

# Statistical Model Identification: Dynamical Processes and Large-Scale Networks in Systems Biology

**Dissertation**  
zur Erlangung des akademischen Grades  
**Doktor-Ingenieur**  
(Dr.-Ing.)

von: Dipl.-Phys. Robert Johann Flassig  
geb. am: 20.06.1981  
in: Pritzwalk

genehmigt durch die Fakultät für Verfahrens- und Systemtechnik der Otto-von-Guericke-Universität Magdeburg

Gutachter: Prof. Dr.-Ing. Kai Sundmacher  
Prof. Dr. rer. nat. Inna Lavrik  
Prof. Dr. rer. nat. Wolfgang Wiechert

eingereicht am: 21. März 2014  
Promotionskolloquium am: 1. September 2014

---

## Abstract

The use of mathematical models for analyzing complex biological processes, including metabolism, signal transduction and gene regulation in mammalian cells or bacteria, is a powerful approach to obtain deep systems understanding. However, this approach is in the need of realistic, predictive mathematical models. During the modeling identification of such complex systems, scientists have to cope with numerous challenges, e.g. limited knowledge about the underlying mechanisms, lack of sufficient dynamic or static experimental data, large experimental and biological variability. This work presents methodological solutions for model identification of (i) ordinary differential equation systems describing dynamic cellular processes and (ii) large-scale biochemical interaction networks based on high-throughput data.

The solution to part (i) addresses the problem of robustly designing a discriminative cell stimulus in the presence of distributed model parameters. Such a robustification of an experimental design is especially important at the beginning of the model identification phase, since many of the parameters are initially badly constraint, and thus most often lead to highly misspecified experiments. Designing an optimal stimulus profile in combination with parameter uncertainty considerations results in a numerically intense dynamic optimization problem. Using the sigma point method, a numerically feasible robust open-loop-control method is developed. As is shown in this work, the developed method is ideally suited for highly nonlinear models with widely distributed model parameters.

The solution to part (ii) presents the modular identification framework TRANSWESD (TRANSitive Reduction in WEighted Signed Digraphs), which is tailored to infer static, large-scale networks from high-throughput data. Within the scope of this thesis, the framework is developed for gene network reconstruction based on one-perturbation at a time data. Being flexible in its design, the reconstruction framework works also on genetical genomics data, where naturally occurring multifactorial perturbations (e.g. polymorphisms) in properly controlled and screened genetic crosses are used to elucidate causal relationships in gene regulatory networks. Although genetical genomics data contain rich information, a clear dissection of cause and effect is even harder to make compared to one-perturbation at

a time data. Still, as is shown in this work, the reconstruction framework performs very well on several different kinds of *in silico* and *in vitro* data sets. Following a simple yet effective paradigm, the framework has been awarded 1<sup>st</sup> and 3<sup>rd</sup> place at independent, international and highly competitive method assessment benchmarks. In this way, a simple yet effective approach is shown to outperform more complex methods with respect to (a) reconstruction quality (especially for small sample sizes) and (b) applicability to high-throughput data, which provides a powerful tool for genome-scale network reconstruction.

## Zusammenfassung

Die Verwendung von mathematischen Modellen für die Analyse komplexer biologischer Prozesse, einschließlich Metabolismus, Signaltransduktion und Genregulation in Säugerzellen oder Bakterien, ist ein leistungsfähiger Ansatz, um ein besseres Systemverständnis zu erhalten. Dieser Ansatz setzt jedoch das Vorhandensein von prädiktiven Modellen voraus. Um zu einem prädiktiven Modell zu gelangen, müssen während der Modellierung solcher komplexen Systeme zahlreiche Herausforderungen wie begrenztes Wissen über die zugrunde liegenden Mechanismen, Mangel an ausreichend zeitaufgelösten oder statischen Messdaten, sowie große experimentelle und biologische Variabilität bewältigt werden. Die vorliegende Arbeit liefert methodische Lösungen zur Modellidentifikation von (i) gewöhnlichen Differentialgleichungssystemen und (ii) großskaligen, biochemischen Interaktionsnetzwerken.

Die Lösung zu Teil (i) adressiert das Problem des modellgestützten, robusten Entwurfs von Zellstimuli, welcher verteilte Modellparameter berücksichtigt und optimale Daten zur Modelldiskriminierung liefern soll. Diese Art der Robustifizierung eines experimentellen Entwurfes ist insbesondere zu Beginn der Modellidentifikationsphase wichtig, da viele Modellparameter initial schlecht bestimmt sind, und deshalb häufig zu suboptimalen Experimenten führen. Der Entwurf von optimalen Stimulusprofilen unter Berücksichtigung von Modellparameterunsicherheiten mündet in ein numerisch rechenintensives, dynamisches Optimierungsproblem. Unter Verwendung von Sigma-punkten wird in dieser Arbeit eine numerisch stabile und effiziente Methodik zur dynamischen Optimierung entwickelt, welche den robusten Entwurf von diskriminierenden Stimulusprofilen erlaubt. Die entwickelte Methodik ist besonders für komplexe, stark nichtlineare mathematische Modelle mit verteilten Parametern geeignet.

Die Lösung zu Teil (ii) ist ein modularer, methodischer Identifikationsansatz namens TRANSWESD (TRANSitive Reduction in WEighted Signed Digraphs), welcher auf die Identifikation von großskaligen biochemischen Interaktionsnetzwerken basierend auf Hochdurchsatzdaten zugeschnitten ist. Die Methodik wird zu Beginn für die Rekonstruktion von genregulatorischen Netzwerken auf der Basis von Einfachperturbationsdaten verwendet. Aufgrund des modularen Konzeptes kann die Rekonstruktionsmethodik nach Anpassung auch auf genomische Daten mit multifaktoriellen Perturbationen

wie beispielsweise Polymorphismen angewandt werden. Obwohl genomische Daten reichhaltige Interaktionsinformationen enthalten, ist eine klare Abgrenzung von Ursache und Wirkung noch schwieriger als im Vergleich zu Einfachperturbationsdaten. Hier erweist sich die entwickelte Methodik als besonders effektiv. Der entwickelte Identifikationsansatz, welcher dem Paradigma *einfach jedoch effektiv* folgt, liefert für verschiedene *in silico* und *in vitro* Datensätze sehr gute Ergebnisse. Bei internationalen, hoch kompetitiven Rekonstruktionswettbewerben erlangte die Methodik mehrere Podiumsplätze. Dadurch konnte gezeigt werden, dass ein einfacher Ansatz komplizierte und rechenintensive Methoden in Bezug auf (a) Rekonstruktionsqualität und (b) Anwendbarkeit auf Hochdurchsatzdaten übertrifft. Damit stellt die entwickelte Rekonstruktionsmethodik ein leistungsfähiges Werkzeug für die Analyse von Hochdurchsatzdaten dar.

## Acknowledgements

In this thesis I present most of my work I have conducted from February 2009 - December 2013 as an employee at the Max Planck Institute for Dynamics of Complex Technical Systems and Otto-von-Guericke University Magdeburg, Germany.

I would like to express my gratitude to Prof. Dr.-Ing. Kai Sundmacher for the very challenging topic, the trust and scientific freedom in letting me explore and contribute to the research field of computational modeling in systems biology. I am very grateful for his support and demands on a wide range of activities, including publications, conference contributions, student supervisions, teaching, third-party fundraising, project reporting, project management as well as project conceptualization. I can say, that all these activities have contributed to a scientific sound and broad education. In addition, special thanks go to Dr.-Ing. Steffen Klamt, for getting me started on biological research, the fruitful discussions and pragmatic advices. I also want to thank Dr. rer. nat. Michael Wulkow for illuminating discussions and insights on modular programming strategies.

For taking the journey of wet-and-dry lab exchange I would like to thank Prof. Dr. rer. nat. Michael Naumann, Dr. rer. nat. Gunter Maubach and Christian Täger.

This work has also benefited from contributions of several students, which have conducted their study research, bachelor, master or Diploma thesis within my supervisions. Special thanks go to Katja Tummler for being a demanding and persistent student, who challenged me at the very beginning of my Ph.D. phase. Further I would like to thank Juliane Diedrich, Sandra Heise, Maxi Soldmann and Iryna Migal for their excellent work. It was a pleasure to work with all of you.

I further express my gratitude to Prof. Dr. rer. nat. Inna Lavrik and Prof. Dr. rer. nat. Wolfgang Wiechert for taking the peer-review of this work. For being a decent head of the Ph.D. defense colloquium, I would like to thank Prof. Dr.-Ing. Udo Reichl.

I very much value the time I could spend with my colleagues and friends from the Max Planck Institute for Dynamics of Complex Technical Systems and Otto-von-Guericke University Magdeburg. In particular I want

to mention Sascha Rollié, Andreas Voigt, Richard Hanke-Rauschenbach, Tanja Vidakovic-Koch, Peter Heidebrecht, Bianka Stein, Christian Borchert, Michael Fricke, Andreas Peschel, Britta Peschel, André Sommer, Benjamin Hentschel, Florian Karst, Astrid Bennsmann, Boris Bennsmann. I further would like to thank the coffee group of the MPI south wing S.3.18 and neighbors for nice post lunch discussions, for looking after my caffeine level but also my correct backbone posture. To this group I count René Schenkendorf, Anke Ryll, Bernhard Kramer, Andre Franz, Katharina Holstein, Michael Mangold and Phillip Erdrich, who together with Kristin Pirwitz I would also like to thank for getting me back into shape in SH3/K. Also I express my gratitude to Melanie Fachel for filling the gap of my initial rudimentary biological understandings.

Substantial social, financial and educational life support came from my parents Christine and Michael Flassig, which I acknowledge hereby. Also not to forget about my dear brother Peter Flassig, who always sets the bar high. Thanks for pushing. As always, I am grateful for every minute we spend together!

There are many other personalities and materials, including the coffee farmers in Central America, the paper from spruces and pines, the German tax payers and the proofreader Ina Michael that have contributed to this work. Thank you all.

Finally my dear darling Sara, thank you for your love and giving us such a good time with Alwin, Hedda and Almuth.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Aim of this work . . . . .	1
1.2	Thesis Guide . . . . .	3
<b>2</b>	<b>Methods for identifying dynamic models of biochemical reaction systems</b>	<b>7</b>
2.1	Dynamic modeling of biochemical reaction systems . . . . .	7
2.2	Parameter inference . . . . .	9
2.2.1	Maximum likelihood . . . . .	10
2.2.2	Confidence intervals . . . . .	12
2.2.3	Parameter identifiability . . . . .	16
2.3	Model discrimination . . . . .	18
2.3.1	Model distinguishability . . . . .	18
2.3.2	Model selection . . . . .	19
2.4	Optimal experimental design . . . . .	22
2.4.1	Working definition: Optimal experimental design . . . . .	22
2.4.2	Experimental design for optimal parameter estimation . . . . .	24
2.4.3	Experimental design for optimal model discrimination . . . . .	24
2.4.4	Robust optimal experimental design . . . . .	25
2.5	Summary . . . . .	27
<b>3</b>	<b>Optimal experimental design in the presence of distributed model parameters</b>	<b>29</b>
3.1	Model overlap as a robust discrimination criterion . . . . .	29
3.2	Estimation of nonlinear PDF mapping . . . . .	31
3.2.1	Estimation based on linearization . . . . .	32
3.2.2	Estimation based on sigma points . . . . .	33
3.3	Robust optimal stimulus design . . . . .	34
3.4	<i>In silico</i> results . . . . .	35
3.4.1	Benchmark using a signaling cascade . . . . .	35
3.4.2	Bistable system . . . . .	38
3.4.3	Summary robust optimal experimental design methodology . . . . .	43

## CONTENTS

---

3.5	A real life application . . . . .	45
3.5.1	Background of the <i>in vitro</i> application . . . . .	45
3.5.2	Model identification . . . . .	46
3.5.3	Model identifiability analysis . . . . .	50
3.5.4	Model predictions . . . . .	55
3.5.5	Discussion of <i>in vitro</i> application . . . . .	57
3.6	Summary optimal experimental design . . . . .	57
<b>4</b>	<b>Methods for identifying structural models of biochemical reaction systems</b>	<b>59</b>
4.1	What are gene regulatory networks? . . . . .	60
4.1.1	Definition of a gene regulatory network . . . . .	60
4.1.2	Reconstructing a gene regulatory network . . . . .	62
4.2	Reconstruction data . . . . .	63
4.3	Methodological approaches for network reconstruction . . . . .	63
4.3.1	Data preprocessing . . . . .	64
4.3.2	Reconstruction methods . . . . .	64
4.4	Summary . . . . .	66
<b>5</b>	<b>TRANSWESD: A reverse engineering algorithm for identifying gene regulatory networks</b>	<b>67</b>
5.1	TRANSWESD . . . . .	67
5.1.1	General workflow of TRANSWESD for one-perturbation at a time data . . . . .	70
5.1.2	Perturbation graph . . . . .	71
5.1.2.1	Single perturbation data . . . . .	71
5.1.3	Identifying significant edges . . . . .	72
5.1.4	Weight association to the edges . . . . .	73
5.1.5	Removing indirect edges: TRANSWESD . . . . .	73
5.1.5.1	Transitive reduction in signed acyclic graphs . . . . .	73
5.1.5.2	Transitive reduction in signed and weighted acyclic graphs . . . . .	75
5.1.5.3	Transitive reduction in signed and weighted cyclic graph . . . . .	76
5.1.6	<i>In silico</i> application: DREAM4 challenge . . . . .	77
5.1.7	Summary TRANSWESD on one-perturbation at a time data . . . . .	80
5.2	Systems genetics . . . . .	81
5.2.1	GRN reconstruction on genetical genomics data . . . . .	82
5.2.1.1	Preprocessing genetical genomics data . . . . .	82
5.2.1.2	Generating the raw perturbation graph from genetical genomics data . . . . .	84
5.2.1.3	Identification of eQTLs and candidate regulator selection . . . . .	85
5.2.1.4	Identifying and removing indirect effects (TRANSWESD) . . . . .	86
5.2.1.5	Sorted edge list . . . . .	86
5.2.2	Applications . . . . .	86

5.2.2.1	<i>In silico</i> application: DREAM5 challenge . . . . .	86
5.2.2.2	<i>In vitro</i> application: genetical genomics data of yeast . . . . .	89
5.2.3	Summary TRANSWESD on systems genetics data . . . . .	91
5.3	Summary . . . . .	93
<b>6</b>	<b>Concluding Remarks</b>	<b>95</b>
6.1	Summary . . . . .	95
6.2	Conclusion and Outlook . . . . .	97
<b>A</b>	<b>Supplementary methodological information</b>	<b>99</b>
A.1	NLP problem formulation for optimal experimental stimulus design . . . . .	99
A.1.1	Direct sequential approach . . . . .	99
A.1.2	Direct simultaneous approach . . . . .	100
A.2	Transformation of log-normal to normal PDF . . . . .	100
A.3	Supplementary <i>in vitro</i> application information . . . . .	102
A.3.1	Model equations . . . . .	102
A.3.2	Parameter Inference . . . . .	103
A.3.3	Profile Likelihood Analysis . . . . .	104
	<b>List of Figures</b>	<b>127</b>
	<b>List of Tables</b>	<b>129</b>

## GLOSSARY

---

# Glossary

## Roman Symbols

$A$	gene-to-marker association	$l_n$	log-likelihood for n samples
$\mathbf{a}$	a vector	$L_2$	space of square-integrable functions
$\mathbf{C}$	variance-covariance matrix	$LR$	likelihood ratio statistics
$\mathbf{C}_{\log}$	estimated variance-covariance matrix of log-normal PDF	$\mathbf{M}$	statistical moment
$\mathbf{D}$	experimental design	$m$	marker
$D_1$	inter-pulse time	$m$	metabolite
$D_2$	(second) pulse dose	$n_i$	integer, used to indicate an amount for samples, parameters, readouts, stimuli etc., which are indicated by the subscript i
$d_{\min}$	threshold for minimal genotypic correlation	$O$	design objective
$\mathbf{E}$	estimated expectation of some PDF	$\mathbf{P}_{\theta}$	probability of the realization $\theta$
$\mathbf{E}_{\log}$	estimated expectation of log-normal PDF	$p$	protein
$E$	set of edges	$Q$	indicates the genotype of a gene/marker
$e$	edge	$r$	Pearson correlation coefficient
$e_M$	relative MSE of moment estimate $\mathbf{M}$	$r^{Q_i T_j}$	genotype-phenotype Pearson correlation coefficient
$F_{df1,df2}$	F distribution with df1,df2 degrees of freedom, also $F$ test statistics	$\mathbf{S}^{+/-}$	sign label matrix
$\mathbf{f}$	right hand side function of an ODE	$s$	sign (+,-)
$G$	graph	$s_M(u(t))$	model $M$ for an input function
$\mathbf{g}$	readout function	$s_i$	scaling parameters in the modeling application
$g$	gene	$\mathbf{S}$	sensitivity matrix
$H$	unit step function	$\mathbf{t}$	time point vector
$\mathbf{h}$	mapping function	$T$	T-criterion for model discrimination
$\mathbf{J}$	Jacobi matrix	$T$	indicates the expression phenotype (etrait) of a gene/marker
$L$	linkage map	$t$	time
$L_n$	likelihood for n samples	$t^{QT}$	threshold for genotype-phenotype Pearson correlation coefficient
		$\mathbf{U}$	stimulus parameter vector
		$\mathbf{u}$	vector of stimulus profile(s)
		$V$	set of vertices
		$w_{ij}$	edge weight
		$\mathbf{W}$	collocation point weighting matrix
		$\mathbf{X}, \mathbf{Y}$	vectors of random variables
		$\mathbf{x}$	vector of system states, can represent a realization of the random variable $\mathbf{X}$

## GLOSSARY

---

$\mathbf{y}$	vector of system readouts, can represent a realization of the random variable $\mathbf{Y}$	$\chi^2$	$\chi^2$ distribution, also $\chi^2$ test statistics
		$\chi_B^2$	Bartlett's $\chi^2$ test statistics
		$\chi_n^2$	residual sum of squares, $n$ indicates the number of samples
<b>Greek Symbols</b>			
$\alpha$	significance level	$\psi$	tuning parameter to control overall association strength of a path
$\beta$	tuning parameter sigma points		
$\delta\theta$	relative parameter change		
$\delta\chi_{df,\alpha}^2$	$\chi^2$ distribution quantile for given degrees of freedom and significance level $\alpha$		
$\Delta_{ij}$	variation measure for a node pair $(i, j)$		
$\delta[\cdot]$	generalized Dirac delta distribution		
$\epsilon$	measurement noise as a random variable		
$\epsilon$	measurement noise realization of $\epsilon$		
$\zeta$	tuning parameter sigma points		
$\eta$	standard deviation scaling factor		
$\Gamma$	weight mapping		
$\theta$	vector of kinetic model parameters, may include readout parameters as well and can also indicate a realization of $\Theta$		
$\Theta$	vector of random variable		
$\kappa$	tuning parameter sigma points		
$\lambda$	lumped tuning parameter sigma points		
$\mu$	vector of true mean		
$\mu_i$	set of markers linked to marker $m_i$		
$\xi$	constant or protein ratio		
$\pi^i$	indicator of perturbation direction of node $i$ , up $\pi^i = +1$ , down $\pi^i = -1$		
$\rho_{\mathbf{X}}(\mathbf{x})$	PDF of $\mathbf{X}$ with realization $\mathbf{x}$		
$\Sigma$	true variance-covariance		
$\Sigma_\epsilon$	diagonal variance-covariance matrix of the measurement noise		
$\Phi$	average overlap of $\Phi(t)$		
$\phi$	sign mapping		
$\Phi(t)$	general overlap as a function of $t$		
			<b>Superscripts</b>
		$\dagger$	optimal
		*	indicates a true value
		$p$	positive, real-valued weighting factor in $\Phi$ , not to be confused with a p-value
		T	transpose
			<b>Subscripts</b>
		0	initial value
		$f$	final value
		exp	experiment
		$i, j, k, l, q$	enumerates a quantity
		PL	profile likelihood
		sim	simulation
		$s, k, t$	sign labels
		TR	transitive reduction
			<b>Other Symbols</b>
		$\hat{\cdot}$	estimated value of $\cdot$
		$\tilde{\cdot}$	indicates a different entity of $\cdot$
		$\mathbb{A}_i$	generic set as a subset of the real valued vector space $\mathbb{R}^{n_i}$
		$\mathbb{D}$	experimental design region
		$\mathcal{E}$	PDF estimation method
		$\mathcal{F}$	parameterization of error variance
		$\mathcal{L}$	linearization estimate
		$\mathcal{M}$	place holder for a generic model, also used as model index
		MC	Monte Carlo estimate
		$\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$	multivariate Normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{C}$
		$\mathbb{R}^+$	one-dimensional vector space over the field of positive real numbers

$\mathbb{R}^n$	n-dimensional vector space over the field of real numbers	eQTL	expression-QTL
$\langle S \rangle$	mean model prediction variance-entropy	<b>FIM</b>	Fisher information matrix
$\mathcal{S}$	sigma point estimate	FN	false negative
$\mathbb{T}$	time point design region	FP	false positive
$\mathbb{U}$	stimulus design region	GMD	gaussian mixture density
$\langle V \rangle$	mean model prediction variance	GRN	gene regulatory network
$\mathbb{Y}$	readout design region	$\gamma$ H2AX	phosphorylated H2AX at serine 139
<b>Abbreviations</b>		H2AX	variant of histone H2A
AD	Anderson-Darling	HR	homologous recombination repair
AIC	Akaike's information criterion	IR	ionizing irradiation
AIL	advanced intercross line	$\max(\cdot)$	maximum operator
aNHEJ	alternative non-homologous end joining	MC	Monte Carlo
ANOVA	analysis of variance	MCMC	Markov chain Monte Carlo
ATM	protein kinase ataxia telangiectasia mutated	MLE	maximum likelihood estimator
ATM-P	phosphorylated protein kinase ataxia telangiectasia mutated	MSE	mean squared error
ATR	ataxia telangiectasia and Rad3-related	ODE	ordinary differential equation
AUPR	area under the precision-recall curve	PIKK	phosphoinositide 3-kinase like kinase family
AUROC	area under the receiver operating characteristics	p53	tumor suppressor protein
CMA-ES	Covariance Matrix Adaptation Evolutionary Strategy	p53-P	phosphorylated tumor suppressor protein
cNHEJ	classical non-homologous end joining	PDF	probability distribution/density function
<b>COV</b>	covariance terms of <b>C</b>	<i>prec</i>	precision, measure of fidelity of a classification
$CR_\alpha$	confidence region at significance $\alpha$	QTL	quantitative trait locus
$cr_\alpha$	mapping into a confidence region at significance $\alpha$	RCS	recombinant congenic strain
$CR_{i,\alpha}$	confidence interval at significance $\alpha$	<i>rec</i>	recall, measure of completeness of a classification
$cr_{i,\alpha}$	mapping into a confidence interval at significance $\alpha$	RIL	recombinant inbred line
CSS	chromosome substitution strain	SNP	single nucleotide polymorphisms
DDR	DNA damage response	SP	sigma point
DNA-PK <sub>cs</sub>	DNA-dependent protein kinase, catalytic subunit	STD[·]	estimated standard deviation of some PDF
DNA-PK <sub>cs</sub> -P	phosphorylated DNA-dependent protein kinase, catalytic subunit	TN	true negative
DSB	double strand break	TP	true positive
		TRANSWESD	TRANSitive Reduction in WEighted Signed Digraphs
		<b>VAR</b>	variance terms of <b>C</b>

## GLOSSARY

---



# 1

## Introduction

### 1.1 Aim of this work

Nature with all its secrets has fascinated human perception since the beginning of night and day. Secrets that lie ahead of, between and beyond ameba and man have been, still are and will always be the source of our wondering. How poor is a world without secrets? Yet the journey of understanding has brought us a long way on diverse fields of investigations. This thesis humbly tries to support the understanding of living systems by contributing methodological solutions to problems arising in mathematical model identification.

