GlaxoSmithKline Vaccines Research on systems biological approaches for the integration of clinical and genomics data. She likes playing clarinet, growing vegetables in her $2m^2$ garden, hugging cats despite her allergy, and sunny weather.