A mathematical or computational model is a generic, well-established tool for analyzing complex natural processes. Often, these processes are based on many elementary interaction mechanisms where complex phenomena emerge from coupling of such to form a large interaction network. Such emergence can often only be understood with the help of mathematical models. As for instance in climate research, mathematical models of climate shaping processes allow simulating the dynamics of certain model states (associated to real world entities) to, for instance, judge the influence of the trend winds on the Baltic Sea level via global warming effects. In this way, system states that are not directly accessible at any place and time (e.g. salinity or temperature of the oceans) can be visualized and understood. Mathematical models of the earth's climate are further used to interpret its past based on ice core or sedimentary data. Model predictions on different scenarios of control policies serve to find best action choices for policy makers for reaching a predefined target, e.g. the 2° C goal (?).

Moving on to the focus of this work, that is model-based analysis and design of biochemical systems, we are facing challenges that arise from emergent biological complexity, from often only vaguely known elementary biological interaction mechanisms paired with inherent biological variability and tricky experimental procedures. Still, the strength of combining mathematical models and biology to an interdisciplinary and integrative research approach - often referred to as *systems biology* - lies in synergist effects for a better understanding of biological functioning as a result of more or less

## 1. INTRODUCTION

---

effortless *in silico* experiments via simulation and prediction of system states. In this way, biological states, which are hard to measure *in vitro*, can be simulated for arbitrary biological scenarios, thus allowing analysis and design of targeted manipulation strategies. To name a few examples: The urge of new scientific insights to understand function and dysfunction of biological processes in a world with an aging population, to ultimately provide biomedical solutions for advanced model-based understanding of diabetes. Models of diabetes that account for patient specific factors bear a huge potential for designing individualized insulin therapies, reducing unwanted side effects (???). Thinking bigger and beyond engineering a personalized drug or therapy, understanding of biochemical systems provides the basis for rational design of biochemical production processes. Driven by the shift from fossil to renewable biomass feedstock, the emerging economy of microbial production is in the need of sophisticated tools for engineering efficient and sustainable technologies that transform biomass into chemicals, material and electricity (?). Although microorganisms can generate an amazingly diverse plethora of valuable products interesting for pharmaceutical and industrial production, they often come at very low yield. Here, a mathematical model of the metabolism of a microbial organism (e.g. *Saccharomyces cerevisiae* and *Escherichia coli*), can serve as a basis for engineering suitable variants of this organism (=strain design) that has an enhanced product yield or altered product spectrum (?). In this way, old-fashioned biotechnology is transformed into a biochemical engineering approach, which builds on rational biochemical systems design for innovative, resource-efficient solutions to a sustainable future of our globe. The approach of systems biology is thus inverted, from understanding biological functioning to intelligent redesign and creation of new biology.

Despite its power, model-based analysis is in the need of a realistic model, i.e. a model that adequately describes data of the true system. This is where the challenge of any model-based approach lies - the generation of a predictive (=realistic) model. Biological systems are inherently diverse and complex. Parametric models thereof are thus often nonlinear mathematical expressions, which typically outnumber the amount of data to a huge extent. This is also referred to as the curse of dimensionality (?) and most often leads to parameter identifiability problems. Additional to identifiability problems, natural biologic variability, experimental variations and limitations directly result in distributed parameters, which must be accounted for to not render model predictions questionable. Therefore, sophisticated theoretical tools are needed for generating and analyzing mathematical models of biological processes. Otherwise it is virtually pointless to put biology into equations. In Fig. 1.1 the engineering view of a biological system (here thought of as a cell) is presented, which is commonly taken in computational biology and in this thesis as well. Most often, biological systems are abstracted in terms of biochemical interaction networks, neglecting physical processes like diffusion or convection when it comes to modeling. From a high level point of view, the cell is viewed as a processing unit, taking inputs, which are processed and converted into proper cellular responses, including signals, products and phenotypes (s. Fig. 1.1).

## 1.2 Thesis Guide

The thesis is centered around two research foci, (i) optimized data generation including model-based experimental design, and (ii) high-throughput data analysis. Research focus (i) addresses the following questions that arise during model identification supported by experimental design:

1. Given data and a set of competing plausible models, which one to select for model-based analysis, prediction and design?
2. How to design stimulus experiments, which generate optimal data with respect to model selection, even though model parameters are highly uncertain?

Within research focus (ii) the following questions are addressed:

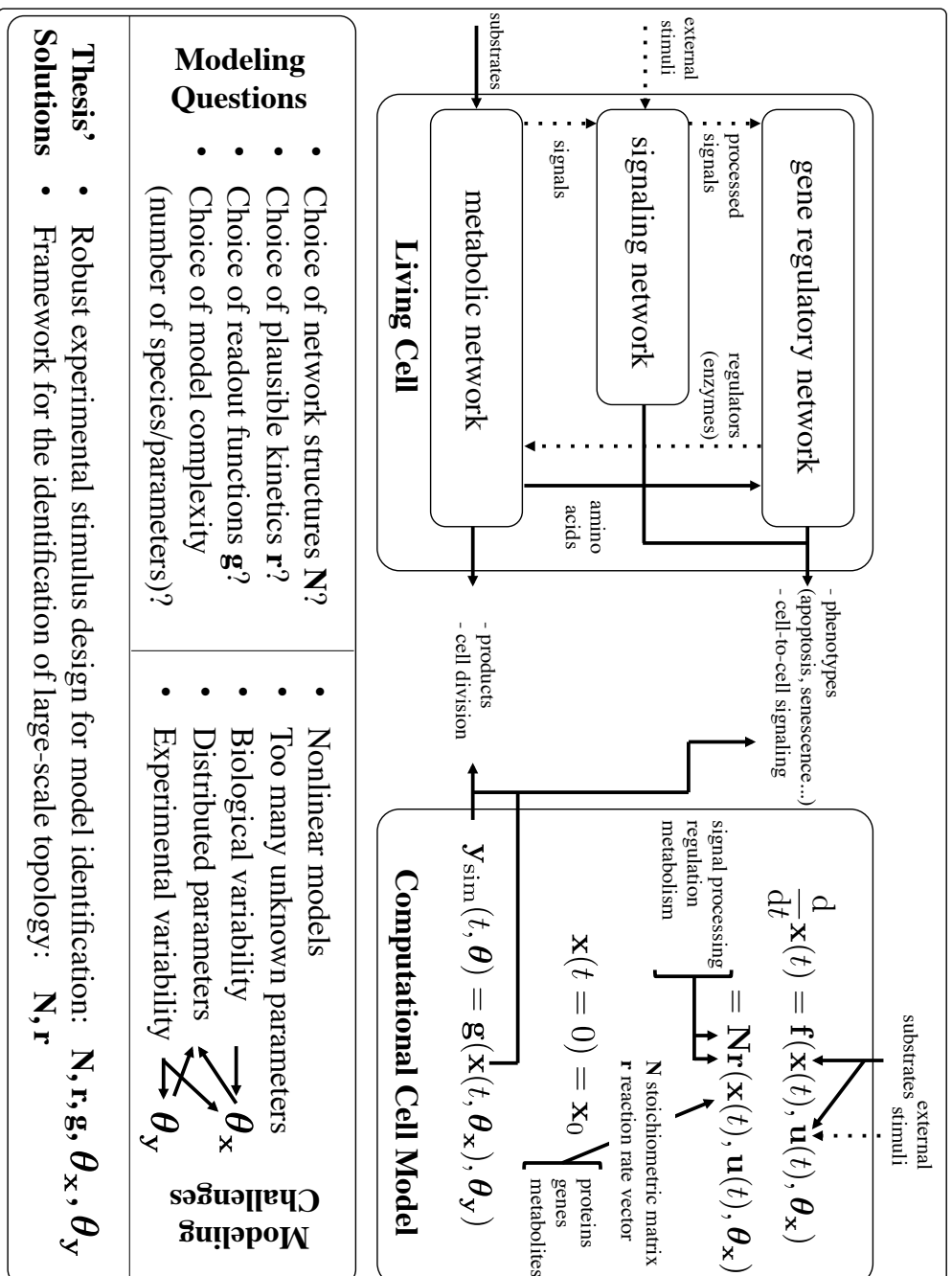
1. Given high-throughput data, how to derive an interaction structure?
2. How to deal with the problem of many interactions between biochemical players but few data samples, i.e.  $n_{\text{network nodes}} \gg n_{\text{samples}}$ ?

Within the scope of (i) a new experimental stimulus design approach is presented, which optimizes experiments and thus data for identifying complex, nonlinear dynamic models in the form of ordinary differential equations (ODEs) describing dynamic biological processes. The advantage of this approach is a numerically efficient consideration of distributed model parameters, which allows robustifying the experimental design. Regarding research focus (ii), a methodology for identifying large-scale structures of biochemical reaction networks for given high-throughput data is presented. This methodology provides a solution to the curse of dimensionality, i.e. few data but many parameters. Ultimately, both research foci provide solutions to biological model identification. In the lower part of Fig. 1.1, the modeling challenges that motivated the contributions of this thesis are put into perspective of the biological system abstraction.

According to the research foci, the thesis is structured into two parts. Part (i) starts with a recall of statistical methods for identifying computational models of complex biological system (Ch. 2). The recall includes a brief discussion on dynamic modeling based on ODEs in the light of distributed determinism, parameter inference, parameter identifiability analysis and model selection. A sound basis of inference methods is important given the setting of few data but many parameters. Here, parameter identifiability analysis has proven to be an important tool to assess the predictive power of a computational model. The methodological introduction is intended to provide the fundamentals of statistical ODE model identification for complex, nonlinear systems. With the methodological fundamentals in mind, the concepts of optimal experimental design (OED) are presented in the last part of Ch. 2.

In Ch. 3 a methodology for designing an experiment aimed at model discrimination in the presence of distributed model parameters is introduced. The method is benchmarked on *in silico* data. The chapter, and thus part (i) of the thesis, is closed

# 1. INTRODUCTION



**Figure 1.1:** The thesis in a picture. The upper part depicts the mathematical abstraction of biological systems. Biological systems are thought to represent processing units, that transform inputs (e.g. stimuli, substrates) into proper outputs (e.g. products, phenotypes). Dashed lines indicate flow of information, whereas solid lines indicate flow of material (which of course can also be interpreted as information). The lower part indicates motivations and contributions of this thesis. Partially adapted from ?.

by illustrating the entire process of model identification, including the application of the robust OED method. Within this application the identified dynamic model that describes DNA damage detection signaling upon ionizing irradiation is analyzed and verifiable predictions are given.

Part (ii) of the thesis starts with a methodological survey on large-scale reconstruction of biochemical reaction networks (Ch. 4). The focus is put on gene regulatory network and it is briefly discussed how to interpret such interaction graphs.

In Ch. 5 TRANSWESD (TRANSitive Reduction in WEighted Signed Digraphs), a methodology for reconstructing biochemical reaction networks is presented. Applications of TRANSWESD to *in silico* and *in vitro* data sets are also presented. Finally, in Ch. 6 the achievements of the thesis are summarized and an outlook on further developments is given. Used abbreviations are always explained in the text, but can also be found in the glossary. Further, bold letters are used to indicate vectors and matrices.

## 1. INTRODUCTION

---

## 2

# Methods for identifying dynamic models of biochemical reaction systems

*From step to jump, from jump to flight.*

---

Otto Lilienthal's<sup>†1896</sup> strategy to become pioneer in gliding flights.

In this chapter a survey on essential methods used in modeling dynamic systems with ordinary differential equations is given. It should serve as a hands-on guide for beginners new to the field of modeling dynamic biological systems and is thus not intended as a comprehensive review on dynamic modeling. The presentation focuses on important, well-established but also recent approaches, which are indispensable for inferring predictive, dynamic models of biological systems. Surfacing challenges for nonlinear model identifications are discussed and recent solutions to some of these challenges are presented. The experienced reader may skip this chapter and move directly to the next one, where a methodological contribution within the field of optimal experimental design is presented. Comprehensive presentations on classical model identification topics can be found in excellent textbooks including ?????.

## 2.1 Dynamic modeling of biochemical reaction systems

This section is partially taken and adapted from Secs. 2.1-2.2 of ?.

Ordinary differential equations provide the modeling basis to describe the dynamics of biochemical reaction networks. The dynamics of the internal states  $\mathbf{x}(t, \mathbf{u}(t), \boldsymbol{\theta}_x) \in \mathbb{A}_{\mathbf{x}} \subset \mathbb{R}^{n_x}$ , e.g. protein concentrations, is thus determined by the solution of an initial

## 2. METHODS FOR IDENTIFYING DYNAMIC MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

value problem of the form

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\theta}_x), \quad (2.1)$$

with initial system states  $\mathbf{x}(t_0) = \mathbf{x}_0$  and right hand side function  $\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\theta}_x)$  describing biological interaction mechanisms, which depends on the system states  $\mathbf{x}(t)$ , (multiple) inputs  $\mathbf{u}(t)$  (=stimulus) and kinetic parameter set  $\boldsymbol{\theta}_x$ . Assuming  $\mathbf{f}$  to be Lipschitz in  $\mathbf{x}(t)$ ,  $\mathbf{u}(t)$  and continuous in  $t$ , the readout variables are determined by

$$\mathbf{y}(t, \boldsymbol{\theta}) = \mathbf{g}(\mathbf{x}(t, \boldsymbol{\theta}_x), \boldsymbol{\theta}_y), \quad (2.2)$$

where the function  $\mathbf{g}$  - assumed to be sufficiently smooth - relates the internal system states to the readouts of the experiment with corresponding readout parameters  $\boldsymbol{\theta}_y$ , which together with dynamic parameters and initial conditions are merged into the model parameter vector  $\boldsymbol{\theta} = [\boldsymbol{\theta}_x, \boldsymbol{\theta}_y]^T$ , with redefined dynamic parameter vector  $\boldsymbol{\theta}_x \equiv [\boldsymbol{\theta}_x, \mathbf{x}_0]^T$ . The dynamic model defined by Eqs. (2.1,2.2) can be understood as a time-dependent mapping from the model parameter space  $\mathbb{A}_{\boldsymbol{\theta}_x} \times \mathbb{A}_{\boldsymbol{\theta}_y} = \mathbb{A}_{\boldsymbol{\theta}} \subset \mathbb{R}^{n_\theta}$  to the model output space  $\mathbb{A}_y \subset \mathbb{R}^{n_y}$ ,

$$\mathbf{h} : \mathbb{R} \times \mathbb{A}_{\boldsymbol{\theta}} \rightarrow \mathbb{A}_y \quad (2.3)$$

$$(t, \boldsymbol{\theta}) \mapsto \mathbf{h}(t, \boldsymbol{\theta}) = \mathbf{y}(t, \boldsymbol{\theta}). \quad (2.4)$$

Although biological systems might follow deterministic rules, repeated measurements, even though with very accurate measurement techniques, will yield different results. The reasons for that are manifold. Additionally to unavoidable measurement errors, biologic variability, i.e. systems with intrinsically distributed parameters, can induce a large spread in the transient dynamics and stationary behavior. In case of the existence of multiple steady states this spreading effect might even be more pronounced. Varying parameters during the measuring procedure and local parameter perturbations by non-stationary noise also contribute to a distributed measurement signal, (??). Complex, nonlinear models of biological systems might also behave chaotic, further contributing to distributed response measurements. Thus, the conventional sharp and deterministic system representation needs to be extended by the notion of distributed determinism, i.e. although the system might completely be deterministic, its perceived signals are distributed realizations of the underlying deterministic mechanisms. This can be done by considering the parameters, and hence, internal states and model responses as random variables  $\boldsymbol{\Theta}$ ,  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively, each characterized by a probability distribution function (PDF). Within this interpretation, the system and hence the model is assumed to naturally possess distributed parameters. Consequently, a distributed response is not solely explained by additive measurement noise and limited quality of the data but also by other sources of variations, which may be represented by distributed model parameter sets. Measurement noise is then also described by a distributed readout parameter, and therefore not explicitly stated in Eq. (2.2).



Let the model parameters be distributed according to some well-defined PDF  $\rho_{\Theta}(\boldsymbol{\theta})$ , with  $\boldsymbol{\theta} \in \mathbb{A}_{\Theta}$  being a realization of  $\Theta$ . The PDF of the random model response  $\mathbf{Y}$  at time  $t$  can be derived from the normalized integral over all possible parameter and corresponding response realizations, weighted with the parameter PDF, i.e.

$$\rho_{\mathbf{Y}}(\mathbf{y}, t) = \frac{\varrho_{\mathbf{Y}}(\mathbf{y}, t)}{\|\varrho_{\mathbf{Y}}(\mathbf{y}, t)\|_1}, \quad (2.5)$$

with

$$\varrho_{\mathbf{Y}}(\mathbf{y}, t) = \int_{\mathbb{A}_{\Theta}} \boldsymbol{\delta}[\mathbf{h}(t, \boldsymbol{\theta}) - \mathbf{y}] \rho_{\Theta}(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (2.6)$$

where  $\boldsymbol{\delta}[\cdot]$  represents the *generalized Dirac delta distribution* (?). The normalization employs the  $L_1$ -norm with respect to  $\mathbf{y} \in \mathbb{A}_{\mathbf{Y}}$ . Note that  $\mathbf{y}$  represents an arbitrary realization of  $\mathbf{Y}$  in  $\mathbb{A}_{\mathbf{Y}}$ , whereas  $\mathbf{h}(t, \boldsymbol{\theta})$  describes the model response at time  $t$  for fixed stimulus time course(s)  $\mathbf{u}(t_0 \rightarrow t)$  and given parameter realization  $\boldsymbol{\theta}$ . Consequently, for every single point in time, the shape of Eq. (2.5) is determined by the parameter PDF, choice of model, Eqs. (2.1,2.2) and stimulus time course. This fact can be used to ground experimental design on the model response PDF. In doing so, such an experimental design is robustified accounting for variabilities in the parameters and model specific mapping of the parameter PDF to the response space in terms of model response PDF. A key challenge within this robust approach is the derivation of the response PDF. In Chapter 3 linearization and sigma point methods are introduced within the scope of robust discriminative stimulus design and compared against Monte Carlo simulations. As will be seen, all three methods represent different approaches for estimating Eq. (2.5) in the case of non-closed form expressions of the integral in Eq. (2.6).

## 2.2 Parameter inference

Parameter inference - sometimes referred to as model calibration - is a necessary step for model identification and model-based prediction. In the model mapping interpretation of the previous section, Eq. (2.4), parameter inference is the inversion of such, subjected to optimality constraints. Thus, in line with what has been pointed out in ?, parameter inference represents a prediction of  $\Theta$  given model and data  $\mathbf{Y}$ , which has to be quantified regarding predictive power, e.g. in terms of confidence intervals. In contrast to parameter estimation of linear models, calibrating ODE models of biochemical reaction networks is a delicate task. This is on the one hand due to the inherently nonlinear character of this model class. For instance, even the solution  $x(t)$  to the simplest ODE model describing a protein degradation process via  $\frac{dx}{dt} = -\theta x$  is highly nonlinear with respect to the parameter  $\theta$ . On the other hand, experimental data tend to be scarce in comparison to the number of model parameters including initial conditions, reaction rate constants and scaling parameters, which is typical of the order of 10 to 100 for small models of for instance cell signaling. On genome scale, one is very quickly well

## 2. METHODS FOR IDENTIFYING DYNAMIC MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

above orders of  $\mathcal{O}(10^4)$ , and until now, dynamic modeling on this scale seems hopeless, in contrast to structural approaches (Ch. 5).

Classical approaches for quantifying the quality of the parameter estimates in form of confidence intervals may work on ODE models, but often fail owing to the requirement of linearity in combination with small sample sizes and high dimensional parameter spaces. It is thus of utmost importance to have sophisticated parameter estimation methods at hand, which efficiently quantify uncertainties of parameter estimates and reversely allow a numerically feasible mapping of uncertainties from the parameter to the model response space.

ODE models of dynamic biological processes can be regarded as meta-mechanistic. Besides describing direct physical interactions, they most often track flow of information. As an example: although posttranslational modifications of proteins can be measured, the direct physical interactions and potential mediators are often unknown. As a consequence, sheer parameter values of (lumped) kinetics are not that much of interest. What is more interesting is whether the parameters of a given model are identifiable. Because then, model predictions can serve the purpose of supporting experimental analysis by looking deeper into the system's details. Of course under the premise that the model structure is correct, which can be partially explored by discriminating amongst plausible models.

In the following, approaches for parameter estimation, quantifying parameter uncertainties and identifiability owing to limited information from measurements are discussed. Intrinsic parameter variability is excluded for ease of presentation. Noting that intrinsically distributed parameters may be thought of as hyper-parameters (=parameter PDFs), these concepts can also be applied to intrinsically distributed parameters.

### 2.2.1 Maximum likelihood

This subsection follows the presentation of ? to outline the maximum likelihood approach in general, which will be of later use in this chapter. For convenience, a single response at one time point as well as a one-dimensional parameter space is discussed, extension of this discussion to the multidimensional case (several readouts, time points and parameters) is an exercise of indexing, but straightforward. Given samples  $y_1, \dots, y_n$  (=measurements including repetitions) of a studied system and some parametric model  $\mathcal{M} = \{\rho_Y(y; \theta), \theta \in \mathbb{A}_\Theta\}$  that is assumed to describe the studied system, maximum likelihood seeks for the model parameter which most likely generated the given samples. For this purpose, the likelihood

$$L_n(\theta) = \prod_{i=1}^n \rho_Y(y_i; \theta), \quad (2.7)$$

or log-likelihood

$$l_n(\theta) = \log(L_n(\theta)) \quad (2.8)$$

is maximized with respect to the parameters. As can be seen in Eq. (2.7), the likelihood is the joint probability of the data, but it is treated as a function of the model parameter  $L_n : \theta \rightarrow [0, \infty)$ . In general,  $L_n(\theta)$  is not normalized with respect to  $\theta$ , i.e.  $\int_{\mathbb{A}_\Theta} L_n(\theta) d\theta \stackrel{i.g.}{\neq} 1$ . The maximum likelihood estimator (MLE)  $\hat{\theta}_n$  is the parameter value that maximizes the likelihood. Since the logarithm is a monotonic function, the MLE of the likelihood is equivalent to the MLE of the log-likelihood. Working with the log-likelihood is convenient in the case of  $y_1, \dots, y_n$  being normally distributed. Then, the log-likelihood  $l_n(\theta)$  is proportional to the standardized residual sum of squares  $\chi_n^2(\theta)$ , and least-squares estimation is equivalent to maximum likelihood estimation. In detail, if  $y_1, \dots, y_n \propto N(\mu_Y, \sigma_Y^2)$  and  $\mathcal{M} = \{\rho_Y(y; \theta) = N(\mu, \sigma_Y^2), \theta = \mu \in \Theta\}$ , i.e. the exact PDF of the sample generating system as well as the variance is known, the likelihood and log-likelihood read

$$L_n(\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma_Y^2}\right), \quad (2.9)$$

$$l_n(\mu) = \sum_{i=1}^n -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_Y^2) - \frac{(y_i - \mu)^2}{2\sigma_Y^2}, \quad (2.10)$$

which can be related to the residual sum of squares

$$\chi_n^2(\mu) = \sum_{i=1}^n \frac{(y_i - \mu)^2}{\sigma_Y^2} = \text{const.} - 2l_n(\mu). \quad (2.11)$$

Only the last term depends on  $\mu$ , and therefore minimizing  $\chi_n^2(\mu)$  with respect to  $\mu$  is equivalent to maximizing  $l_n(\mu)$ . The MLE estimator properties are consistency, functional invariance and asymptotic normality. For details see ?.

MLEs of ODE model parameters are most commonly obtained by assuming that the error on the data  $y_{i, \text{exp}}$  is normally distributed with known error variance  $\sigma_{i, \text{exp}}^2$ . Then, the MLE is obtained by minimizing the residual sum of squares

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{A}_\Theta} \chi^2(\theta) \quad (2.12)$$

$$= \arg \min_{\theta \in \mathbb{A}_\Theta} \sum_{i=1}^n \frac{(y_{i, \text{exp}} - y_{i, \text{sim}}(\theta))^2}{\sigma_{i, \text{exp}}^2}, \quad (2.13)$$

where the subscript for the number of total measurements  $n$  of the MLE is dropped. In case of unknown error variance, one has to maximize  $l(\tilde{\theta})$ , with  $\tilde{\theta} = (\theta, \sigma_1, \dots, \sigma_n)$ , where the error variance can be estimated from a parameterized error model  $\sigma_{i, \text{exp}} = \mathcal{F}(y_i(\theta), \theta_{\mathcal{F}})$ . Minimization of the residual sum of squares (Eq. (2.13)) is the most frequently used objective in parameter estimation.

## 2. METHODS FOR IDENTIFYING DYNAMIC MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

### 2.2.2 Confidence intervals

Confidence intervals or, more generally, regions are used to quantify predictions (e.g. parameters) with respect to type 1 errors at a given confidence level giving a feeling on the prediction quality. For a given structurally identifiable model in the light of data, confidence regions result from limited information content owing to limited measurement capabilities including sampling granularity, experimental noise or number of different readouts. For proper statistical analysis, systematic errors and/or confounding effects are to be eliminated and structural identifiability has to be ensured by the experimental design. Then, a consistent estimate is possible; meaning that estimates tend to their true values, see e.g. ?. According to ?, a general definition for a joint confidence region CR for a parameter estimate  $\hat{\boldsymbol{\theta}}$  at  $1 - \alpha$  confidence level,  $\alpha \in [0, 1] \subset \mathbb{R}^+$ , can be stated by using the mapping

$$\text{cr}_\alpha : \mathbb{A}_{\mathbf{Y}} \rightarrow \mathbb{A}_{\boldsymbol{\Theta}} = \text{a region in } \mathbb{R}^{n_{\boldsymbol{\Theta}}} \quad (2.14)$$

$$\mathbf{Y} \mapsto \text{cr}_\alpha(\mathbf{Y}) = \boldsymbol{\Theta}, \quad (2.15)$$

that satisfies

$$\mathbf{P}_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \in \text{cr}_\alpha(\mathbf{Y})) \geq 1 - \alpha. \quad (2.16)$$

Here,  $\text{cr}_\alpha$  represents the (non)linear transformation from the random variable  $\mathbf{Y}$  to the random variable  $\boldsymbol{\Theta}$  via a parameter estimation procedure.  $\mathbf{P}_{\boldsymbol{\Theta}}$  is the probability of realization  $\boldsymbol{\theta}$ . In linear regression, the mapping  $\text{cr}_\alpha$  corresponds to the *normal equations*. Equivalently, a confidence interval  $\text{CR}_i$  for parameter estimate  $\hat{\theta}_i$  can be defined by the mapping

$$\text{cr}_{i,\alpha} : \mathbb{A}_{\mathbf{Y}} \rightarrow \mathbb{A}_{\boldsymbol{\Theta}_i} = \text{an interval in } \mathbb{R} \quad (2.17)$$

$$\mathbf{Y} \mapsto \text{cr}_{i,\alpha}(\mathbf{Y}) = \boldsymbol{\Theta}_i, \quad (2.18)$$

that satisfies

$$\mathbf{P}_{\boldsymbol{\Theta}_i}(\theta_i \in \text{cr}_{i,\alpha}(\mathbf{Y})) \geq 1 - \alpha. \quad (2.19)$$

Assuming that  $\mathbf{Y}$  and its transformed form  $\boldsymbol{\Theta} = \text{cr}_\alpha(\mathbf{Y})$  can be represented by some PDF, say  $\rho_{\mathbf{Y}}(\mathbf{y})$  and  $\rho_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ , a confidence region of  $\hat{\boldsymbol{\theta}} \in \mathbb{A}_{\boldsymbol{\Theta}}$  can be constructed by determining the integration domain  $\tilde{\mathbb{A}}_{\boldsymbol{\Theta}}$  that satisfies  $\int_{\tilde{\mathbb{A}}_{\boldsymbol{\Theta}} \in \text{CR}_\alpha} \rho_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1 - \alpha$ , which is equivalent to inverting the corresponding parameter CDF to find the  $1 - \alpha$  confidence quantiles. However for nonlinear models, closed form expressions for the parameter PDF are usually not at hand. Therefore one is left to either construct a parameter PDF or corresponding percentiles of the parameter PDF via appropriate sampling and/or approximations of  $\text{cr}_\alpha(\mathbf{Y})$ . There are, however, three main challenges: (i) sampling strategies have to cope with the curse of dimensionality of the parameter space, (ii) approximations of  $\text{cr}_\alpha(\mathbf{Y})$  face problems that arise from the nonlinear character of the transformation and (iii) dimensions of parameter and response space rarely coincide, resulting in under- or overdetermination of the image or pre-image speaking in the mapping interpretation of parameter inference, which is related to the identifiability of a prediction (e.g. parameter).

**Note on confidence intervals** By reconstructing a parameter PDF or percentiles thereof, it is not possible to measure the distance between estimated and (unknown) true parameter. What is meant here is the distance between best estimate and alternative estimates. The assumption then made is, that the distribution of distance between best and alternative estimates corresponds to the distribution of the distances between true parameter and best estimates (?). Constructed confidence regions thus correspond to observed ones, which for large numbers of observations converge to the true confidence regions, given the model is correct and constructed confidence regions are exact and not approximates. From a frequentist point of view, true parameters are thought to be constant and confidence intervals quantify type 1 errors for a given level of confidence with respect to their estimates. From a Bayesian point of view, confidence intervals represent probability statements with respect to the parameters, i.e. parameters are thought to be distributed (?).

**Linear approximation** It is instructive to start with confidence regions for linear, dynamic models in the form

$$\mathbf{y}(t_i, \boldsymbol{\theta}) = \mathbf{y}(t_i, \hat{\boldsymbol{\theta}}) + \mathbf{J}(t_i, \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \boldsymbol{\varepsilon}_i \quad (2.20)$$

with additive, white measurement noise  $\boldsymbol{\varepsilon} \propto \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$  and matrix  $\mathbf{J}$  (suggestive for Jacobian) one has from linear least squares theory (???)

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \propto \mathcal{N}(0, \mathbf{C}_{\Theta}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_i})). \quad (2.21)$$

The corresponding confidence region represents an ellipsoid and can be derived in closed form,

$$\text{CR}_{\alpha} = \left\{ \boldsymbol{\theta} : (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \mathbf{C}_{\Theta}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_i})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq n_{\boldsymbol{\theta}} F_{n_{\boldsymbol{\theta}}, n_{\mathbf{y}} - n_{\boldsymbol{\theta}}, 1 - \alpha} \right\}. \quad (2.22)$$

For each parameter one has the confidence interval

$$\text{CR}_{i, \alpha} = \left( \hat{\boldsymbol{\theta}}_i - z_{\alpha/2} \sqrt{[\mathbf{C}_{\Theta}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_i})]_{ii}}, \hat{\boldsymbol{\theta}}_i + z_{\alpha/2} \sqrt{[\mathbf{C}_{\Theta}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_i})]_{ii}} \right), \quad (2.23)$$

with  $z_{\alpha/2}$  being the  $(1 - \alpha/2)$  percentile of the standard Normal distribution. If the variance of the noise is estimated rather than known, Student's t-distribution is used instead of the standard Normal distribution. The parameter covariance matrix  $\mathbf{C}_{\Theta}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_i})$  is given by

$$\mathbf{C}_{\Theta}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_i}) = \mathbf{FIM}^{-1}, \quad (2.24)$$

with Fisher information matrix  $\mathbf{FIM} = \sum_{t_i} \mathbf{J}(t_i, \hat{\boldsymbol{\theta}})^T [\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_i}]^{-1} \mathbf{J}(t_i, \hat{\boldsymbol{\theta}})$ . The **FIM** corresponds to  $\mathbf{X}^T(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})^{-1}\mathbf{X}$  for linear regression models. For clearance,  $\mathbf{X}$  is given in the conventional nomenclature of linear regression, representing factors. Here, it does not represent a random state variable.

## 2. METHODS FOR IDENTIFYING DYNAMIC MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

Confidence regions for nonlinear models can be constructed by linearizing the model output. Here it is assumed that variations in the model output owing to parameter variations can be approximated by

$$\mathbf{y}^{\mathcal{L}}(t_i, \boldsymbol{\theta}) \approx \mathbf{y}(t_i, \hat{\boldsymbol{\theta}}) + \mathbf{J}(t_i, \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \boldsymbol{\varepsilon}_i, \quad (2.25)$$

for ODEs, the Jacobian  $\mathbf{J}$  can be derived by solving the sensitivity equations as described in Sec. 3.2.1. Then, results from linear regression theory apply and the confidence region/intervals are given by Eqs. (2.22), (2.23), whereas the variance-covariance matrix is approximated by the Cramer-Rao inequality

$$\mathbf{C}_{\boldsymbol{\Theta}}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_i}) \geq \frac{\partial \mathbf{E}[\hat{\boldsymbol{\theta}}]}{\partial \boldsymbol{\theta}} \mathbf{FIM}^{-1} \frac{\partial \mathbf{E}[\hat{\boldsymbol{\theta}}]}{\partial \boldsymbol{\theta}}^{\text{T}} \quad (2.26)$$

assuming additive, white noise  $\boldsymbol{\varepsilon} \propto \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$  and appropriate regularity conditions to hold (??). If the parameter estimate is additionally unbiased, i.e.  $\mathbf{E}[\hat{\boldsymbol{\theta}}] = \hat{\boldsymbol{\theta}}$ , one has

$$\mathbf{C}_{\boldsymbol{\Theta}}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_i}) \geq \mathbf{FIM}^{-1}, \quad (2.27)$$

an ideal lower bound for the parameter variance-covariance matrix. Depending on the nonlinearity, confidence regions based on this linear approximation are strong over- or underestimations, for instance in case of small  $\frac{\partial \mathbf{E}[\hat{\boldsymbol{\theta}}]}{\partial \boldsymbol{\theta}}$  (see ? for an example).

**Asymptotics** Asymptotic confidence regions can be constructed by using the asymptotic properties of the MLE. For an MLE one has  $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \propto \mathcal{N}(0, \mathbf{C}_{\boldsymbol{\Theta}})$ . The variance-covariance matrix of the parameters is given by (??)

$$\mathbf{C}_{\boldsymbol{\Theta}} = \left( -\mathbf{E} \left[ \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \right] \right)^{-1} \stackrel{\boldsymbol{\varepsilon}_i \propto \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_i})}{\downarrow} \mathbf{FIM}^{-1}. \quad (2.28)$$

The confidence region is again given by Eq. (2.22). As is indicated in Eq. (2.28), asymptotically both confidence regions are equivalent for nonlinear models. For linear models, they are equivalent. For small sample size, however, estimated confidence regions might differ significantly since Eq. (2.28) comprises second order derivatives of the model output with respect to the parameters (?). In ? it is argued that the simplest approximation via **FIM** seems often the most appropriate choice given estimation accuracy, numerical effort and stability.

**Likelihood** Based on the likelihood, an exact confidence region can be defined according to (?)

$$\text{CR}_{\alpha} = \{\boldsymbol{\theta} : L_n(\boldsymbol{\theta}) \leq \xi L_n(\hat{\boldsymbol{\theta}})\}, \quad (2.29)$$

and likewise the confidence interval

$$\text{CR}_{i,\alpha} = \{\boldsymbol{\theta}_i : L_n(\boldsymbol{\theta}_i) \leq \xi L_n(\hat{\boldsymbol{\theta}}_i)\}, \quad (2.30)$$

where  $\xi > 1$  represents some constant of statistical significance. This definition is simply the inversion of the likelihood ratio test. Depending on the number of data and available knowledge about noise realization, a statistically significant value for  $\xi$  can be derived, e.g. based on  $F$  or  $\chi^2$  statistics (?). Since in general the likelihood function is not given in closed form, several Monte Carlo and bootstrap based methods exist to derive likelihood-based confidence regions (??).

Another elegant concept, which is related to likelihood-based confidence regions, is based on the profile likelihood. Profile likelihood-based confidence intervals are grounded on marginalized likelihoods subjected to the constraint of maximizing the likelihood. Once determined, the profile likelihood can also be used for parameter identifiability analysis (Sec. 2.2.3), model prediction uncertainty estimation (Sec. 3.5, ?) and experimental design (?). The profile likelihood and confidence regions are determined as follows: Starting from additive, white noise, the weighted residual sum of squares (Eq. 2.11) can be used to construct likelihood-based confidence intervals according to

$$\text{CR}_\alpha\{\boldsymbol{\theta}_i : \chi^2(\boldsymbol{\theta}_i) - \chi^2(\hat{\boldsymbol{\theta}}_i) \leq \delta\chi_{\text{df},\alpha}^2\}, \quad (2.31)$$

with  $\delta\chi_{\text{df},\alpha}^2$  being the  $\alpha$  quantile of the  $\chi^2$  distribution for df degrees of freedom. Point-wise confidence intervals are obtained by  $\text{df} = 1$ , simultaneous ones via  $\text{df} = n_{\boldsymbol{\theta}}$  (??). Importantly here to notice, parameter dependencies are neglected. In contrast, the profile likelihood uses the following modification, which allows projecting the entire likelihood information of the high dimensional parameter space onto one parameter coordinate, treating the remaining parameters as nuisance parameters but accounting for parameter interdependencies via

$$\chi^2(\boldsymbol{\theta}_i)_{\text{PL}} = \min_{\boldsymbol{\theta}_{i \neq j} \in \mathbb{A}_{\boldsymbol{\theta}}} \chi^2(\boldsymbol{\theta}). \quad (2.32)$$

Here  $\chi^2(\boldsymbol{\theta}_i)_{\text{PL}}$  of parameter  $\boldsymbol{\theta}_i$  represents values of the smallest residual sum of squares (=largest likelihood levels) when moving  $\boldsymbol{\theta}_i$  away from  $\hat{\boldsymbol{\theta}}_i$ . Then, profile likelihood-based confidence intervals correspond to Eq. (2.31), with  $\chi^2(\cdot)$  being replaced by  $\chi_{\text{PL}}^2(\cdot)$  (?). ? illustrated, that in the small sample case, profile likelihood-based confidence regions have better coverage range than asymptotic-based ones. This results from the fact that likelihood ratio statistic converges faster to its asymptotic  $\chi^2$  distribution than the equivalent Wald statistic (??). Likewise for the log-likelihood and weighted residual sum of squares in case of standard conditions to hold. It is important to note that often and especially in the small sample case the residual sum of squares Eq. (2.11) needs not to follow a  $\chi^2$  distribution. Regarding confidence intervals, it is then necessary to numerically derive the distribution of Eq. (2.11) - see next paragraph for possible approaches - and adjust the threshold parameter accordingly.

## 2. METHODS FOR IDENTIFYING DYNAMIC MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

**Sample-based approach** Sample-based approaches repeat the parameter estimation procedure several times. From the resulting set of parameter estimates, a joint or marginal parameter PDF as well as percentiles may be constructed. Simple Monte Carlo based approaches simulate the model for a sufficiently large set of independent and identically distributed parameter samples. Then, for each of the simulation realizations a new parameter estimation is performed. Based on the resulting population of best estimates, a confidence region can then be constructed (??).

Bootstrapping is an alternative to Monte Carlo, which replaces the initial simulation step via bootstrapping the data itself (??). Here pseudo data are generated from the original data by drawing samples from the original data set with equal probability and replacement. Then, for each bootstrap sample a parameter estimate is derived resulting in a population of best parameter estimates, which again represent the parameter PDF. Markov chain Monte Carlo (MCMC) methods have become very popular over the last years, as they allow drawing correlated samples - thus increasing convergence compared to simple MC approaches - from an arbitrary PDF to learn a Markov chain, which represents the parameter PDF by its stationary distribution (??). One might also use the deterministic sigma point method, as illustrated in ?. Here, deterministically chosen samples from the data are transformed by the parameter estimation and used to derive an expected parameter estimate and corresponding confidence interval, including information on the bias of the estimate. The sigma point method was also used in this thesis to design robust discrimination experiments as is discussed in detail in Sec. 3.2.2.

### 2.2.3 Parameter identifiability

When it comes to model predictions on internal dynamics, or domains beyond data coverage, model identifiability is of crucial importance. Model identifiability analysis investigates whether a model parameter can be uniquely determined based on a given input-output setting. It is often referred to as parameter identifiability analysis. In an idealized setting, where the model  $\mathcal{M}(\boldsymbol{\theta})$  corresponds to the data generating process in addition to assuming perfect measurement conditions, i.e. no noise, free choice of input and measurement time points, a definition for structural identifiability can be given as follows. A model  $\mathcal{M}(\boldsymbol{\theta})$  with output  $\mathbf{y}(t, \boldsymbol{\theta}, \mathbf{u})$  is uniquely identifiable at  $\boldsymbol{\theta}^*$  (true value), if there is only one solution to the equation

$$\mathbf{y}(t, \hat{\boldsymbol{\theta}}, \mathbf{u}) = \mathbf{y}(t, \boldsymbol{\theta}^*, \mathbf{u}), \quad \forall t \in \mathbb{R}^+, \forall \mathbf{u} \in \mathbb{A}_{\mathbf{u}} \text{ and } \hat{\boldsymbol{\theta}} \in \mathbb{A}_{\boldsymbol{\theta}}, \quad (2.33)$$

namely  $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$ . Because the true value is rarely known, structural identifiability has either a global or local character. In the later case, there exists a countable set of solutions  $\hat{\boldsymbol{\theta}}$  to Eq. (2.33), which ensures that a neighborhood around  $\boldsymbol{\theta}^*$  can be defined, in which  $\hat{\boldsymbol{\theta}}$  is unique (?). For non-identifiable parameters, there exists an uncountable set of parameters, which yield the same model input-output behavior, whereas predictions on internal states - states that are not directly observed - may be completely different.



Consequently, such an identifiability analysis has two obvious reasons. First, non-identifiable parameters hint at a necessary re-design of the applied or envisioned input-output setup. If this is not possible a model re-parameterization can resolve non-identifiabilities, or at last, an exclusion of non-identifiable parameters from the parameter estimation procedure avoids poor convergence of the parameter estimation procedure. A second reason results from the purpose of modeling. Model-based predictions on internal states that are related to non-identifiable parameters have to be considered with care. Further, non-identifiabilities can be resolved by accounting for parameter constraints, e.g. by imposing the Wegscheider condition for a thermodynamic consistent model calibration (??) or by parameterizing the model in a thermodynamic safe way (??).

There exist several approaches for identifiability analysis, which can be classified into theoretical (structural, a priori) and practical (a posterior) identifiability analysis. Theoretical identifiability analysis serves the aforementioned first reason, whereas practical identifiability analysis is mainly used to address the second reason in the light of data. Theoretical identifiability analysis is solely based on structural properties of a given model, trying to clarify the uniqueness of a parameter set as defined in Eq. (2.33). Approaches include Laplace transform (?), Taylor/Lie series (?), similarity transformations and differential algebra approaches (??). Here, the right choice of method is not straight forward and heavily depends on model size, complexity and degree of nonlinearity. During my Ph.D. work, we have been testing software packages provided by ?? within the scope of ODE modeling in ? and a student research project by Katja Tummler (?). Whereas the method from ? - implemented in the computer algebra system REDUCE (?) - seems only applicable to small, nonlinear systems, the method from ? is also applicable to larger systems at the cost of only providing a probabilistic statement about identifiability. However, given its fast computation time even for larger systems, the method by ? seems favorable.

Besides structural a priori analysis, a posterior analysis allows assessing practical identifiability of an estimated parameter set for given input-output data with limited information content resulting from limited sampling rates, limited number of different input profiles and corruption by noise. It is closely related to the construction of confidence regions and yields in most of the cases a local identifiability statement corresponding to definition in Eq. (2.33). For model identification, a practical identifiability analysis is mandatory to understand prediction limitations for a given model. For biological models, identifiability methods based on the shape estimation of the likelihood function, e.g. Hessian or Fisher information (???), only work well in cases where identifiability problems arise from linear interdependencies between parameters, which is often not the case for nonlinear models as for instance illustrated by ?.

A well-proven approach, which works well for (non-)linear models is the profile likelihood approach. Here, the profile likelihood is used to determine whether parameters are practically (non-)identifiable (?). It is also possible to reveal structural

## 2. METHODS FOR IDENTIFYING DYNAMIC MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

non-identifiabilities. In the later case, profile likelihoods are perfectly flat, since non-identifiable parameter variations are either compensated by proper adjustment of other parameters (over parameterization) or structurally not observable (?). Practical non-identifiable parameters are characterized by an initial increase of the  $\chi^2$  function (decrease of the profile likelihood), which at least for one direction gradually flattens out when moving away from the MLE. An identifiable parameter has an increasing  $\chi^2$  (decreasing profile likelihood) when de-/increasing a parameter from its MLE, which eventually hits the critical confidence level defined by Eq. (2.31). Examples of profile likelihoods are given in one of the applications in Sec. 3.5. Further details on concepts and methods for identifiability analysis can be found in ????????

### 2.3 Model discrimination

In the previous section, concepts of parameter estimation, confidence regions and identifiability have been discussed for a given model structure. However, when modeling a system from scratch, modelers typically start with a set of several alternative models. Then, model discrimination can be used to select the most plausible model or to at least establish a plausibility hierarchy amongst the competing models. This sections recalls and illustrates important concepts of model discrimination, which can be understood as a generalization to parameter identification and identifiability (?).

#### 2.3.1 Model distinguishability

Structural parameter identifiability analysis has been discussed in Sec. 2.2.3 as an important step in model identification to asses whether a given input-output setup allows identifying a parameter set for a defined model structure in a unique way. With regard to model discrimination, a related concept known as model output distinguishability applies (?). Here, it is desired to know, whether a given input-output setup allows discriminating between two given models structures  $\mathcal{M}(\boldsymbol{\theta})$  and  $\tilde{\mathcal{M}}(\tilde{\boldsymbol{\theta}})$ . According to ?, the following definition of structural model output distinguishability can be given:

$\tilde{\mathcal{M}}$  is structurally output distinguishable from  $\mathcal{M}$  if, for almost any  $\boldsymbol{\theta} \in \mathbb{A}_{\boldsymbol{\theta}}$  the equation

$$\tilde{\mathbf{y}}(t, \tilde{\boldsymbol{\theta}}, \mathbf{u}) = \mathbf{y}(t, \boldsymbol{\theta}, \mathbf{u}), \quad \forall t \in \mathbb{R}^+, \forall \mathbf{u} \in \mathbb{A}_{\mathbf{u}}, \text{ and } \tilde{\boldsymbol{\theta}} \in \mathbb{A}_{\tilde{\boldsymbol{\theta}}} \quad (2.34)$$

has no solution for  $\tilde{\boldsymbol{\theta}}$ . Further, that  $\tilde{\mathcal{M}}$  is structurally output distinguishable from  $\mathcal{M}$  does not imply that  $\mathcal{M}$  is structurally output distinguishable from  $\tilde{\mathcal{M}}$ . If both directions hold, then  $\mathcal{M}$  and  $\tilde{\mathcal{M}}$  are structurally output distinguishable. In the case of three or more competing models, a pairwise comparison has to be performed. Typically, an *exhaustive summary* is derived, which allows eliminating time and input functions, see e.g. ?. Still, owing to the fact that parameter spaces of the competing models may be of different dimensions, testing model output distinguishability directly from the above definition is a non-trivial task. In practice, Laplace transform approaches can be helpful, but also

other modified methods based on the ones used for parameter identifiability analysis (???)

Finally, although structural parameter identifiability and structural output distinguishability are closely related, identifiability is neither a necessary nor sufficient condition for structural output distinguishability. Consequently, optimal experimental design aimed at model discrimination is also applicable to unidentifiable models. Also note that identifiability analysis seeks to prove uniqueness of the solution to Eq. (2.33), whereas output distinguishability analysis aims at proving non-existence of a solution to Eq. (2.34).

### 2.3.2 Model selection

Besides structural distinguishability analysis, model selection or discrimination refers to several other aspects of the model identification procedure. At the very beginning of the modeling process, an appropriate modeling approach has to be chosen for the desired model purpose, which comprises defining a desired degree of model complexity (level of detail, number of variables, computational effort, physico-chemical rigor) (?). By doing so, the modeler already discriminates between the set of all possible models associated to the chosen modeling approach and the remaining model classes. In the following, it is assumed that a modeling purpose, a set of data and models are given. The task of model discrimination is then to identify a model or set of model structures, which serve the modeling purpose and is consistent with the data (???). In this setting, model discrimination is in fact model adequacy testing (=falsification procedure) sorting out models that do not adequately describe existing data (?). From the remaining models, one might then discriminate, again based on trading off goodness-of-fit and model complexity. Finally, if the data do not suffice to identify one final model or at least to obtain a statistically significant hierarchy amongst the remainder-models, new data using OED should be generated (s. application in Sec. 3.5).

As for confidence intervals, model discrimination is based on comparing distances between models' predictions and data. Here either classical hypothesis testing or approaches from information theory are available. Whereas classical hypothesis testing seeks to sort out models based on test statistics, which allow specifying p-values given a test distribution. For the finite sample case, proper statistical model testing is however limited to nested models (?). Alternatively, Bayesian or information theory approaches can be used, which easily extend to non-nested models (?). Both approaches have been developed in the 70s. ? proposed a simple measure of divergence, whereas ? thought of a discrimination function, i.e. a model response PDF, which has been extended by ? to incorporate prior model probabilities.

In the case of nested models, there exists a structural hierarchy between models such that one model is a special case of a larger model. This is for instance found in linear regression. In such a case, one might either use an F- or likelihood ratio test (?). Both tests are asymptotically equivalent, whereas the likelihood ratio test has more power

## 2. METHODS FOR IDENTIFYING DYNAMIC MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

(?). Having two models  $\mathcal{M}(\hat{\theta})$  and  $\tilde{\mathcal{M}}(\hat{\tilde{\theta}})$ , with parameter MLE  $\hat{\theta}$ ,  $\hat{\tilde{\theta}}$  and corresponding  $\chi_{\mathcal{M}}^2(\hat{\theta})$ ,  $\chi_{\tilde{\mathcal{M}}}^2(\hat{\tilde{\theta}})$  values, whereas  $\tilde{\mathcal{M}}$  is nested in  $\mathcal{M}$ , the test statistics

$$F = \frac{\chi_{\tilde{\mathcal{M}}}^2(\hat{\tilde{\theta}}) - \chi_{\mathcal{M}}^2(\hat{\theta})}{n_{\theta} - n_{\tilde{\theta}}} \frac{n_{\mathbf{Y}} - n_{\theta} - 1}{\chi_{\mathcal{M}}^2(\hat{\theta})} \quad (2.35)$$

follows an F-distribution,  $F \propto F_{n_{\theta} - n_{\tilde{\theta}}, n_{\mathbf{Y}} - n_{\theta} - 1}$ , and asymptotically ( $n_{\mathbf{Y}} \rightarrow \infty$ ) becomes a  $\chi_{n_{\theta} - n_{\tilde{\theta}}}^2$  distribution. Based on this statistics, it is then possible to sort out all models, which differ from the expected statistics for a predefined level of statistical significance and are thus not adequate, either due to lack of fit or over-fitting (noise). The likelihood ratio test follows the same principle with test statistic

$$LR = 2(L(\tilde{\mathcal{M}}(\hat{\tilde{\theta}})) - L(\mathcal{M}(\hat{\theta}))), \quad (2.36)$$

which is  $\chi_{n_{\theta} - n_{\tilde{\theta}}}^2$  distributed. Just as for the F-test, models need to be nested and  $\mathcal{M}(\hat{\theta})$  must belong to the class of the true model. Following the line of reasoning for these two test procedures, it is thus possible to discriminate models, which do not fall into a certain range of model complexity consistent with the data. In stepwise linear regression, this is applied by forward and backward elimination (?) of parameters. Forward elimination would correspond to an engineering strategy (seek minimal model required for adequate description of the data), backward elimination would correspond to a scientific strategy (seek the most complex model still supported by the data) (?).

In the case of non-nested models an analogous reasoning is not straight forward. In contrast to linear regression, where model complexity is measured by the number of parameters or degrees of freedom, it is not clear which model is to be used as the most complex/simple one to which all others should be compared. In this case, Bartlett's  $\chi^2$  test of homogeneity of variances amongst the models can help to sort out models with large error variances in the following way (?). The test statistic

$$\chi_{\text{B}}^2 = \frac{\sum_{\mathcal{M}=1}^{n_{\mathcal{M}}} (n_{\mathbf{Y}} - n_{\theta_{\mathcal{M}}}) \log(\chi_{\text{tot}}^2 / \chi_{\mathcal{M}}^2)}{1 + \frac{1}{3(M-1)} \left( \sum_{\mathcal{M}=1}^{n_{\mathcal{M}}} \frac{1}{n_{\mathbf{Y}} - n_{\theta_{\mathcal{M}}}} - \frac{1}{\sum_{\mathcal{M}=1}^{n_{\mathcal{M}}} (n_{\mathbf{Y}} - n_{\theta_{\mathcal{M}}})} \right)}, \quad (2.37)$$

should follow a  $\chi^2$  distribution with  $n_{\mathcal{M}} - 1$  degree of freedoms and  $\chi_{\text{tot}}^2 = \sum_{\mathcal{M}=1}^{n_{\mathcal{M}}} \chi_{\mathcal{M}}^2$  the total error variance over all models. If  $\chi_{\text{B}}^2$  is rejected, then remove the model with the largest error variance, re-evaluate  $\chi_{\text{B}}^2$  and possibly remove additional models until  $\chi_{\text{B}}^2$  cannot be rejected anymore. Alternatively, a J-test (?) or adjusted likelihood ratio statistic can be used but remains questionable for the finite sample size as it still relies on a reference distribution, which is only valid asymptotically. For details on adjusted likelihood ratio tests see ???. In the most general case, i.e. non-nested models, misspecified models and non-Gaussian observational noise, an appropriate test statistic can be estimated via simulation without any need for an asymptotic argument. Here bootstrapping has become an important method owing to increased availability

of computational power. For details and examples on bootstrapping based selection methods see for instance [?].

Information-based model selection criteria follow a different paradigm. Here, model discrimination is understood as identifying an evident-based (=plausibility) hierarchy amongst the models that are supported by the data. This is in principle also possible with hypothesis testing using p-values, but the classical view is to classify into significant and non-significant models. Information-based model selection disregards the rather unlikely assumption of one correct model in the modeling pool (*...all models are wrong, some are useful...* [?, p. 74]). Model discrimination is thus understood as building an order from best to worst model trading of goodness-of-fit and model complexity. Information-based selection often builds on the principle of Occam's razor, which follows the principle of parsimony, i.e. preferring the least complex hypothesis still compatible with given data. Here Akaike's information criterion (AIC) as an estimator of the relative expectation of Kullback-Leibler distance based on Fisher's maximized log-likelihood is the most prominent model discrimination criterion and often used in biological science [?]. However, as illustrated in [?], AIC can be very sensitive to noise. It is thus important to also consider the variability of any information-based criterion, see Sec. 2.4.4. Further details on information-based model selection and discussions including critical AIC values for discrimination, Bayesian information criterion, minimal description length or Mallows's C can be found in [?]. A comparison between F-test and AIC selection is for instance given by [?].

Finally, residual analysis provides an additional tool for discriminating models. Assuming standard conditions to hold, residuals of the fitted models should follow a standard Normal (error variance known a priori) or t distribution (error variance estimated from samples). Therefore, one can look at a qq-plot of residuals vs. standard Normal/t distribution or derive p-values from a Normality/t-test, e.g. Kolmogorov-Smirnoff or Anderson-Darling tests. Furthermore, such tests allow identifying outliers, which may point to model weaknesses, convergency problems of the parameter estimation or experimental errors. They may also reveal non-overlapping features or misspecification of two competing models. The most appropriate model should have a small residual sum of squares with Normal residual distribution and small number of parameters. An example of such analysis is given in [?], where it is shown that minimal  $\chi^2$  values do not necessarily ensure normal residual distribution owing misspecification in the model. In Sec. 3.5, Anderson-Darling testing was also applied to justify the model choice.

**Note on Bayesian analysis** The aforementioned methods belong to frequentist approaches, i.e. inference is based on fixed, deterministic, parametric models in the light of varying data. This also holds for the case of distributed parameters, since here a fixed parameter PDF is assumed. In contrast, the Bayesian school has a stochastic model interpretation and inference is based on the prior, or on the belief in a given hypothesis. Data are then used to modify the belief, whereas frequentists create belief out of data. Bayesian methods rely on Bayes' theorem, which relates prior and post belief in form of

## 2. METHODS FOR IDENTIFYING DYNAMIC MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

a distribution function via the normalized likelihood. An advantage but also a catch of the Bayesian approach is the possibility of incorporating prior knowledge via the prior. If however this prior knowledge is poor, then flat priors have to be used, which in turn means that the posterior distribution is approximately the normalized likelihood function. Then Bayesian analysis is more or less equivalent to frequentist analysis. Further details can be found in ?.

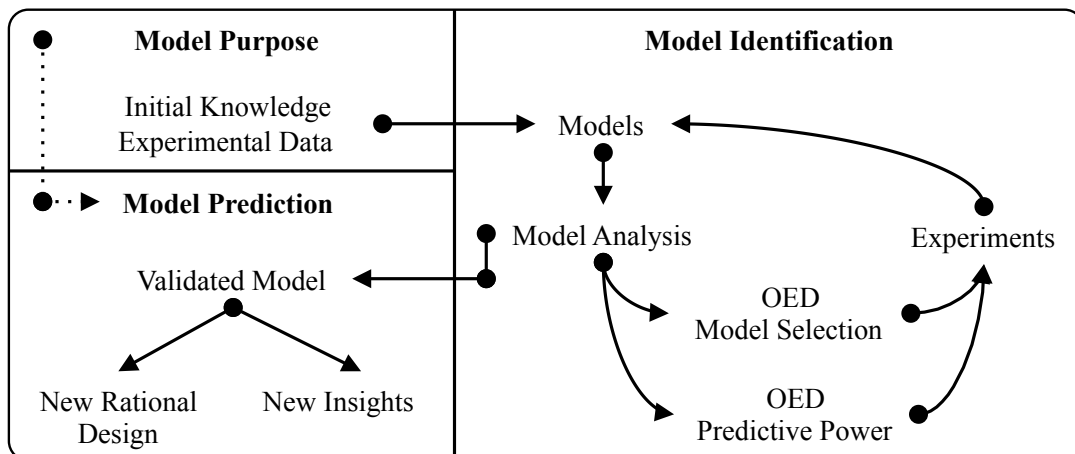
### 2.4 Optimal experimental design

Data analysis is as old as mankind has started to explore and analyze its surroundings. Structured analysis of detailed observation by means of mathematics has been developed ever since and gave rise to a plethora of mathematical methodologies, including simple correlation, linear regression or analysis of variances (ANOVA). But no matter how sophisticated mathematical methodologies for data analysis may be, they cannot overcome the lack of information contained in the data. From a naive point of view one might think that increasing the number of repetitions may serve the goal to increase data quality in terms of information content. In principle, this is one way, since according to the central limit theorem the precision of averaged data is improved as  $\text{STD}[E[y]]_{n=m} = \frac{\text{STD}[E[y]_{n=1}]}{\sqrt{m}}$  with  $n, m \in \mathbb{N}^+$  and STD standard deviation. But additionally, each set of repetitions can be improved to contain maximal information with respect to a specific objective or data analysis goal.

Well-designed experiments are the most substantial ingredient for informative data and successful model identification. Methodologies for design of experiments have been developed from the beginning of the 1920s including seminal work by ????. Back then factorial, blocking and randomization strategies have been used to plan experiments for optimal model identification using empirical linear regression models, which allowed to derive several closed form expressions for optimal experimental plans. In contrast, models in form of ordinary differential equations do not allow to explicitly state an optimal design. Here, owing to the lack of closed-form expressions, an iterative model-based optimization approach has to be followed, which is often hampered by highly uncertain parameters resulting in large uncertainties in the model predictions. For large-scale network reconstruction in genetical genomics, experimental design in its classical form is used to design strains with the aim to generate a study population yielding most informative data with respect to gene-gene but also gene-phenotype interactions. Several design examples are given in Sec. 4.2.

#### 2.4.1 Working definition: Optimal experimental design

An experimental design specifies a set of independent experimental variables that influence the system of interest. The idea of OED is to optimally choose these independent experimental variables. Optimality refers to some performance score, which represents an objective, being optimally adjusted by the corresponding optimal design  $\mathbf{D}^\dagger$ , whereas



**Figure 2.1:** Model identification based on a cyclic iteration between model analysis and experiments, including OED. Starting from the purpose the model should serve - usually help answering a scientific question or achieve an engineering goal - models are generated based on initial knowledge and data at hand. Then, in an iterative workflow, this knowledge is refined by adjusting the models until convergence to a validated model, which serves the initial model purpose by verifiable model-based predictions (scientific insights, rational design).

$\mathbf{D} \in \mathbb{D}$  represents an experimental design within the feasible design region  $\mathbb{D}$ , for instance  $\mathbf{D} \in \mathbb{D} = \mathbb{T} \times \mathbb{U} \times \mathbb{Y}$  encompassing selection of discrete measurement time points  $t_k \in \mathbb{T}$ , stimulus design  $\mathbf{u}(t) \in \mathbb{U}$  and readout design  $\mathbf{g} \in \mathbb{Y}$ . In general, different objectives yield different optimal designs.

Experimental design strategies may be divided into qualitative and quantitative approaches. In Secs. 2.2.3, 2.3.1, the concepts of a priori identifiability and input-output distinguishability have been introduced. Qualitative experimental designs for model identification aim at resolving potential identifiability or distinguishability problems based on the model structure only, by identifying suitable new input or output variables. ? have developed an approach for qualitative OED that allows identifying measurement signals that are most informative for parameter estimation based on adjacency matrix of the extended system. In this work we focus on sequential, quantitative experimental design, i.e. for given data, models are (re)analyzed, possibly modified and experiments are planned and performed for further model identification until convergence is reached, Fig. (2.1). Quantitative experimental design can further be grouped into off- and online designs. In offline designs, all acquired data are used to optimize future experiments. In contrast, online designs are optimized during an experimental run, taking advantage of new measurements during the experimental run ?? . This however requires an experimental setup, which allows instant data collection and processing coupled to an optimizer that feeds back adjustments to the design variables based on updated model predictions. In this thesis we are concerned with offline optimization,

## 2. METHODS FOR IDENTIFYING DYNAMIC MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

which is still the most prevailing setting found in nowadays biological research activities.

### 2.4.2 Experimental design for optimal parameter estimation

Much work on optimal experimental design for biological systems with distributed parameters focuses on information maximization with respect to parameter identification, e.g. [10]. Here, for a given pool of plausible ODE models, OED aimed at best parameter estimation predicts experimental conditions, which yield time course data that decouple model parameters and at the same time contain maximal information for all parameter values. This is analogous to improving the condition on the design matrix in combination with reduced covariances for linear models. Classical approaches use the Fisher information (Sec. 2.2) to find experimental designs that are A, D, E, optimal, which represent a selection of different criteria condensing the Fisher information into one numerical value, see for instance [11]. Note that A, D and E are not to be confused with the nomenclature of this thesis. A-optimal designs maximize the trace of the Fisher information matrix or minimize that of the parameter variance-covariance matrix. D-optimal designs maximize the determinant of the Fisher information matrix or minimize that of the parameter variance-covariance matrix. E-optimal designs maximize the smallest eigenvalue of the Fisher information matrix or minimize the largest eigenvalue of the parameter variance-covariance matrix. In a recent study, [12] demonstrated that designs based on the FIM, which operate in the parameter space, may be outperformed by designs that directly minimize model prediction variances.

### 2.4.3 Experimental design for optimal model discrimination

An experimental design aimed at model discrimination is typically generated at a point, where existing data do not provide further discriminative information for a pool of competing models. Research on discriminative experimental design dates back to the 1960s, including work from [13]. [14] formulated a divergence criterion as the square difference between two competing model predictions. [15] derived a divergence measure starting from the concept of entropy. [16] introduced the notion of T-optimality (T means test) for two competing regression models. [17] have build on these works to develop a modified T criterion and extension to multiple response setups, where model prediction uncertainties are also accounted for. This represents a robustification of the experimental design against parameter variations. Finally, [18] have extended the criterion from [16] to the dynamic case. For two competing models  $\mathcal{M} = \{i, j\}$  with dynamic output  $\mathbf{y}_{\text{sim},\mathcal{M}}(t_k, \mathbf{D})$ , the modified T criterion reads [18]

$$\mathbf{T}_{ij}(\mathbf{D}) = \frac{1}{n_t} \sum_{k=1}^{n_t} (\mathbf{y}_{\text{sim},i}(t_k, \mathbf{D}) - \mathbf{y}_{\text{sim},j}(t_k, \mathbf{D}))^T \mathbf{S}(t_k, \mathbf{D})^{-1} (\mathbf{y}_{\text{sim},i}(t_k, \mathbf{D}) - \mathbf{y}_{\text{sim},j}(t_k, \mathbf{D})) \quad (2.38)$$

$$\mathbf{S}(t, \mathbf{D}) = 2\mathbf{S}_{\text{exp}}(t, \mathbf{D}) + \mathbf{S}_i(t, \mathbf{D}) + \mathbf{S}_j(t, \mathbf{D}). \quad (2.39)$$



Here,  $\mathbf{S}_{\text{exp}}(t, \mathbf{D})$  represents the variance-covariance matrix of experimental errors,  $\mathbf{S}_{\mathcal{M}}(t, \mathbf{D})$  the variance-covariance of the expected response based on model  $\mathcal{M} = \{i, j\}$ . For a single response system one has

$$T_{ij}(\mathbf{D}) = \frac{1}{n_t} \sum_{k=1}^{n_t} \frac{(y_{\text{sim},i}(t_k, \mathbf{D}) - y_{\text{sim},j}(t_k, \mathbf{D}))^2}{2\sigma_{\text{exp}}^2 + \sigma_{\text{sim},i}^2(t_k, \mathbf{D}) + \sigma_{\text{sim},j}^2(t_k, \mathbf{D})}. \quad (2.40)$$

Apparently, for two rivaling models, one needs to find a design  $\mathbf{D}$  that maximizes  $T_{ij}(\mathbf{D}) > 1$ , since then the variance of the divergences between the expected model responses is explained in terms of error variance of the experiment and variance of the expected responses. One may interpret  $T_{ij}(\mathbf{D})$  as the absolute value of a z-score, which needs to exceed one standard deviation in order to have statistical significance. As noted by ?, even under Normality assumption of the model responses,  $T$  is not properly distributed as an F distribution owing to correlations between the model divergencies ( $y_{\text{sim},i}(t_k, \mathbf{D}) - y_{\text{sim},j}(t_k, \mathbf{D})$ ). If the expected model response  $\mathbf{E}[\mathbf{y}_{\text{sim},\mathcal{M}}(t_k, \mathbf{D})]_{\mathbf{e}}$  is easily computed, it should be preferred over  $\mathbf{y}_{\text{sim},\mathcal{M}}(t_k, \mathbf{D})$ .

In the case of multivariate, multi-modal and non-Gaussian response distributions the modified  $T$  criterion partially fails to adequately represent differences in the model predictions. Multi-modalities are not accounted for, which frequently occur when modeling multi-site phosphorylation events in signal transduction systems with ODEs, e.g. ?. Here, a further generalization of the  $T$  criterion to the model overlap has been given by ???. The generalization is based on directly comparing model response PDFs. Details on the model overlap are given in Ch. 3.

#### 2.4.4 Robust optimal experimental design

A major challenge of experimental design focused at model identification is that it relies on predictions from models that yet have to be identified. Therefore, it is to be questioned, whether an OED derived from model predictions is superior to ad hoc choices based on the experience of experimenters. This becomes even more problematic when dealing with uncertainties in the data and thus model parameters. Sources of uncertainties in the data comprise biological variability but also complex measurement techniques and sub-optimally performed experiments (including insufficient observability of parameters). In order to overcome this problem, model-based experimental design can be *robustified*. This means that the performance score of an experimental design is less sensitive to the different kinds of uncertainties. In detail, robustness of the experimental design is achieved by considering

- (i) pure uncertainty of the model itself,
- (ii) distributed model predictions that arise from distributed model parameters,
- (iii) measurement noise and
- (iv) design variabilities (e.g. variations of the applied stimulus)

## 2. METHODS FOR IDENTIFYING DYNAMIC MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

during the conduction of the experiment. Notice that in applications, (ii) and (iii) will jointly contribute to a distributed parameter space, and thus distributed model responses. The difference is that (ii) is an intrinsic and (iii) an extrinsic source of variability in the parameter and response space, respectively. In the context of OED modeling of biological systems, several authors have demonstrated that robustification of the design against parameter uncertainties strongly improves the designed experiments and experimental data quality (??????????).

Notice that variations of the measurement system, e.g. temperature, pressure, initial conditions or cell cycle state, can have a strong influence on the state of the biological system. That is, cell activity and related measurements can be altered completely under the same experimental design. Consideration and reduction of such covariates (=confounding effects) is of utmost importance, in order to draw valid conclusions from measurements. This is typically achieved by increasing the number of experimental replicates and randomization (?), but also by focusing on *in vitro* analysis of *one* specific cell type, which is cloned and cultivated throughout the experiments under constant conditions. Still, the variability in the replicate data may be high. ? have demonstrated that even sister cells can respond differently under stress conditions owing to natural occurring differences in protein levels.

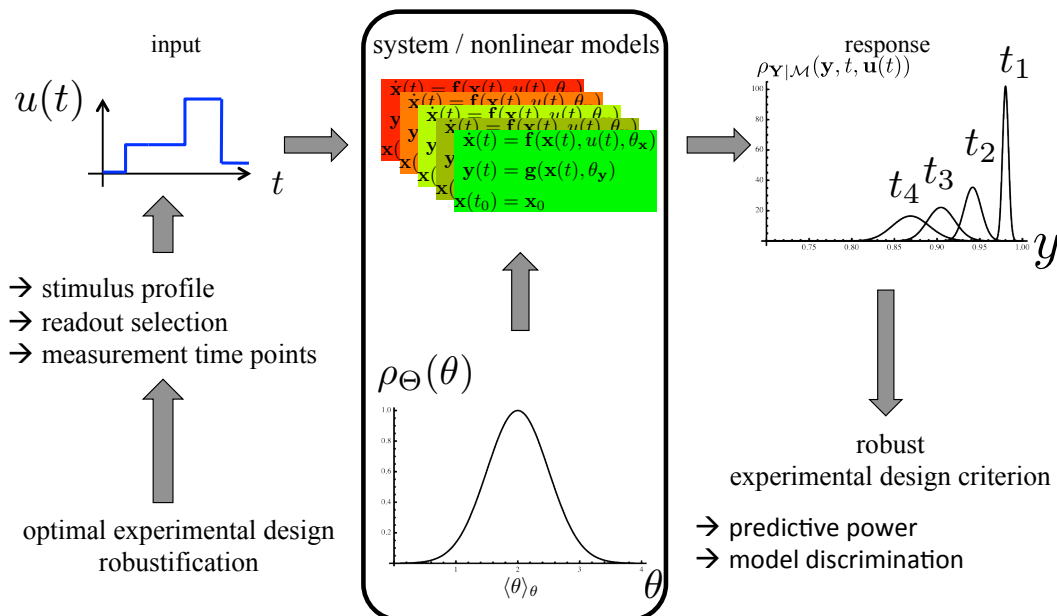
A robust experimental design that accounts for prediction uncertainties (i-iii) is obtained by optimizing the expected objective  $O$

$$\mathbf{E}[O(\mathbf{D})]_{\mathcal{M},\varepsilon,\theta} = \sum_{\mathcal{M}} \mathbf{P}_{\mathcal{M}} \int_{\mathbb{A}_{\varepsilon}} \int_{\mathbb{A}_{\theta}} \rho_{\varepsilon}(\varepsilon) \rho_{\Theta_{\mathcal{M}}}(\theta_{\mathcal{M}}) O(\mathbf{D}, \theta_{\mathcal{M}}, \varepsilon) d\varepsilon d\theta, \quad (2.41)$$

where  $\mathbf{P}_{\mathcal{M}}$  represents the probability of model structure  $\mathcal{M}$ , which can be derived from prior or - in case of new experimental data - posterior model analysis. Further, prediction uncertainties that result from uncertain parameters and measurements are accounted by their respective PDFs, i.e.  $\rho_{\Theta_{\mathcal{M}}}(\theta_{\mathcal{M}})$  and  $\rho_{\varepsilon}(\varepsilon)$ . Although point (iii) is typically independent of the design it should be included in the robustification to predict whether a specific experimental design will yield significant results under the given measurement noise. To account for design variabilities, one might use

$$\mathbf{E}[\mathbf{E}[O(\mathbf{D})]_{\mathcal{M},\varepsilon,\theta}]_{\mathbf{D}} = \int_{\mathbb{D}} \rho_{\mathbf{D}}(\tilde{\mathbf{D}}) \mathbf{E}[O(\tilde{\mathbf{D}})]_{\mathcal{M},\varepsilon,\theta} d\tilde{\mathbf{D}}, \quad (2.42)$$

where the design variabilities are described by  $\rho_{\mathbf{D}}(\tilde{\mathbf{D}})$ . The subscript  $\mathbf{D}$  indicates, that  $\rho_{\mathbf{D}}(\tilde{\mathbf{D}})$  itself is a function of the actual design  $\mathbf{D}$ . For practical application, one will typically use  $\rho_{\mathbf{D}}(\tilde{\mathbf{D}}) \propto \mathcal{N}(\tilde{\mathbf{D}}, \mathbf{C}(\mathbf{D}))$ , where variance-covariance  $\mathbf{C}(\mathbf{D})$  will depend on the design. For biological experiments, this reflects the variance of the used devices, conduction complexity, experimental reproducibility but also experimental skills of the wet lab for the design  $\mathbf{D}$ . As should be clear, robust OED is a delicate task, comprising integration in high dimensional spaces that is embedded in an optimization framework trading off best expectation at minimal variance, for instance by performing multi-objective (see Sec. 3.5), worst case or minimax optimization, e.g. ??.



**Figure 2.2:** Robustification of a stimulus design by accounting for the parameter PDF. The kinetic parameter  $\theta$  may also represent a design variable and its associated PDF then quantifies variability in the design itself, e.g. variability of the stimulus profile or measurement time points during conduction of the experiment.

Ch. 3 presents a methodology that addresses points (ii) and (iii) of OED robustification for nonlinear models focusing on discriminative stimulus design. Figure 2.2 illustrates this robustification concept with respect to model parameter uncertainties for an optimal experimental stimulus design. One should note that a design based on the objective in Eq. (2.42) will only be optimal on average and one should therefore also have a look at the objective's variance. Therefore, the developed robust experimental design methodology is based on a scalar criterion, where expectation and variance of a design objective are merged into one single scalar - the model overlap. An extension to a multi-objective experimental design is straightforward as is illustrated in the real life application (s. Sec. 3.5).

## 2.5 Summary

In this chapter a survey on the most essential approaches for inferring predictive ODE models has been given. The ODE modeling approach has been illustrated and its extended interpretation from single trajectories to distributed determinism was given. Further structural concepts regarding parameter estimation and model discrimination,

## 2. METHODS FOR IDENTIFYING DYNAMIC MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

which solely rely on model structure and input-output setups were discussed. Given data at hand, it was shown how to assess the quality of parameter estimates and model structures by means of statistical tests. For model discrimination, information-based concepts have also been discussed. Finally, it was discussed how to plan experiments to support the just mentioned inference methods with optimized data. Here, well-established OED methods for the inter-related objective of best parameter estimation and model discrimination have been illustrated. The next chapter focuses on generating data that support model discrimination using the overlap concept.

## 3

# Optimal experimental design in the presence of distributed model parameters

*We are at the very beginning of time for the human race. It is not unreasonable that we grapple with problems. But there are tens of thousands of years in the future. Our responsibility is to do what we can, learn what we can, improve the solutions, and pass them on.*

---

Richard Feynmann  
The Value of Science, 1955

In this chapter, a robust design methodology that allows designing stimulus experiments aimed at best model discrimination taking the model response PDF into account is introduced. After introducing the method, its performance is analyzed by using two *in silico* examples from cell signaling. Further, an application of the method to a real life case is discussed. In this real life case a dynamic model describing DNA double strand break signaling upon ionizing irradiation had to be identified (?). The method has been published in ??. It is based on the model overlap concept (???) and sigma points (?). Results presented in this chapter are taken from ???, so are text passages which have been adopted and modified to fit the presentation in this chapter.

### 3.1 Model overlap as a robust discrimination criterion

Closely related to the modified T criterion is the model overlap, which is a robust discrimination criterion measuring dissimilarities of model response PDFs. It allows

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS

---

estimating the discriminative power of a design for the case of multi-variate and multi-modal PDFs. The *general overlap* shall be defined as the probability product kernel of two multivariate PDFs  $\rho_{\mathbf{Y}|i}^p, \rho_{\mathbf{Y}|j}^p \in L_2(\mathbb{A}_{\mathbf{Y}})$

$$\Phi(t, \mathbf{u}(t)) = \int_{\mathbb{A}_{\mathbf{Y}}} \rho_{\mathbf{Y}|i}(\mathbf{y}, t, \mathbf{u}(t))^p \rho_{\mathbf{Y}|j}(\mathbf{y}, t, \mathbf{u}(t))^p d\mathbf{y} \quad (3.1)$$

with the densities being raised to some power  $p \in \mathbb{R}^+ \setminus \{0\}$  (?). It provides a bounded, positive-definite measure of similarity between distributions on the set  $\mathbb{A}_{\mathbf{Y}}$  (?), whereas the parameter  $p$  controls the weighting of regions with small vs. large densities. This measure is used in vector machine learning to measure statistical distances for the sake of discriminative learning. ? proposed to use Eq. (3.1) with  $p = 1/2$  - known as Bhattacharyya's affinity between distributions - for discriminating nonlinear regression models. In this case ( $p = 1/2$ ), the overlap can be interpreted as the scalar product between two PDFs measuring *cosine* similarity (?). Bhattacharyya's affinity is closely related to Hellinger's distance, which represents a symmetrized approximation to the Kullback-Leibler divergence (?). The general overlap thus comprises several criteria from information theory (?), which all measure (dis)similarity between two PDFs.

Using model response PDFs  $\rho_{\mathbf{Y}|\mathcal{M}}(\mathbf{y}, t, \mathbf{u}(t))$  from Eq. (2.5) for two competing models  $\mathcal{M} = \{i, j\}$  the overlap provides one with a measure of model dissimilarities for a given experimental design. From the general overlap, the average overlap of the time course is

$$\Phi = E[\Phi(\mathbf{D})]_t = \frac{1}{n_t} \sum_{k=1}^{n_t} \Phi(t_k, \mathbf{u}(t_k)), \quad (3.2)$$

where  $\mathbf{D} \in \mathbb{D}$  represents an experimental design point as described in Sec. 2.4.1. For  $p = 1$  the overlap is the expected model response likelihood of model  $i$  with response PDF  $\rho_{\mathbf{Y}|i}(\mathbf{y}, t, \mathbf{u}(t))$  under model  $j$  with  $\rho_{\mathbf{Y}|j}(\mathbf{y}, t, \mathbf{u}(t))$  and vice versa. In this case, assuming one of the models to be true, the overlap yields the expected likelihood of the other model depending on the experimental design  $\mathbf{D}$ . Consequently, an optimal model discrimination design  $\mathbf{D}_{\dagger}$  minimizes Eq. (3.2).

In the following, the overlap as defined in Eq. (3.2) with  $p = 1$  is used. In this work, it is referred to as *model overlap* as this directly represents the time-averaged, expected likelihood of one model under the other. As for the T criterion, assuming one of the models to be true, expected p-values can be derived via a likelihood ratio test to assess whether a new optimized design potentially allows model discrimination. In detail (for two competing models): From the ratio of the estimated model overlap ( $p = 1$ , expected likelihood) at the new design point and model overlap from the initial/previous design can be used to conclude via a likelihood ratio test, if the new design yields significant discriminative information. Alternatively, one may also derive a hierarchy of p-values for each model separately, by comparing initial/previous data-model overlap to the new model overlap, assuming that one model is generating the data, thus equivalent to the true system. The overlap can also be combined with Bayes theorem to derive expected model posteriors, given appropriate model priors.

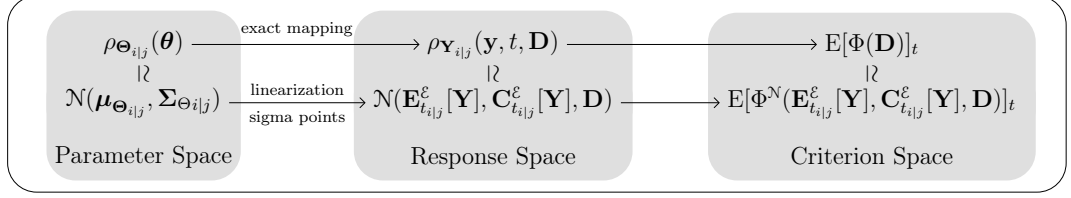


Figure 3.1: Approximation of nonlinear PDF mapping.

## 3.2 Estimation of nonlinear PDF mapping

If the solution  $\mathbf{h}(t, \boldsymbol{\theta})$  to the dynamic systems Eq. (2.1) can be obtained in closed form, it is straightforward to derive the model response PDF for a given parameter PDF using Eq. (2.5). However, in most of the cases the model response  $\mathbf{h}(t, \boldsymbol{\theta})$  for a specific parameter realization is obtained by numerical integration. Here, besides random sampling techniques based on Monte-Carlo simulations, the approximate model response PDF may also be obtained by deterministic sampling, e.g. by enumeration via optimized latin hypercubes of the parameter space and application of Eq. (2.6). For a very large number of samples, the true model response PDF can be constructed from these samples, which can be used for a subsequent evaluation of Eq. (3.2) to judge the quality of a given design. Such procedures become computational inefficient for an increasing number of model parameters and cannot be used in an optimization framework. Therefore we suggested to approximate parameter/response PDFs via Normal PDFs and estimate the Normal response PDFs with the sigma point method, instead of the linearization method.

**1. Normality approximation** From initial data, one may obtain accurate estimates of the true parameter PDF, e.g. by MCMC sampling, which can be approximated by unimodal Normal PDFs  $\rho_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \simeq \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ , possibly multivariate. The model response PDF can also be approximated by

$$\rho_{\mathbf{Y}|\mathcal{M}}(\mathbf{y}, t, \mathbf{u}(t)) \simeq \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{Y}|\mathcal{M}}(t, \mathbf{u}(t)), \boldsymbol{\Sigma}_{\mathbf{Y}|\mathcal{M}}(t, \mathbf{u}(t))\right).$$

Note that for skewed or multimodal PDFs one should apply a transformation, e.g. Box-Cox or inverse hyperbolic sine transformation in order to achieve normality of the PDF (??) or apply Gaussian mixture densities (GMD), e.g. ?. In this way, one is not restricted to normal PDFs and the robustification via the overlap can then account for multi-modalities. The Normality assumption dramatically reduces the computational costs as only the two first statistical moments (i.e. expectation and variance-covariance) need to be estimated. In the case of a GMD representation, one of course has to estimate expectation and variance-covariance for each GMD component. The task of solving Eq. (2.5) to obtain model response PDFs and subsequent integration of Eq. (3.2) to evaluate the discriminative power of a given design in an optimization framework is then reduced

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS

---

to estimating the time course of mean vector  $\boldsymbol{\mu}_{\mathbf{Y}|\mathcal{M}}(t, \mathbf{u}(t))$  and variance-covariance matrix  $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathcal{M}}(t, \mathbf{u}(t))$  of two model response PDFs for given parameter expectation  $\boldsymbol{\mu}_{\boldsymbol{\Theta}|\mathcal{M}}$  and variance-covariance  $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}|\mathcal{M}}$ , see Fig. 3.1. In the following, as true mean and variance-covariance of the parameters are unknown, these are replaced by their sample-based estimates, i.e.  $\boldsymbol{\mu}_{\boldsymbol{\Theta}}$  by  $\mathbf{E}[\boldsymbol{\Theta}]$  and  $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$  by  $\mathbf{C}[\boldsymbol{\Theta}]$ .

**2. Normality estimation** Estimates of response expectation and variance-covariance can be obtained by linearizing the system at additional computational costs that scale linearly with the number of parameters using forward sensitivity analysis. But this approach can become sub-optimal or yield even misleading designs as is illustrated in Sec. 3.4. On the additional expense of  $\mathcal{O}(n_{\theta}^2)$  estimates may be improved by a quadratic response approximation of the system, which may become infeasible for larger systems, as do Monte Carlo based approaches. In ? the sigma point method has been shown to perform very well for experimental design aimed at parameter optimization. Further, as it has an additional computational expense of  $\mathcal{O}(n_{\theta})$  comparable to linearization we chose the sigma point method as an alternative for estimating the response PDF for a robustified discrimination design based on the overlap concept.

#### 3.2.1 Estimation based on linearization

The classical approach to estimate model response variabilities is linearization of the nonlinear model mapping  $\mathbf{h}(t, \boldsymbol{\theta})$  with respect to the parameters. The linearization of the model response is given by applying the chain rule to Eq. (2.2)

$$\mathbf{y}^{\mathcal{L}}(t, \boldsymbol{\theta}) = \mathbf{h}(t, \mathbf{E}[\boldsymbol{\Theta}]) + \mathbf{S}(t, \mathbf{y})^{\text{T}} \Big|_{\boldsymbol{\theta}=\mathbf{E}[\boldsymbol{\Theta}]} (\boldsymbol{\theta} - \mathbf{E}[\boldsymbol{\Theta}]), \quad (3.3)$$

with response sensitivity matrix  $\mathbf{S}(t, \mathbf{y}) = \frac{\partial \mathbf{h}(t, \boldsymbol{\theta})}{\partial \mathbf{x}} \mathbf{S}_{\mathbf{x}}(t, \mathbf{x}) + \frac{\partial \mathbf{h}(t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$  and state sensitivity matrix  $\mathbf{S}_{\mathbf{x}}(t, \mathbf{x}) = \frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}$ , which can be obtained by solving

$$\frac{d}{dt} \mathbf{S}_{\mathbf{x}}(t, \mathbf{x}) \Big|_{\boldsymbol{\theta}=\mathbf{E}[\boldsymbol{\Theta}]} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \mathbf{S}_{\mathbf{x}}(t, \mathbf{x}) \Big|_{\boldsymbol{\theta}=\mathbf{E}[\boldsymbol{\Theta}]} + \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\mathbf{E}[\boldsymbol{\Theta}]} \quad (3.4)$$

with initial condition  $\mathbf{S}_{\mathbf{x}}(t_0, \mathbf{x}_0)$  along the systems dynamics, which is known as the forward sensitivity matrix equation. The additional computational effort is of order  $\mathcal{O}(n_{\theta_{\mathbf{x}}})$ , as  $n_{\theta_{\mathbf{x}}}$  additional ODEs have to be solved in Eq. (3.4), since  $\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}_{\mathbf{y}}} = \mathbf{0}^{n_{\mathbf{x}} \times n_{\theta_{\mathbf{y}}}}$ . One may also formulate an adjoint system to derive the state sensitivities in a backward manner or use numerical differentiation.

Having determined the parameter sensitivities of the system, the linear estimates of expectation and variance-covariances of the model response PDF can be calculated to yield

$$\mathbf{E}_t^{\mathcal{L}}[\mathbf{Y}] = \mathbf{h}(t, \mathbf{E}[\boldsymbol{\Theta}]) \quad (3.5)$$

$$\mathbf{C}_t^{\mathcal{L}}[\mathbf{Y}] = \mathbf{S}(t, \mathbf{y}) \mathbf{C}[\boldsymbol{\Theta}] \mathbf{S}(t, \mathbf{y})^{\text{T}}. \quad (3.6)$$



For nonlinear models, the estimate of the expectation is typically biased, i.e.,  $\mathbf{B}_i = \mathbf{E}_t^{\mathcal{L}}[\mathbf{Y}] - \boldsymbol{\mu}_{\mathbf{Y}} \neq \mathbf{0}$  and errors are introduced at second and higher orders. The quality of the predicted variance-covariance cannot readily be judged as the errors are of fourth and higher order, whereas the contributions depend on the system. Notice that the linear design approach yields a local estimate in the parameter space, i.e. parameter dependent coexisting stable states will be missed, resulting in significant estimation errors in both moments (Sec. 3.4.2). The estimators are exact for linear systems, as higher order terms vanish.

#### 3.2.2 Estimation based on sigma points

? introduced the sigma point method for advanced Kalman filtering and state estimation. It is based on the idea that with a fixed set of parameters (sigma points), it is easier to approximate a nonlinearly transformed PDF by a Gaussian distribution than the nonlinear transformation itself. ? show that expectation and variance-covariance of a random variable  $\mathbf{Y}$ , given by a transformation  $\mathbf{Y} = \mathbf{h}(t, \boldsymbol{\Theta})$ , possibly nonlinear, of a random variable  $\boldsymbol{\Theta}$  with expectation  $\mathbf{E}[\boldsymbol{\Theta}]$  and variance-covariance  $\mathbf{C}[\boldsymbol{\Theta}]$  can be estimated according to the following procedure:

1. Select  $2n_{\boldsymbol{\theta}} + 1$  sigma points in the original domain according to

$$\boldsymbol{\theta}^{(0)} = \mathbf{E}[\boldsymbol{\Theta}]; \quad \boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(0)} \pm \sqrt{n_{\boldsymbol{\theta}} + \lambda} \sqrt{\mathbf{C}[\boldsymbol{\Theta}]}^{(i)},$$

where  $\sqrt{\mathbf{C}[\boldsymbol{\Theta}]}^{(i)}$  is the  $i^{\text{th}}$  column of the square root of the variance-covariance matrix. Further one has  $\lambda = \zeta^2(n_{\boldsymbol{\theta}} + \kappa) - n_{\boldsymbol{\theta}}$ , with tuning factors  $\zeta$  and  $\kappa$  (see below).

2. Propagate these points through the model

$$\mathbf{y}_t^{(i)} = \mathbf{h}(t, \boldsymbol{\theta}^{(i)}).$$

3. Estimated expectation and variance-covariance of the transformed variable based on the sigma points are given by the linearly weighted sums

$$\mathbf{E}_t^{\mathcal{S}}[\mathbf{Y}] = \sum_{i=-n_{\boldsymbol{\theta}}}^{n_{\boldsymbol{\theta}}} w^{(i)} \mathbf{y}_t^{(i)} \tag{3.7}$$

$$\begin{aligned} \mathbf{C}_t^{\mathcal{S}}[\mathbf{Y}] &= (1 - \zeta^2 + \beta) \left( \mathbf{y}_t^{(0)} - \mathbf{E}_t[\mathbf{Y}] \right) \left( \mathbf{y}_t^{(0)} - \mathbf{E}_t[\mathbf{Y}] \right)^{\text{T}} + \\ &+ \sum_{i=-n_{\boldsymbol{\theta}}}^{n_{\boldsymbol{\theta}}} w^{(i)} \left( \mathbf{y}_t^{(i)} - \mathbf{E}_t[\mathbf{Y}] \right) \left( \mathbf{y}_t^{(i)} - \mathbf{E}_t[\mathbf{Y}] \right)^{\text{T}} \end{aligned} \tag{3.8}$$

with weights  $w^{(0)} = \frac{\lambda}{n_{\boldsymbol{\theta}} + \lambda}$ ,  $w^{(\pm i)} = \frac{1}{2(n_{\boldsymbol{\theta}} + \lambda)}$  and additional tuning parameter  $\beta$  (see below).

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS

---

According to ?, the error of the expectation estimate is of fourth and higher order, whereas the variance-covariance estimates have an error of fourth and higher order. This however only holds for scalars, i.e.  $n_\theta = 1$  as pointed out by ?. For  $n_\theta > 1$ , the sigma point parameters  $\zeta, \beta, \kappa$  can be used to tune the estimated moments by including a priori knowledge about the PDFs, i.e.,  $\beta$  and  $\kappa$  allow to account for higher order moments of the parameter PDF and should be set to  $\beta = 2$  for an initial Gaussian, whereas for  $n_\theta > 3$  one should choose  $\kappa = 0$ . Further,  $\zeta$  controls the sigma point spread and should lie within  $0 < \zeta \leq 1$ , ?. The sigma points have several advantages:

- no need to calculate derivative information (neither Jacobian nor Hessian have to be available or need to exist), which makes this method numerically robust and applicable to a wide range of system classes,
- use of curvature information of the system,
- deterministic sampling method with computational effort that scales linearly with the number of distributed variables, i.e.  $\mathcal{O}(n_\theta)$ ,
- since each sigma point is independently propagated, parallelization can easily be applied to speed up estimate calculation of the transformed expectation and variance-covariance.

As has been pointed out by ?, the sigma point method can be understood as a statistical linearization. This corresponds to a nonlocal evaluation of the moment propagation and allows retaining higher order information as is done by Gaussian integration. There exists another approach based on Stirling-Polynoms, which has been independently developed by ? and ?. It is derived from a Taylor expansion of the nonlinear transformation. As shown by ?, both approaches perform equally well, although some slight differences exist, e.g. number of tuning parameters.

### 3.3 Robust optimal stimulus design

The problem of finding an optimal stimulus design can be stated as an optimal control problem. Given a nonlinear dynamic system of the form Eq. (2.1,2.2) and corresponding parameter set (expectation and variance-covariance), an optimal stimulus is an admissible control defined over an interval  $[t_0, t_f]$ , say experimental time window, at which a cost function assumes its infimum (or supremum) with the set of all admissible controls. Robustness of such a control with respect to distributed model responses can be achieved by incorporating expectation and variance-covariance into a robust design criterion (e.g. model overlap). Within the sigma point approach, variabilities in the stimulus conductions can also be accounted by interpreting a design  $\mathbf{u}(t)$  as the time dependent mean of a distributed variable. Then, for a design  $\mathbf{u}(t)$ , model response PDF is determined by the propagation of sigma points given by mean and variance-covariances of (i) model parameters and (ii) stimulus. The problem of finding an optimal control

may be solved by (i) Hamilton-Jacobi-Bellman, (ii) variational, (iii) NLP-based or (iv) flatness-based approaches, e.g. ???. The following two direct NLP-based approaches are used (?), which can easily be combined with the methods discussed in Secs. 3.2.1 and 3.2.2 for mapping distributed parameters onto the design criterion:

- *Direct Sequential Approach*: A control vector parameterization in combination with numerical integration of the model equations. This approach is suited for design problems without nonlinear path constraints and stable behavior with respect to variations in the control and parameters.
- *Direct Simultaneous*: A full discretization of the problem, e.g. control vector and state/response vector parameterization based on orthogonal collocation on finite elements. If the design problem includes nonlinear path constraints, this solution approach can be beneficial, since feasibility of the solution is ensured at the collocation points of each finite element.

Both NLP approaches are typically non-convex, i.e. there exist several local and possibly one global optimal design solution. Therefore, resulting solutions to the NLP problem are local optima. Global optimality of the design can be achieved - but is not ensured - by (i) performing local optimizations from many different initial starting points and/or (ii) deterministic/stochastic/heuristic global optimizers (???). Note that optimal design solutions need not necessarily be global in real life applications. Local optimal solutions can be very close to the global solution with respect to the design criterion. Therefore, non-convexity allows to account for further experimental constraints - restricting the degrees of freedom in the design space - without losing, e.g. discriminative power.

### 3.4 *In silico* results

Results presented in the following two subsections are based on ??.

#### 3.4.1 Benchmark using a signaling cascade

The highly conserved mitogen-activated protein kinase signaling cascade (?) with two different hypothesized negative feedbacks was used as a nonlinear test system for benchmarking the two design approaches with respect to estimation accuracy and design quality. The respective ODE systems - adapted from ? - of two model candidates  $\mathcal{M} \in \{A, B\}$  that describe the change in protein concentration of the phosphorylated forms are

$$\begin{aligned} \frac{d}{dt}x_{1\mathcal{M}}^*(t) &= \frac{k_{1\mathcal{M}}u(t)x_{1\mathcal{M}}(t)}{K_{1\mathcal{M}} + x_{1\mathcal{M}}(t)} - \frac{v_{2\mathcal{M}}x_{1\mathcal{M}}^*(t)}{K_{2\mathcal{M}} + x_{1\mathcal{M}}^*(t)} - r_{1\mathcal{M}} \\ \frac{d}{dt}x_{2\mathcal{M}}^*(t) &= \frac{k_{3\mathcal{M}}x_{1\mathcal{M}}^*(t)x_{2\mathcal{M}}(t)}{K_{3\mathcal{M}} + x_{2\mathcal{M}}(t)} - \frac{v_{4\mathcal{M}}x_{2\mathcal{M}}^*(t)}{K_{4\mathcal{M}} + x_{2\mathcal{M}}^*(t)} - r_{2\mathcal{M}} \\ \frac{d}{dt}x_{3\mathcal{M}}^*(t) &= \frac{k_{5\mathcal{M}}x_{2\mathcal{M}}^*(t)x_{3\mathcal{M}}(t)}{K_{5\mathcal{M}} + x_{3\mathcal{M}}(t)} - \frac{v_{6\mathcal{M}}x_{3\mathcal{M}}^*(t)}{K_{6\mathcal{M}} + x_{3\mathcal{M}}^*(t)} \end{aligned}$$

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS

---

with model  $\mathcal{M} = A$ :

$$r_{1A} = k_{9A}x_4^*(t)x_{1A}^*(t); r_{2A} = k_{10A}x_{3A}^*(t)x_{2A}^*(t)$$

$$\frac{d}{dt}x_{4A}^*(t) = \frac{k_{7A}x_{3A}^*(t)x_{4A}(t)}{K_{7A} + x_{3A}(t)} - \frac{v_{8A}x_{4A}^*(t)}{K_{8A} + x_{4A}^*(t)}$$

and model  $\mathcal{M} = B$ :

$$r_{1B} = \frac{k_{9B}x_{3B}^*(t)x_{1B}^*(t)}{K_{9B} + x_{1B}^*(t)}; r_{2B} = \frac{k_{10B}x_{3B}^*(t)x_{2B}^*(t)}{K_{10B} + x_{2B}^*(t)}$$

no  $x_{4B}(t), x_{4B}^*(t)$ .

For both models it was assumed that

$$x_{i\mathcal{M}}^{\text{tot}}(t) = x_{i\mathcal{M}}(t) + x_{i\mathcal{M}}^*(t)$$

$$x_{i\mathcal{M}}^*(t_0) = 0$$

with the total concentration of each species  $x_{i\mathcal{M}}^{\text{tot}}$  as an additional model parameter and  $i \in \{1, 2, 3, (4)_B\}$ . The measurement response signals were defined as

$$y_{1\mathcal{M}}(t) = x_{2\mathcal{M}}^*(t) + \varepsilon \quad \text{and} \quad y_{2\mathcal{M}}(t) = x_{3\mathcal{M}}^*(t) + \varepsilon, \quad (3.9)$$

where  $\varepsilon$  represents additive measurement noise which is assumed to be normally distributed with zero mean and variance  $\sigma_\varepsilon^2$ . We assumed that the response signals could be measured at  $n_t$  specific time points. Based on an initial stimulus design, the model parameters were adjusted, so that both model responses matched up to a small error, which did not allow preferring one over the other model to mimic the starting point of an OED for model discrimination. Parameter values and further details on the implementation are given in the supplementary material of ?. Since biological systems often follow a log-normal distribution (??), a log-normal transformation to the response was applied in order to improve the Normality approximation used in the estimation approaches for the response PDF (Secs. 3.2.1, 3.2.2). Therefore, the response signal Eq. (3.9) used for the overlap calculation was redefined as

$$\tilde{y}_{i\mathcal{M}}(t) = \log(y_{i\mathcal{M}}(t) + \lambda), \quad (3.10)$$

with  $i = 1, 2, \lambda > 0$ . For each model, the dynamic parameters were assumed to have a log-normal distribution, with nominal value being the expectation  $\mathbf{E}_{\log}[\Theta] = \theta$  and diagonal covariance matrix  $\sqrt{\mathbf{C}_{\log}[\Theta]} = \text{diag}(\eta \mathbf{E}_{\log}[\Theta])$ , with scaling parameter  $\eta$ . The measurement noise is typically independent on the stimulus design, and was thus held constant at  $\sigma_\varepsilon = 0.01$ . The sigma points for the log-normal parameter PDF were obtained in the following way: In the parameter space, the normal equivalents of log-normal expectation and covariance were derived to calculate the normal sigma points, which were then exponentiated. The log-normal sigma points were propagated through the model, including Eq. (3.10) to obtain the normal estimates via Eqs. (3.7,3.8). Further details can be found in the appendix A.2. In the following the tilde is dropped from the redefined response signal in Eq. (3.10).

**Table 3.1:** Monte Carlo verifications of randomly selected stimulus profiles for different  $\eta$  and number of samples  $n$ . Corresponding expectation and standard deviation STD of the overlap (scaled by  $10^5$ ) were taken over 100 runs.

$n$	$\eta = \sigma_i / \mathbb{E}[\Theta_i]$		0.1		0.3		0.4	
	$\mathbb{E}[\Phi^{\mathcal{M}}]$	STD $[\Phi^{\mathcal{M}}]$						
50	$\mathbb{E}[\Phi^{\mathcal{M}}]$	STD $[\Phi^{\mathcal{M}}]$	2	0.05	3.18	0.17	3.24	0.04
100	$\mathbb{E}[\Phi^{\mathcal{M}}]$	STD $[\Phi^{\mathcal{M}}]$	2	0.05	3.17	0.06	3.27	0.04
1000	$\mathbb{E}[\Phi^{\mathcal{M}}]$	STD $[\Phi^{\mathcal{M}}]$	2	0.02	3.14	0.02	3.27	0.02
10000	$\mathbb{E}[\Phi^{\mathcal{M}}]$	STD $[\Phi^{\mathcal{M}}]$	2	0.02	3.15	0.007	3.27	0.005

Following the direct sequential approach, the stimulus (single input) was parameterized as

$$u(\mathbf{U}, t) = u_k \quad \text{for } t_k \leq t \leq t_{k+1}, \quad (3.11)$$

with  $[\mathbf{U}]_k = (u_k, dt_k)^\top$ ,  $k \in \{0, \dots, n_u\}$ , whereas  $dt_{n_u} = 0$ . Here,  $u_k$  represents the amount of stimulus between the time point  $t_k$  and  $t_{k+1} = dt_k + \sum_{j=0}^k t_j$ . If for the last time point  $t_{n_u} < t_f$ , we put  $u(\mathbf{U}, t_{n_u} < t \leq t_f) = u_{n_u}$ . On the other hand, if  $t_{n_u} > t_f$ , the design was given a penalty. Depending on the estimation method  $\mathcal{E} \in \{\mathcal{L} = \text{linearization}, \mathcal{S} = \text{sigma points}\}$  the resulting optimization problem for discriminating between models  $A$  and  $B$  was formulated as an NLP problem, i.e.

$$\mathbf{U}_\dagger^\mathcal{E} = \arg \min_{\mathbf{U} \in \mathcal{UCD}} \Phi^\mathcal{E}(\mathbf{U}) = \mathbb{E} \left[ \Phi^{\mathcal{N}} \left( \mathbf{E}_{t_{A|B}}^\mathcal{E}[\mathbf{Y}], \mathbf{C}_{t_{A|B}}^\mathcal{E}[\mathbf{Y}], \mathbf{U} \right) \right]_t \quad (3.12)$$

subject to systems dynamic, additional constraints and method  $\mathcal{E}$  to estimate  $\mathbf{E}_{t_{A|B}}^\mathcal{E}[\mathbf{Y}]$  and  $\mathbf{C}_{t_{A|B}}^\mathcal{E}[\mathbf{Y}]$ . We further have  $\mathbb{E}[\cdot]_t = \frac{1}{n_t} \sum_{i=1}^{n_t} \cdot_i$ , i.e. average of  $\cdot$  over a set of discrete measurement time points. Further details can be found in the appendix A.1.

The number of optimization parameters was  $n_{\text{utot}} = 39$ , which allowed 20 stimulations  $u_k$  with 19 stimulus durations  $dt_k$ . Since the problem is non-convex, a hybrid optimization strategy, consisting of the evolutionary-based CMA-ES algorithm (?), in combination with a subsequent gradient-based optimizer was used. Owing to the stochastic nature, the hybrid optimization is performed 40 times for each parameter variance level, which is derived from the scaling parameter  $\eta$ . The benchmark is based on a Monte Carlo verification of the resulting optimal stimulus designs. For each optimal design, the overlap, including expectation and variance-covariance of the model responses, was calculated based on sampling the parameter space  $10^4$  times for each model and corresponding optimal stimulus design. The MC sample size was derived by comparing the change in expectation and variance of the overlap for different sample

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS

---

sizes (s. Tabl. 3.1) for a reference design. The relative mean squared error (MSE) of the moment estimate is given by

$$e_{\mathbf{M}}^{\mathcal{E}} = \frac{1}{2n_t n_{\mathbf{y}}} \sum_{\mathcal{M}=A,B} \sum_{i=1}^{n_{\mathbf{y}}} \sum_{t=1}^{n_t} \left( \frac{\mathbf{M}^{\mathcal{E}} - \mathbf{M}^{\mathcal{MC}}}{\mathbf{M}^{\mathcal{MC}}} \right)^2 \quad (3.13)$$

with  $\mathbf{M}^{\mathcal{E}}$  being the moment estimates of the best designs (expectation  $\mathbf{E}_{t_{A|B}}^{\mathcal{E}}[Y_i]$ , variance-covariance split into variance  $\mathbf{VAR}_{t_{A|B}}^{\mathcal{E}}[Y_i]$  and covariance terms  $\mathbf{COV}_{t_{A|B}}^{\mathcal{E}}[Y_i]$ ). The Monte Carlo reference is represented by  $\mathbf{M}^{\mathcal{MC}}$ .

Table 3.2 illustrates that for all parameter variance levels, both methods have negligible relative MSE in the mean response estimates (maximal MSE:  $e_{\mathbf{E}}^{\mathcal{E}} < 10^{-7}$ ;  $e_{\mathbf{E}}^{\mathcal{S}} < 10^{-9}$ ). In contrast, the relative MSEs for linearization increases with the parameters variance levels up to 0.03 for the variance and 0.18 for the covariance estimates. Here, the sigma point approach performs better with maximal relative MSE of the variance estimate 0.007 and covariance estimate 0.096. In this application both approaches estimate mean responses of the models very well, although the maximal MSE of the sigma points is still two orders of magnitudes smaller than the maximal MSE for linearization. For the (co)variance estimates, the sigma point approach consistently outperforms the linearization approach for increasing parameter variance level.

In the lower part of Tab. 3.2 discriminative powers of the resulting designs for different parameter variance levels are compared. From the Monte Carlo verifications ( $\mathcal{E} = \mathcal{MC}$ ), it is apparent that for small variances, both methods yield designs that have the same discriminative power ( $\Phi_{\dagger}^{\mathcal{E}\mathcal{MC}} \equiv \Phi^{\mathcal{MC}}(\mathbf{U}_{\dagger}^{\mathcal{E}})$  vs.  $\Phi_{\dagger}^{\mathcal{S}\mathcal{MC}} \equiv \Phi^{\mathcal{MC}}(\mathbf{U}_{\dagger}^{\mathcal{S}})$ ). However, for widely distributed parameters (starting at  $\eta = 0.3$ ) sigma point based designs perform up to 1.3 times better than linearization-based designs and their estimates coincide with the MC validation, which is not the case for linearization. For both methods, optimization time for one design is  $1.3 \pm 0.1$  h on a standard desktop computer (4 GB ram, 3 GHz quad core processor), whereas the validation time ( $10^4$  MC samples) of a single optimal design is  $0.4 \pm 0.1$  h.

#### 3.4.2 Bistable system

The Schlögl model is a canonical example of a biochemical system exhibiting bistability (?). It describes an autocatalytic, tri-molecular reaction, which may occur in biochemical systems such as cell metabolism or signaling. Two model alternatives  $\mathcal{M} = \{A, B\}$  for the rate of concentration change of specie  $x$  are given by

$$\frac{d}{dt}x_{\mathcal{M}}(t) = k_1 a s_{\mathcal{M}}(u(t))x_{\mathcal{M}}^2(t) - k_2 x_{\mathcal{M}}^3(t) - k_4 x_{\mathcal{M}}(t) + k_3 b, \quad (3.14)$$

$$s_A(u(t)) = u(t) \quad \text{or} \quad s_B(u(t)) = \frac{1}{2}(u(t) + u^2(t)) \quad (3.15)$$

and distributed initial condition  $X_0 \propto \mathcal{N}(\mathbf{E}[x_0], (\eta\mathbf{E}[x_0])^2)$ . The model alternatives differ in the input layer  $s_{\mathcal{M}}(u(t))$  (scaled to relative units by an arbitrary reference stimulus

**Table 3.2:** Relative MSEs of moment estimates and overlap (scaled  $10^5$ ) of the best designs based on linearization/sigma point estimation and corresponding Monte Carlo verification.

$\eta = \sigma_i/E[\theta_i]$	0.01		0.1		0.2		0.3		0.4	
$e_{\mathbf{E}}^{\mathcal{L}}$	0	0	0	0	0	0	0	0	0	0
$e_{\mathbf{E}}^{\mathcal{S}}$	0	0	0	0	0	0	0	0	0	0
$e_{\text{VAR}}^{\mathcal{L}}$	0	0	0.002	0	0.001	0	0.01	0.007	0.03	0.007
$e_{\text{VAR}}^{\mathcal{S}}$	0	0	0.002	0	0.001	0	0.01	0.007	0.03	0.007
$e_{\text{COV}}^{\mathcal{L}}$	0	0	0.02	0.002	0.059	0.015	0.106	0.046	0.181	0.096
$e_{\text{COV}}^{\mathcal{S}}$	0	0	0.02	0.002	0.059	0.015	0.106	0.046	0.181	0.096
$\Phi_{\dagger}^{\mathcal{L}}$	0.2	0.2	2	2	2	2	2	3	3	3
$\Phi_{\dagger}^{\mathcal{S}}$	0.2	0.2	2	2	2	2	3	3	4	3
$\Phi_{\dagger}^{\mathcal{LMC}}$	0.2	0.2	2	2	2	2	3	3	4	3
$\Phi_{\dagger}^{\mathcal{SMC}}$	0.2	0.2	2	2	2	2	3	3	4	3

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS

---

level). Parameters  $a$  and  $b$  represent the concentration of two reaction partners  $a$  and  $b$  of specie  $x$ , which both are assumed to be in constant exchange with a material bath. The parameter values are taken from ?. For an initial, sub-optimal experiment with stimulus  $u(t) = 1$ , models  $A$  and  $B$  cannot be distinguished, given  $y_{\mathcal{M}}(t) = x_{\mathcal{M}}(t) + \varepsilon$  to be the response signal. Additionally to a distributed initial condition  $x_0$ , constant additive measurement noise with zero mean and  $\sigma_{\varepsilon}^2 = 0.1$  was assumed. The stimulus is thought to control the concentration in the reservoir of species  $a$  to find an optimal discriminative stimulus.

In a first analysis, the stimulus is parameterized as

$$u(t) = u_1 (H[u_2 - t] + (1 - H[2u_2 - t]H[3u_2 - t])); \quad H[t] = \begin{cases} 1, & \text{if } t \leq 0 \\ 0, & \text{else,} \end{cases} \quad (3.16)$$

which allows analyzing the overlap landscape as a function of  $u_1$  and  $u_2$  (see Fig. 3.2). Both parameters are subjected to box constraints mimicking experimental limitations, i.e.  $u_{i|lb} \leq u_i \leq u_{i|ub}$ ,  $i = \{1, 2\}$ .

In Fig. 3.2A the design dependent overlap landscapes, Eq. (3.2), for  $n_t = 30$  and different levels of parameter variances derived from  $\sigma_{x_0} = \eta E[x_0]$  are shown. For small parameter variances  $\eta = 0.01$ , the overlap landscapes of linearization, sigma point and Monte Carlo estimations coincide. In this case, all methods would yield the same robust optimal design  $\mathbf{D}_{\dagger}^{\mathcal{E}}$ , again with  $\mathcal{E} = \mathcal{L}$  linearization,  $\mathcal{E} = \mathcal{S}$  sigma points and  $\mathcal{E} = \mathcal{MC}$  Monte Carlo. When the uncertainty of the initial condition is increase ( $\eta = 0.3$ ), the overlap landscapes differ quantitatively for all estimation methods. However, the optimal designs still coincide (optimal design point at the maximum of the (negative) overlap landscape). For  $\eta = 0.6$  overlap landscapes and the best design points differ between linearization and sigma point predictions, i.e.  $\mathbf{D}_{\dagger}^{\mathcal{L}} \neq \mathbf{D}_{\dagger}^{\mathcal{S}}$ . From the Monte Carlo reference, one has  $\mathbf{D}_{\dagger}^{\mathcal{S}} = \mathbf{D}_{\dagger}^{\mathcal{MC}}$ , i.e. the optimal design point estimated with the sigma points coincides with Monte Carlo prediction. Further, the predicted value of the overlap at the optimal point agree, which is not the case for the linearization estimate. In panel B in Fig. 3.2 time courses of the estimated moments are shown for the best designs  $\mathbf{D}_{\dagger}^{\mathcal{E}}$ . Evaluating the discrimination performance of the optimal linear design with sigma points and Monte Carlo reveals a larger expected overlap  $\Phi_{\dagger}^{\mathcal{LMC}} = \Phi_{\dagger}^{\mathcal{LS}} > \Phi_{\dagger}^{\mathcal{L}}$ . In contrast to the sigma point approach, linearization cannot capture two coexisting states at the same time. As a result, linearization misses the true location of the expected response, due to the existence of a second stable state. Further, the time course prediction of the sigma points for the linear optimal design almost matches the Monte Carlo reference. Evaluating the discrimination performance of the optimal sigma point design with linearization and Monte Carlo reveals good agreement between sigma points and Monte Carlo, whereas linearization overestimates the variance at early time points. Here again, the predicted overlap coincide for sigma points and Monte Carlo.

The reason why linearization fails in this case lies in the local property of this approach. The optimal sigma point design minimizes the overlap by switching optimally

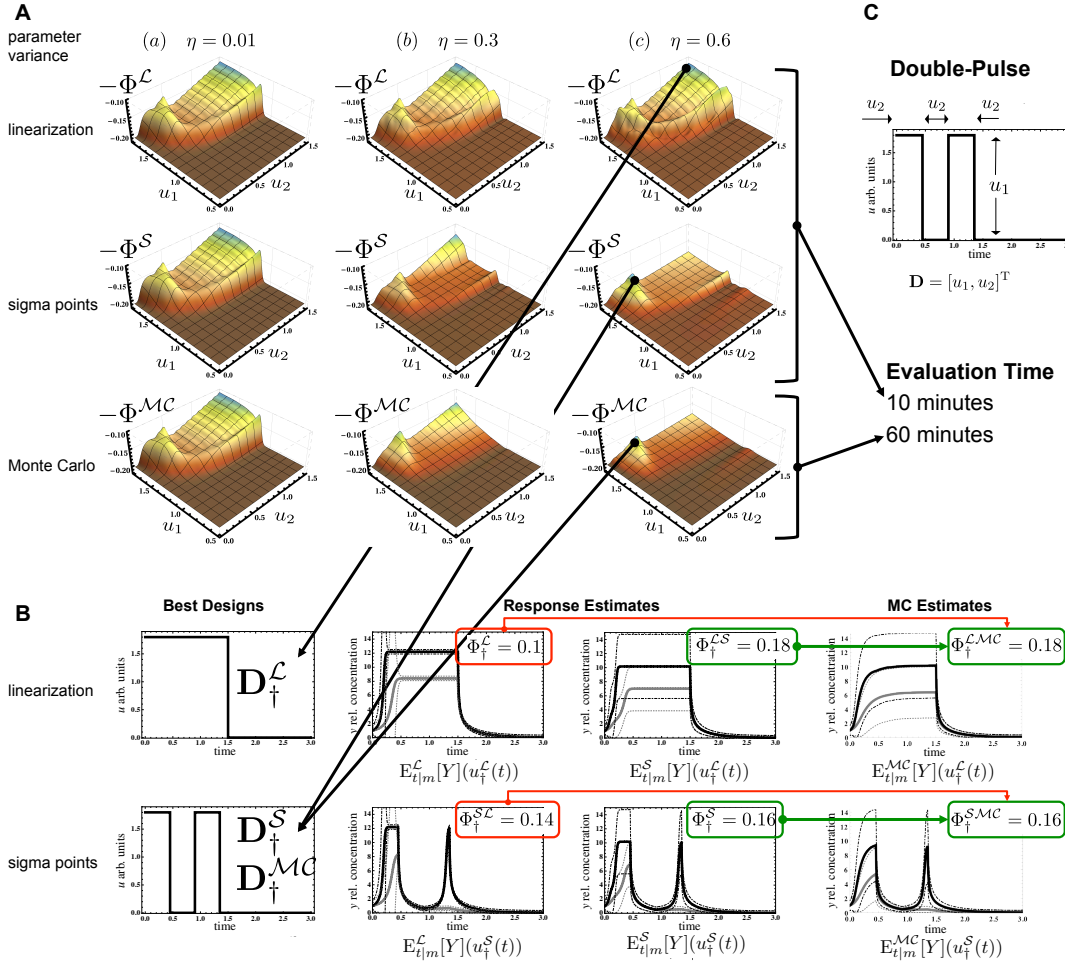


between the two stable states while keeping the variance minimal, whereas the linear design misses the bistability.

The next analysis illustrates the differences between linearization and sigma point approach when deriving a more complex or flexible stimulus design allowing more or less continuous stimulus changes over time. In this example, nonlinear constraints to account for possible control limitations have been included. In detail, it was assumed that subsequent stimulations can be applied, but a minimal time period has to pass in between. Such nonlinear constraint optimization can efficiently be solved within the direct simultaneous approach, which here was applied using orthogonal collocation on 100 finite elements (each with 3 collocation points) to discretize control and system states. The objective of the resulting non-convex NLP problem was the same as in Eq. (3.12), however, subjected to different constraints, i.e. system dynamics in form of a nonlinear algebraic equation system and additional constraints. For details see appendix A.1. For the linear design strategy, sensitivity equations (Eq. (3.4)) were implemented. For the sigma point design, constraints have to simultaneously hold for all  $(2n_{\theta} + 1)$  sigma points. The solver AMPL in combination with the optimizer CONOPT was used to solve the above NLP problem (?). For a given optimization setup ( $\eta$  and estimation method) the solution took about 2 minute on a standard desktop computer. Since CONOPT yields local solutions, the optimization was performed for 1000 different randomized initial designs for a given optimization setup, from which the best solution was selected.

In Fig. 3.3 the resulting stimuli designs for  $\eta = 0.35$  based on linearization, sigma point estimation and corresponding Monte Carlo reference simulations are shown. Re-examination of the optimized linear design with MC simulations revealed a large underestimation of the estimated overlap:  $\Phi_{\dagger}^{\mathcal{L}} = 0.004$  vs.  $\Phi_{\dagger}^{\mathcal{L}MC} = 0.17$ , i.e. misleading discriminative power by 2 orders of magnitude. The local estimation property of the linear approach yields a highly biased expectation and underestimation of the variance with relative MSE of 0.44 for the expected response and 6.66 for the variance (see Fig. 3.3, estimated response of model  $B$ , (B) vs. (C)). The sigma point based design (D) in Fig. 3.3 has a relative MSE of 0.15 for the expected response and 0.38 for the variance. The overlap estimate of the sigma point design closely matches the MC validation ( $\Phi_{\dagger}^{\mathcal{S}} = 0.04$  vs.  $\Phi_{\dagger}^{\mathcal{S}MC} = 0.03$ ). Further, the sigma point design performs 5.7 better than the linear design  $\Phi_{\dagger}^{\mathcal{S}MC} = 0.03$  vs.  $\Phi_{\dagger}^{\mathcal{L}MC} = 0.17$ ). As can be seen in Fig. 3.3, the non-local propagation property of the sigma points enables the optimizer to find a stimulus that drives the expected response of model  $B$  to the upper steady state. This behavior was already observed for the simpler parameterization of the stimulus in Eq. (3.16) (estimated time courses in Fig. 3.2B vs. Fig. 3.3). However initially, in the more flexible parameterization, the optimal stimulus profile first moves both models into the lower stable state by simply not stimulating at all, to then push the expected response of model  $B$  to the upper stable state. This is due to the fact that the simpler parameterization has to immediately stimulate the models at the beginning. From the model equations (3.14,3.15), this behavior seems also plausible, since any stimulus  $u > 1$  will

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS



**Figure 3.2:** (A) Comparison of estimated overlap landscape over the design space for different noise levels (a)-(c) and estimation methods ( $\mathcal{L}$ =linearization,  $\mathcal{S}$ =sigma points,  $\mathcal{MC}$ =Monte Carlo with  $10^3$  samples for 837 design point evaluations). The approximate evaluation times for each landscape are obtained with a 1.6 Ghz Intel Core i5, 4 GB at 1.333 MHz ram speed. In panel (B) the best designs predicted from linearization/sigma points and the corresponding time courses of the model predictions (model *A* solid gray lines, model *B* solid black lines) with 95% confidence bands (dashed lines) are indicated for linearization, sigma point and Monte Carlo estimates. The Monte Carlo simulations are also given for each design so are the estimated overlaps. Panel (C) illustrates the stimulus parameterization. Figure adapted from ?.

have a higher impact on model  $B$  than on model  $A$ , simply due to the quadratic impact on the right hand side function in Eq. (3.14). Note also that owing to a more flexible stimulus design compared to Eq. (3.16), the predicted overlap is smaller.

From these two examples one can see the benefit of the nonlocal estimation property of the sigma point method. The optimizer is aware of the bistability and uses it to separate the two model responses for optimal discrimination. As linearization and sigma points have the same computational effort, the sigma point method should be favored providing a powerful estimation method when designing discrimination experiments for complex, nonlinear biochemical ODE models having widely distributed parameters and associated multiple stable states, even though the estimation of the overlap is restricted to expectation and variance.

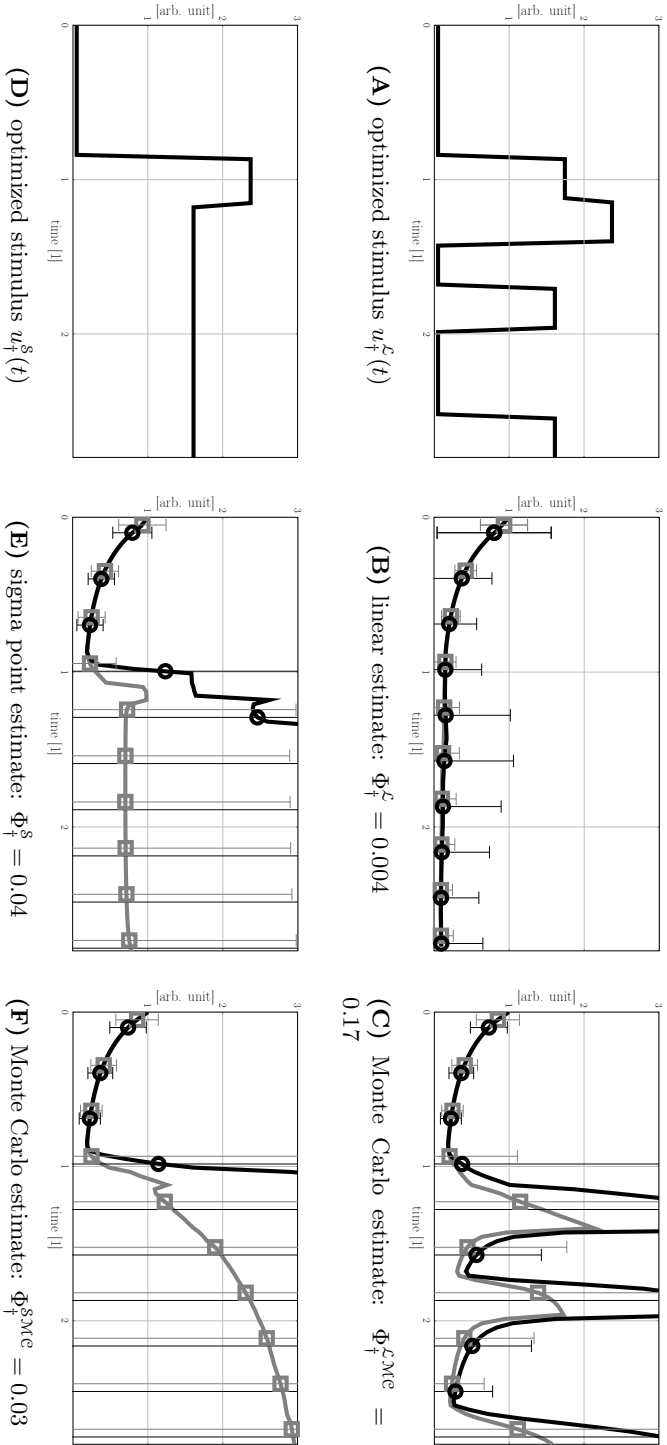
An application that extends this approach to approximate multi-modality of the response PDF via GMDs is given in the Bachelor thesis of ?, which we have supervised during my Ph.D. phase. In another Bachelor thesis, we have investigated the hierarchy amongst shape classes of parameterized stimuli including ramp, step, double-step, pulse and double-pulse profiles (?). This investigation hints at the following: best discrimination is achieved with stimuli that have a bang-bang like characteristics, i.e. the stimulus is either on its upper or lower bounds. Bang-bang like solutions to an optimal control problem can result from the Hamiltonian being linear in the control (details see for instance ?). This was the case in ?. The bang-bang property in optimal control can be related to optimal designs in linear regression. For linear regression, best results are achieved when placing the design points at their boundaries.

### 3.4.3 Summary robust optimal experimental design methodology

Using the overlap, discrimination of different models with distributed parameters is based on comparing model response PDFs. Here we explored the use of the sigma point method for estimating model response PDFs. For benchmarking the performance of the sigma points, the response PDF was derived with the sigma point method and was compared against the classical linearization approach in terms of estimation accuracy of the nonlinearly propagated parameter PDFs. Both approaches were presented and compared owing to the same linear scaling of the additional numerical costs with respect to the number of distributed parameters. The comparison in the light of optimal experimental stimulus design for robust model discrimination is based on Monte Carlo validation of the predicted optimal designs assuming perfect experimental conduction, i.e. no design variability. As noted above, for a given distribution of design variability, the applied sigma point method may also be used. In the next section, the presented stimulus design methodology is applied in a sequential experimental design approach for model identification.

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS

---



**Figure 3.3:** Comparison of robust stimulus design. (A)-(C) Optimal robust stimulus derived from linearization  $u_f^L(t)$ , corresponding linear and MC estimates of expected responses of models A (gray line, square) and B (black line, circle) for  $\eta = 0.35$  with indicated standard deviation bands. (D)-(F) Sigma point based results. Figure adapted from ?.

### 3.5 A real life application

This section exemplifies a sequential experimental design, which has been performed in an iterative work between experiments and modeling. This *in vitro* study is applied on DNA damage signaling, which is of utmost importance for understanding how cells maintain genomic integrity and how cancer therapy by means of  $\gamma$ -irradiation can be optimized. Here, the modified T criterion Eq. (2.38) was used, as it directly yields a statistical measure for the discriminative power of the optimal design. For one-dimensional responses it is equivalent to the overlap up to some constants (??). The following results have been submitted for publication in ?.

#### 3.5.1 Background of the *in vitro* application

Living cells are constantly affected by DNA damage, resulting from ionizing  $\gamma$ -irradiation (IR), genotoxic or replication stress and reactive oxygen species. DNA damage, including single and double strand breaks (DSB), base modification, deletions or point mutations, seriously affects genome stability and cell integrity if not properly detected and repaired by the DNA damage response (DDR) (?). Upon DNA damage, higher order chromatin has to be made accessible by various modifications before DSB can be detected and repaired (?). Among several DNA-damage associated histone modifications, phosphorylation of H2AX is widely accepted as an indicator of DSB. H2AX becomes rapidly phosphorylated at serine 139 ( $\gamma$ H2AX) to generate foci at the DSB site (?). The assembly of chromatin remodeling complexes at the DSB site greatly depends on  $\gamma$ H2AX and enables the accessibility of the damaged DNA to repair proteins (?). Depending on the stimulus,  $\gamma$ H2AX is induced by different members of the phosphoinositide 3-kinase like kinase (PIKK) family; ataxia telangiectasia mutated (ATM), ataxia telangiectasia and Rad3-related (ATR) and DNA-dependent protein kinase catalytic subunit (DNA-PK<sub>cs</sub>). ATR phosphorylates H2AX upon replicative stress (?), whereas ATM and DNA-PK<sub>cs</sub> are responsible for this phosphorylation upon DSB induced by IR (?). The interplay between ATM and DNA-PK<sub>cs</sub> in IR-induced H2AX phosphorylation remains puzzling because although ATM is required (?), DNA-PK<sub>cs</sub> can substitute for it (?). Here we analyzed the interplay of DNA-PK<sub>cs</sub> and ATM to the phosphorylation of histone H2AX with a computational model. This enabled us to look at the dynamics of the very first minutes post irradiation damage without the need for direct measurements of protein activities, reducing confounding effects from experimental manipulations. The challenge of informative experiments with respect to proteins of interests is therefore shifted to the challenge of generating a predictive dynamic model.

? have created a dynamic model solely focused on DNA-PK<sub>cs</sub> to predict dose and dose-rate effects on  $\gamma$ H2AX dynamics. ATM dynamics in the context of DNA damage has been modeled, albeit on theoretical grounds (?). A dynamic model for DNA-PK<sub>cs</sub>/ATM interactions with regard to  $\gamma$ H2AX activation integrating biochemical time course data was missing. Therefore, an iterative workflow was established to identify a predictive dynamic model involving DNA-PK<sub>cs</sub>/ATM mediated H2AX phosphorylation.

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS

---

Starting from several models, optimal experimental design was applied to optimize experiments for model identification. The identified model was used to analyze the dynamic contribution of ATM and DNA-PK<sub>cs</sub> to H2AX phosphorylation.

The following subsections show results regarding (i) model identification workflow, (ii) model analysis and (iii) model predictions. As stated in the beginning of this section, results including figures and tables are taken from ?. For reasons of condensed and focused presentation, details on data processing procedures, data in raw and processed form as well as experimental materials and methods are not included herein, but can be found in ? and supplementary information thereof.

#### 3.5.2 Model identification

**Defining network structures for  $\gamma$ H2AX activation upon IR** The network structures (Fig. 3.4A) have been constructed based on meta-analysis (????) focusing on the initial activation dynamic within the nucleus. DDR initiates with recognition of damaged DNA (DDNA1). Ku7080 as a sensor for non-homologous end joining (cNHEJ) associates to the damage site (RC11) forming the DNA-PK complex (RC12). Then, the catalytic subunit of DNA-PK is either phosphorylated by active ATM or/and autophosphorylated at the T2609 cluster to initiate cNHEJ (?). The MRN complex (Mre11-Rad50-Nbs1), a sensor for homologous recombination repair pathway (HR), can also co-localize to the damage site to promote ATM autophosphorylation at Ser1981. Failure of DNA repair via cNHEJ potentially allows HR proteins to access the damage site. This is modeled by splitting the initial DSB pool (DDNA) into DDNA1 and DDNA2, whereby DDNA2 is associated to HR and/or alternative non-homologous end joining (aNHEJ) (?). Phosphorylation of H2AX can be achieved by active DNA-PK<sub>cs</sub> or active ATM.

Four dynamic models in the form of ordinary differential equation systems were derived from the network structures in Fig. 3.4A describing various interplays between ATM, DNA-PK<sub>cs</sub> and  $\gamma$ H2AX. The ODE systems were implemented in MATLAB using the solver CVODES (?). The equation systems and further details on the choice of kinetic rate laws are given in the appendix A.3. After the poor discrimination performance of OED 0 and OED I data (see following paragraph), the models were extended to also describe p53 activation dynamics. The tumor suppressor p53 is an important effector protein during DDR. Phosphorylation of p53 at Ser15 by ATM promotes its release from MDM2 and results in p53 activation (?). Activation of p53 by DNA-PK<sub>cs</sub> has also been described (?). However, DNA-PK<sup>-/-</sup> MEFs show normal p53 activation (?). In our study (?), evidence for a DNA-PK<sub>cs</sub> contribution to the p53 phosphorylation was not found. This agrees with earlier data (?). Therefore, p53 activation was implemented as an ATM-dependent process only. As described in detail in the appendix A.3, 19 kinetic and 8 scaling parameters were estimated by maximizing the likelihood function, whereas variances have been estimated from data replicates. Parameter estimation was performed for each model in an iterative manner, according to the 3 datasets

(OED 0/I/II). Optimization of the likelihood function was performed iteratively, using a hybrid strategy. A genetic algorithm from the global optimization toolbox of [?](#), which was used to obtain a population of suitable starting solutions for a local optimizer, was combined with a gradient-based optimization.

**Experimental design for model identification** For model calibration purpose, an initial time course of H2AX phosphorylation in response to IR was studied in MDCK cells in a dose-dependent manner using 0.5, 1, 2, 5, 40 Gy.  $\gamma$ H2AX levels increased with IR dose, while concurrently signal attenuation was delayed (see Fig. 3.4B). These results agree with data from [?](#). From the competing network structures, we derived ordinary differential equation models and calibrated them as described above. Simulations of the initial data set for all models are shown in Fig. 3.4C. Based on  $\chi^2$  statistics, none of the models could be rejected at a significance level of  $\alpha_{0.05} = 0.05$  (Table 3.3, OED 0). P-values of Anderson-Darling (AD) residual statistics also indicated that all models seemed adequate for the initial data. In [?](#) we have shown that pulse and double-pulse profiles are powerful stimulus profile classes providing discriminative power that is comparable to more complex stimulus, however at moderate experimental effort. Therefore, to discriminate between models, we subsequently designed (i) an IR double-pulse (Fig. 3.5 A-D) and (ii) an IR double-pulse in combination with kinase inhibitors (Fig. 3.6). The IR double-pulse design  $\mathbf{D} = [D_1 D_2]^T$  was parameterized with 2 design variables, namely inter-pulse time  $D_1$  and second pulse dose  $D_2$ , where the first pulse was fixed at 1 Gy (Fig. 3.5A). The measurement time points were fixed to  $\mathbf{t} = [0 15 35 60 160 240 370 420 450]^T$  minutes for all subsequent designs. The first 6 time points were chosen from simulating OED 0 conditions to fully capture rising and falling flanks of the initial  $\gamma$ H2AX peak, whereas the remaining time points were placed based on the estimated second signal peak. The objective was to maximize  $O = [\mathbf{T}_{\text{red}} \langle V \rangle \langle S \rangle]^T$ . Herein  $\mathbf{T}_{\text{red}}$  is the reduced, modified T criterion of Eq. (2.38) to measure the average discriminative power along all model pair combinations ([?](#)), whereas  $\langle V \rangle$ ,  $\langle S \rangle$  represent mean model prediction variance and variance-entropy. The latter two criteria measure parameter information and distribution within the  $\gamma$ H2AX signal via

$$\langle V \rangle = \frac{1}{n_t n_{\mathcal{M}} n_y} \sum_{i=1}^{n_t} \sum_{j=1}^{n_{\mathcal{M}}} \sigma_{\text{sim},j}^2(t_i, \mathbf{D}) \quad (3.17)$$

$$\langle S \rangle = \sum_{j=1}^{n_{\mathcal{M}}} \sum_{i=1}^{n_t} -\tilde{\sigma}_{\text{sim},j}^2(t_i, \mathbf{D}) \log \tilde{\sigma}_{\text{sim},j}^2(t_i, \mathbf{D}), \quad (3.18)$$

with  $\sum_{j=1}^{n_{\mathcal{M}}} \sum_{i=1}^{n_t} \tilde{\sigma}_{\text{sim},j}^2(t_i, \mathbf{D}) = 1$ . For OED I (Fig. 3.5), the optimal design  $\mathbf{D}_I^\dagger$  was chosen by trading off maximal  $\mathbf{T}_{\text{red}}$ ,  $\langle V \rangle$  and  $\langle S \rangle$  (Fig. 3.5B). Recalibration of all models to data from OED 0 and I, and additional inclusion of p53-P data (Fig. 3.5E) from titration experiments did not allow for model discrimination (all p-values  $> \alpha_{0.05}$  for both fit statistics; Tab. 3.3), but reduced prediction variances (Tab. 3.4).

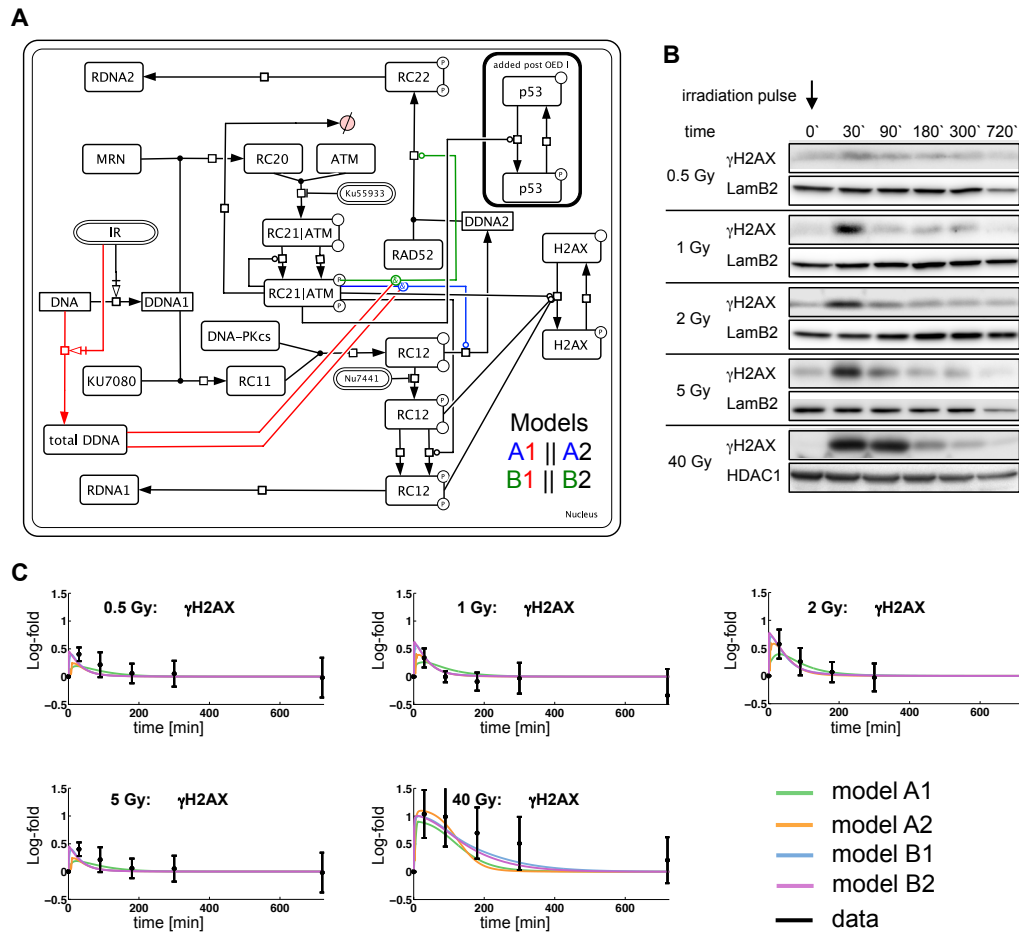
### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS

---

**Table 3.3:** Fit statistics for initial (OED 0) and optimized experiments (OED I and II) Anderson-Darling p-values are indicated as AD.  $AD_{3\sigma}$  indicates p-values of AD statistics where residuals larger than  $3\sigma$  have been excluded. The number of data points  $n_{data}$  do not include the time point  $t=0$  [min].  $n_{\theta}$  and  $n_s$  indicate the number of estimated kinetic and scaling parameters.

OED	$N_{data}$	$n_{\theta}$	$n_s$	Fit Statistics	Model A1	Model A2	Model B1	Model B2
0	114	19	2	$\chi^2$	93.45	91.74	92.79	91.69
				p-value $\chi^2$	4.09E-01	4.59E-01	4.28E-01	4.60E-01
				p-value $AD_{3\sigma}$	3.44E-02	1.21E-02	3.04E-02	2.32E-02
I	147	19	7	p-value AD	3.44E-02	1.21E-02	3.04 E-02	2.32E-02
				$\chi^2$	135.98	131.53	125.84	125.64
				p-value $\chi^2$	1.37E-01	2.04E-01	3.16E-01	3.21E-01
				p-value $AD_{3\sigma}$	1.38E-01	1.84E-01	9.22E-02	5.64E-02
				p-value AD	2.12E-01	1.84E-01	9.22E-02	5.64E-02
				$\chi^2$	290.6	208.2	286.22	479.1
II	237	19	8	p-value $\chi^2$	1.35E-04	4.83E-01	2.60E-04	0
				p-value $AD_{3\sigma}$	1.97E-05	6.52E-02	3.11E-02	1.12E-01
				p-value AD	3.86E-08	5.22E-29	3.21E-32	5.46E-14





**Figure 3.4:** Network structures and initial data (OED 0). (A) The network structures of four different models are shown as an interaction graph. Interactions are modeled via state transitions (arrows with squares), enzyme catalysis (lines with circles) and complex formation (joined lines). Stimulus and inhibitors have round-edge boxes. Four mechanisms have been considered for branching (A1, A2, B1, B2). A and B refer to the location of the catalytic activity of ATM and index 1 and 2 refer to the kinetic law used. For index 1 branching to DDNA2 is catalyzed by the total amount of damaged DNA. Index 2 does not use the total amount of damaged DNA. (B) MDCK cells were irradiated with different doses and the insoluble nuclear extracts were analyzed by immunoblot. Lamin B2 or HDAC1 served as loading control. (C) Model simulation and quantified experimental data for OED 0 using the estimated band intensities of  $\gamma$ H2AX. Data represent mean  $\pm$  2STD of 3-5 independent experiments. Reproduced from ?.

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS

---

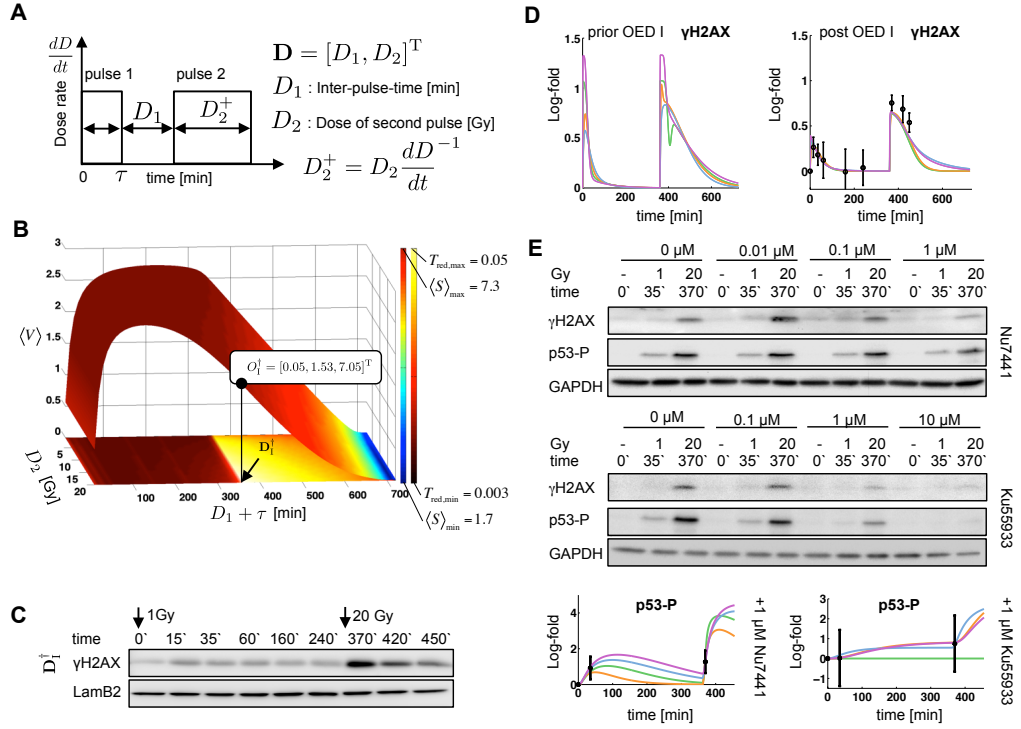
**Table 3.4:** Design criteria for the experimental runs, OED 0 (initial), and optimized OED I, II. For details on the criteria see text.

Criterion		OED I				OED II			
		prior OED I		post OED I		prior OED II		post OED II	
T	$T_0$	107.13	6.5	45.1	0.3	4.6E03	44.7	1.5E03	51.5
$T_{\text{red}}$	$T_{\text{red},0}$	0.05	3E-3	0.02	1E-04	28.2	0.3	9.3	0.3
$\langle V \rangle$	$\langle V \rangle_0$	1.53	4E-08	0.52	2E-07	2.2	6E-08	0.6	1e-05
$\langle S \rangle$	$\langle S \rangle_0$	7.05	2.26	7	2.29	20.1	7.5	5.1	3.1

Kinase inhibitors were employed for OED II to better dissect DNA-PK<sub>cs</sub> and ATM contributions. Titration of two highly specific inhibitors, namely Nu7441 and Ku55933 for DNA-PK<sub>cs</sub> and ATM, respectively, identified the optimal concentration for each. Further, we used the phosphorylation of p53 at S15 as a read-out to show the specificity of the inhibitors. Two successive pulses with different intensities (1 and 20 Gy) show in the immunoblot that the contribution of DNA-PK<sub>cs</sub> to this particular phosphorylation of p53 is marginal (Fig. 3.5E). This confirms earlier data (??). OED II was designed for three different inhibitor settings, namely Nu7441 and/or Ku55933 (Fig. 3.6). The estimated optimal design  $\mathbf{D}_{\text{II}}^{\dagger}$  potentially allowed for discrimination (Tab. 3.4, T criterion for OED II; Fig. 3.6). The initial  $\gamma$ H2AX peak showed a comparable reduction for both inhibitors. Phosphorylation of H2AX after the second pulse seemed to decay more rapidly for inhibited ATM compared to inhibited DNA-PK<sub>cs</sub>. Both inhibitors together showed synergistic effects on  $\gamma$ H2AX (Fig. 3.6B). According to the fit statistics of OED II (Tab. 3.3) only model A2 could not be rejected in terms of  $\chi^2$ . However, we found significant AD p-values for all four models, whereas models A2 and B2 had non-significant AD<sub>3 $\sigma$</sub>  p-values, which account only for residuals smaller than 3 $\sigma$ . This behavior may be attributed to outliers in one of the experimental conditions (Fig. 3.6C) owing to experimental variations or deficits of the models in describing experimental conditions of OED II. We selected model A2 as the final model for further analysis, since it was the only model with p-values of  $\chi^2$  and AD<sub>3 $\sigma$</sub>  statistics exceeding  $\alpha_{0.05}$  for all 3 experimental runs.

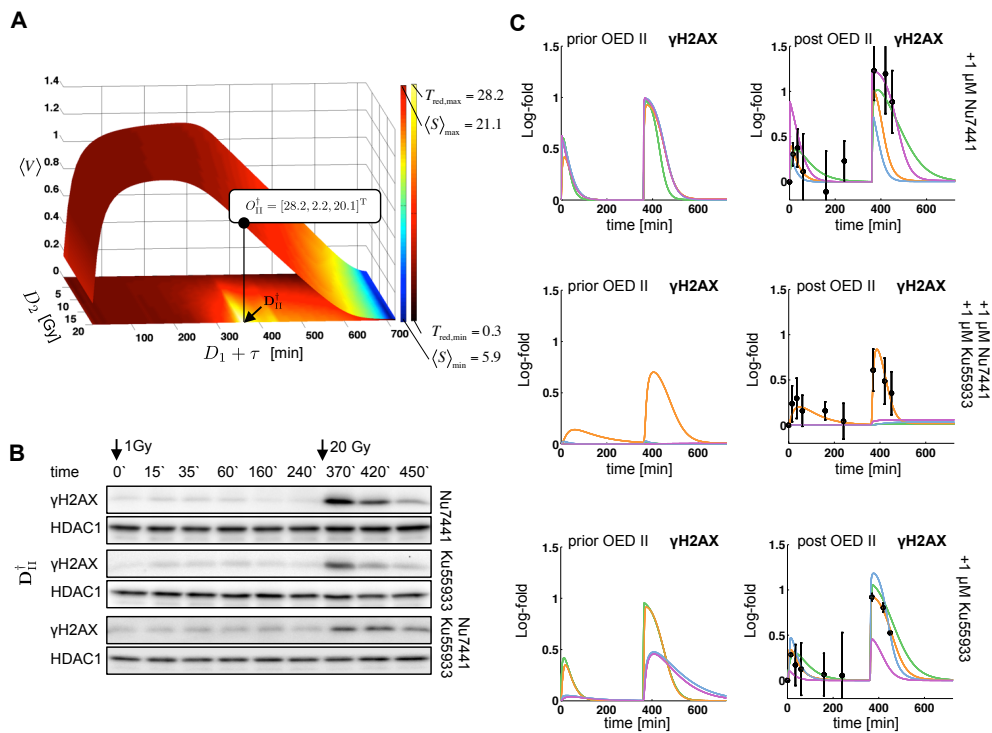
#### 3.5.3 Model identifiability analysis

Before DNA-PK<sub>cs</sub>, ATM and  $\gamma$ H2AX dynamics can be analyzed with model A2, an identifiability analysis was performed based on the profile likelihood to assess the uniqueness of the model prediction for unmeasured states. The profile likelihood samples can also be used to derive approximate prediction uncertainty bands via the envelope of all model trajectories associated to the parameter samples from the confidence interval. Since only relative data were at hand,  $\xi = \frac{[H2AX_{\text{tot}}]}{[Ku7080_{\text{tot}}]}$  and the readout scaling parameters are non-identifiable. This means, that the model cannot be used to predict absolute



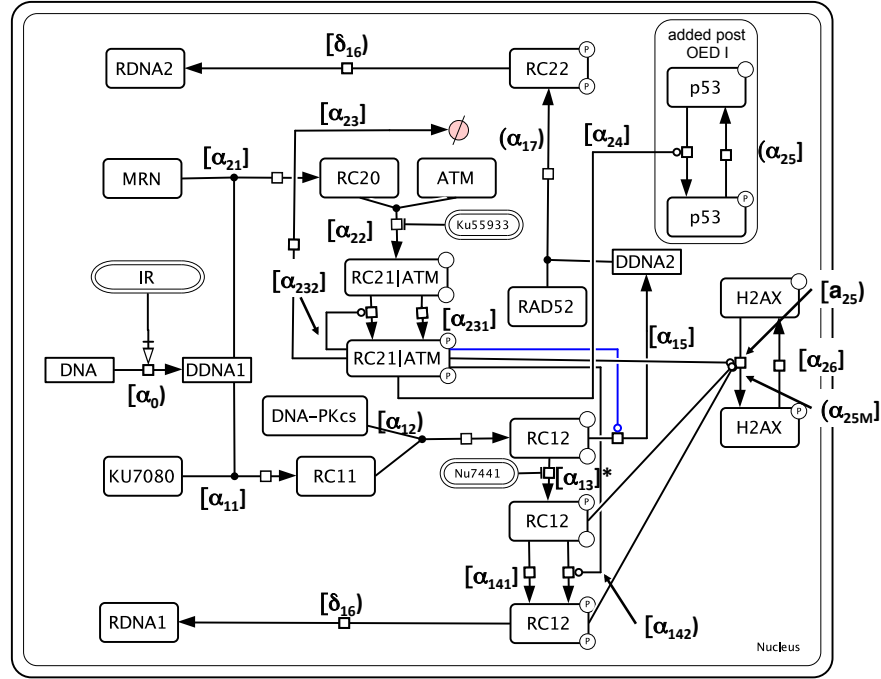
**Figure 3.5:** Parameterization of the stimulus design, design criteria and respective immunoblots. (A) Parameterization of the stimulus design for OED I/II. (B) Design criteria predicted from the model simulations are plotted over the feasible design space. The optimal design point for OED I  $D_1^\dagger$  and corresponding criteria are indicated. (C) A representative immunoblot from an experiment based on  $D_1^\dagger$  is shown. MDCK cells were irradiated as indicated and the insoluble nuclear extracts were analyzed by immunoblot. Lamin B2 served as loading control. (D) Corresponding model simulation implements the acquired data for  $\gamma$ H2AX (C), model colors as in Fig. 3.4. Data represent mean  $\pm$  2 STD of 3 independent experiments. (E) MDCK cells were irradiated as indicated. Inhibitors Ku55933 and Nu7441 were used at different concentrations and whole cell lysates were analyzed for p53-P and  $\gamma$ H2AX. GAPDH served as loading control. Model simulation and quantified experimental data for p53-P are shown. Data of a single experiment. Reproduced from ?.

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS



**Figure 3.6:** Design criteria and respective immunoblots. Parameterization of the stimulus design as in Fig. 3.5. (A) The optimal design  $D_{II}^{\dagger}$  is obtained as in Fig. 3.5B. (B) Sample data where MDCK cells were incubated with  $1 \mu\text{M}$  of the indicated inhibitor and irradiated as indicated. The insoluble nuclear extracts were analyzed by immunoblot. HDAC1 served as loading control. (C) The corresponding model simulations compare the acquired data for  $\gamma$ H2AX before and after OED II (mean  $\pm$  2 STD of 2-4 independent experiments). Colors as in Fig. 3.5. Reproduced from ?.

values of protein concentration. However, quantitative predictions regarding protein dynamics are possible. This is due to the fact, that the scaling parameters do no in-



**Figure 3.7:** The model structure for model A2 is shown including reaction parameters and the identifiability status: parameter  $p$  is  $[p]$  identifiable,  $[p]^*$  identifiable but exceeding the upper optimization bound,  $[p]$  non-identifiable at the upper limit,  $(p)$  non-identifiable at the lower limit,  $(p)$  structurally non-identifiable. The identified interaction that belongs to model A2 and could be discriminated with respect to the other models is indicated in blue. Reproduced from ?.

fluence the right hand side of the ODE system. Like the authors of ?, we thus treated scaling parameters as nuisance parameters. Regarding the kinetic parameters we found that 8 kinetic parameters are not fully identifiable for the given data and optimization constraints (upper and lower bounds such, that the parameters can vary 4 orders of magnitudes). Six of these parameters were non-significant at the upper bound, whereas the other two were non-significant at the lower bound. One parameter was structurally non-identifiable. The non-identifiable parameters were not decisive for the question of kinase contribution to H2AX phosphorylation. In Fig. 3.7, the reaction scheme and corresponding parameters including their identifiability property are shown. A discussion on each non-identifiable parameter and its meaning regarding model prediction interpretation is given in the following.

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS

---

Parameter  $\alpha_0$  has a non-identifiable upper bound for the given parameter estimation setup. The parameter represents the number of DNA double strand breaks per dose generated for a given dose rate. This means that the model structure has enough degrees of freedom to compensate higher but not too low DNA double strand breaks per dose rates. Thus, a minimal rate of DNA damage is needed to trigger the signaling. Compensation abilities by the model owing to limited information in the data is also apparent from the many parameter variations in terms of relative parameter change along the profile likelihood of  $\alpha_0$ , see appendix A.3.3. The parameter can be interpreted as a damage impact scaling parameter setting the scale of the downstream parameters. The qualitative behavior of protein dynamics is thus not changed.

Parameter  $\alpha_{12}$  represents the complex formation step between Ku7080 and DNA-PK<sub>cs</sub>. According to the profile likelihood bounds, a minimal rate of complex formation is needed, whereas the upper bound is unconstrained. This means that complex formation may be arbitrary fast, thus this reaction step may be neglected (model reduction). However, we kept this step in the model, as it represents a verified interaction (????). Here, the model is in the need of data that represent  $\alpha_{12}$ .

Although  $\alpha_{13}$  is practically non-identifiable for the given optimization setup, if it is increased above the upper optimization constraint, it then becomes identifiable. This means, that in principle the parameter is identifiable.

Parameter  $\alpha_{142}$  describes the catalysis of the second phosphorylation step of DNA-PK<sub>cs</sub> by ATM and has an unconstrained upper bound. This means, that catalysis of ATM seems to be necessary (lower bound is constrained), however, several parameters can compensate increased catalytic activity of this reaction (see relative change of the parameters along the profile likelihood in the detailed figures given in the appendix A.3.3). For instance  $\alpha_{141}$ , which represents the parallel reaction not catalyzed by ATM, anti-correlates with  $\alpha_{142}$ . Note that  $\alpha_{142}$  is identifiable owing to the data set where ATM is inhibited, which in turn makes the contribution of  $\alpha_{142}$  negligible small and thus uncovers  $\alpha_{141}$ .

Parameter  $\delta_{16}$  is used to model the final repair step for both, cNHEJ and HR/aNHEJ. This parameter has a lower bound, ensuring a minimal turnover of  $RC21^{pp}$ , which is related to the measurement signal. Since the upper bound of  $\delta_{16}$  is unconstrained, both repair steps can be arbitrarily fast in the model.

Parameter  $\alpha_{17}$  represents the reaction from Rad52 to RC22. As no measurement information is provided for this specific step, this reaction is thus unconstrained for the given data. Note that the subsequent  $\delta_{16}$  reaction has a lower bound, since it is also used in the DNA-PK<sub>cs</sub> part.

Parameters  $\alpha_{25M}$  and  $a_{25}$  are both related to the activation of  $\gamma$ H2AX. Parameter  $a_{25}$  has an unconstrained upper bound, whereas  $\alpha_{25M}$  is unconstrained on the lower bound.

Parameter  $\alpha_{25}$  represents the degradation of p53-P and can in principle be arbitrarily fast.

Having the identifiability characteristics of each parameter and associated states in mind, one can now move to model predictions.

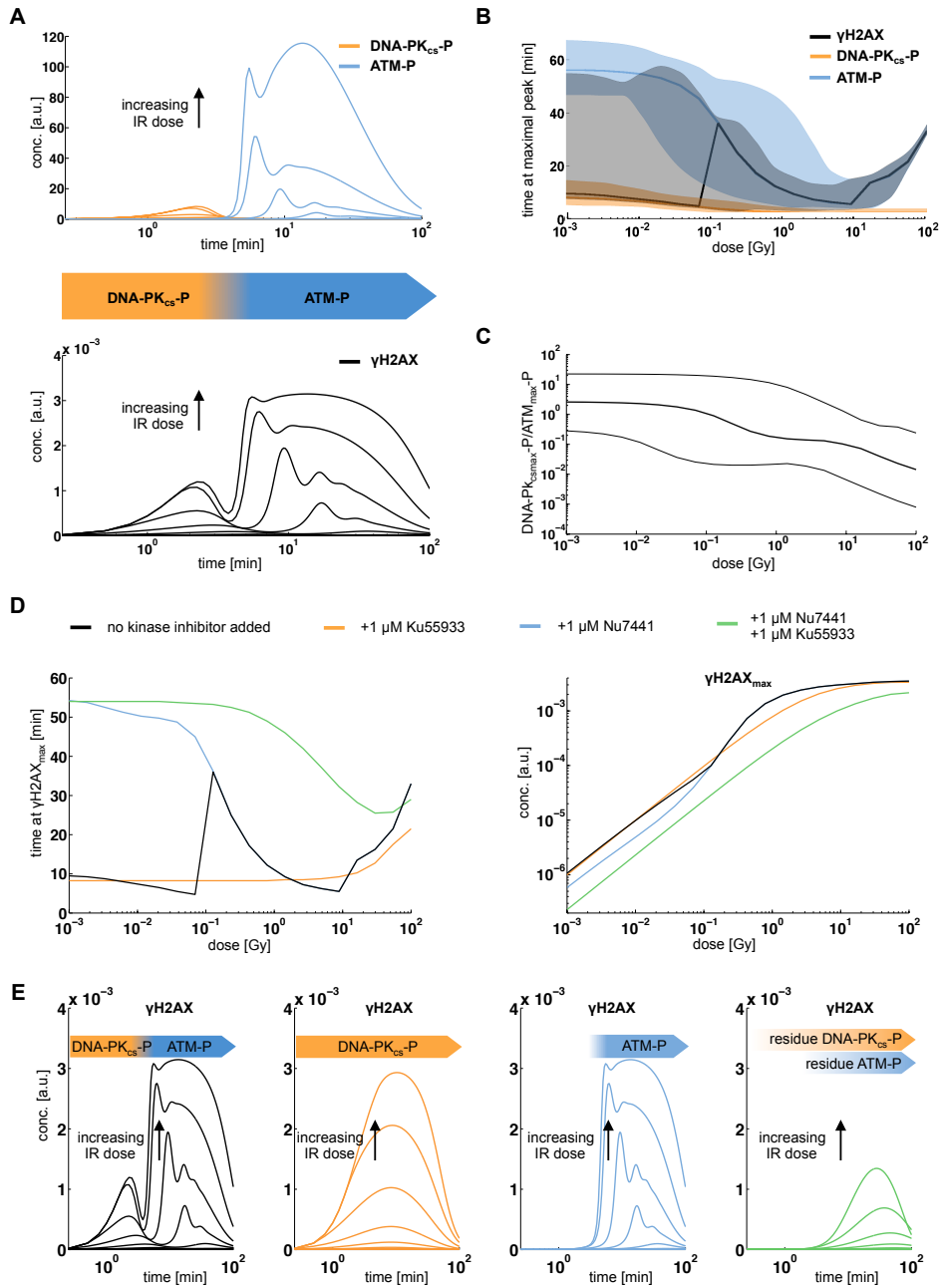
#### 3.5.4 Model predictions

To investigate contributions of DNA-PK<sub>cs</sub> and ATM to the H2AX phosphorylation, we analyzed times of maximal peak activity post irradiation. We simulated a single IR pulse from 1 mGy to 100 Gy (Fig. 3.8A-C). Active DNA-PK<sub>cs</sub> (DNA-PK<sub>cs</sub>-P) responds directly after irradiation within 2-10 minutes and shows fast signal attenuation. Response time of active ATM (ATM-P) in terms of maximal activity is delayed with respect to H2AX and dose-dependent ranging from 10 minutes to about 56 minutes. These model predictions are in line with the literature: DNA-PK<sub>cs</sub> activation peaks at 10 minutes after IR treatment, whereas ATM has its peak activity at around 20 minutes (?).

According to the model predictions, phosphorylation of H2AX is biphasic, following a dose independent temporal activation order: The first activation phase of  $\gamma$ H2AX right after stimulation is associated to DNA-PK<sub>cs</sub>, whereas the second phase is linked to ATM-P (Fig. 3.8A). The  $\gamma$ H2AX signal decays on the scale of hours and correlates with ATM-P. This kind of dynamics, fast initial and prolonged response is known from coherent feed forward loops, which serve as a signal persistence detector (?). At doses below 1 dGy, peak level of  $\gamma$ H2AX is dominated by DNA-PK<sub>cs</sub>, whereas above 1 dGy it is dominated by ATM (Fig. 3.8B, C). For larger dose levels, ATM auto-phosphorylation results in a prolonged activation phase, with  $\gamma$ H2AX peak activity shifted from 10 minutes at 10 Gy to 40 minutes at 100 Gy.

Simulations of  $\gamma$ H2AX dynamics with inhibited DNA-PK<sub>cs</sub> or/and ATM show that exclusive inhibition of ATM is nearly compensated by DNA-PK<sub>cs</sub> replacing the ATM associated activation phase of  $\gamma$ H2AX by a prolonged DNA-PK<sub>cs</sub> associated phase (Fig. 3.8D left; 3.8E black vs. orange). In contrast, DNA-PK<sub>cs</sub> inhibition leads to loss of the DNA-PK<sub>cs</sub> associated activation phase. Owing to slower activation kinetics, ATM cannot compensate this delay (Fig. 3.8D left; 3.8E black vs. blue). At doses where DNA-PK<sub>cs</sub> dominates,  $\gamma$ H2AX peak activity is delayed by roughly 45 minutes. Simulations of simultaneous inhibition of DNA-PK<sub>cs</sub> and ATM show a 3- to 10-fold reduction in  $\gamma$ H2AX peak level, depending on IR dosage, whereas exclusive inhibition of either DNA-PK<sub>cs</sub> or ATM is not as much affecting peak activity of  $\gamma$ H2AX (Fig. 3D right and 3E). For all inhibition scenarios, the biphasic phosphorylation kinetics of H2AX is lost.

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS



**Figure 3.8:** Model predictions for the dynamic contribution of DNA-PK<sub>cs</sub> and ATM to  $\gamma$ H2AX. (A) Simulated time courses of active DNA-PK<sub>cs</sub> and ATM and resulting biphasic  $\gamma$ H2AX activity for IR pulses of different dose levels (1 mGy to 100 Gy). At larger dose, ATM shows a damped oscillation as a result of a positive feedback (autophosphorylation). (B) Model prediction of the corresponding dose response in terms of time points at maximal activity of  $\gamma$ H2AX, DNA-PK<sub>cs</sub> and ATM. Shaded areas indicate 95% confidence regions of the model predictions estimated from simulation along the profile likelihood. (C) Ratio of maximal DNA-PK<sub>cs</sub>-P to ATM-P. Thin lines indicate 95% confidence region of the model predictions, estimated as in (B). (D,E) Simulations of *in silico* experiments with indicated inhibitors (color code) illustrating the co-regulation of DNA-PK<sub>cs</sub> and ATM, resulting in a partial redundancy. Reproduced from ?.



### 3.5.5 Discussion of *in vitro* application

This application illustrates an iterative workflow combining experimental work, computational modeling and experimental design methodologies to shed light on the interplay of two PIKK family members (DNA-PK<sub>cs</sub> and ATM) to the rapid histone H2AX phosphorylation in the context of DNA damage sensing upon  $\gamma$ -irradiation. By performing optimized dynamic stimulation experiments, an extensive set of time-resolved data was generated to identify a computational model for analyzing DNA-PK<sub>cs</sub>-P, ATM-P and  $\gamma$ H2AX dynamics. The parameter identifiability analysis revealed that the computational model could be used to predict internal state dynamics, e.g. phosphorylation of DNA-PK<sub>cs</sub> and ATM. With a predictive model at hand, it was then possible to investigate the fast phosphorylation kinetics of DNA-PK<sub>cs</sub>, ATM and H2AX post irradiation without the need of direct kinase activity measurements, thus reducing confounding effects from experimental manipulations.

The model simulations show that H2AX phosphorylation is biphasic, with initial and succeeding phase associated to DNA-PK<sub>cs</sub> and ATM, respectively, in which the individual contributions to peak level of  $\gamma$ H2AX are dose-dependent. It is tempting to link the dose-dependent biphasic response of  $\gamma$ H2AX observed *in silico* to the known biphasic signaling responses of cNHEJ and HR, that is fast DNA-PK<sub>cs</sub>- and slower ATM-related repair activity (?). In fact, following DNA-PK<sub>cs</sub> inhibition ? have shown that HR activity is increased. Further, ? showed that DNA-PK<sub>cs</sub> enzymatic activity inhibits HR in a titratable fashion. From simulating DNA-PK<sub>cs</sub> inhibition one may hypothesize that this is a consequence of delayed  $\gamma$ H2AX activation, associated chromatin remodeling and DNA repair initiation of cNHEJ. This fact may be exploited by cancer therapy development. One can further conclude that DNA-PK<sub>cs</sub> and ATM have distinct roles in H2AX phosphorylation equipping cells with a signal persistence detection function, i.e. fast initial response (DNA-PK<sub>cs</sub>) and delayed signal attenuation (ATM). This ensures reliable damage detection and repair signaling.

## 3.6 Summary optimal experimental design

Biological variability in combination with experimental measurement noise leads to distributed response signals, which is one of the main challenges when modeling biological systems deterministically with ODEs. To account for this variability, the parameter set needs to be extended to a parameter distribution. In this way, natural variabilities in the dynamic parameters as well as measurement noise can be readily accounted for. However, an exact quantification is computationally expensive and infeasible in an optimization framework for large systems. Therefore, approximate descriptions of the PDFs and the nonlinear mapping process between parameter and model response space are used. This chapter presented a nonlinear design approach based on sigma points within the application of model-based OED aimed at model discrimination. Its application and performance were illustrated using two numerical approaches from optimal

### 3. OPTIMAL EXPERIMENTAL DESIGN IN THE PRESENCE OF DISTRIBUTED MODEL PARAMETERS

---

control and several nonlinear model examples. Using the model overlap and modified T criterion as a robust design criterion, it was shown that in the case of nonlinear models with widely distributed parameter PDFs, the sigma point predictions and designs consistently outperform the linearization approach. In the case of bi-(multi)stability, the benefit of the nonlocal propagation property was illustrated. The sigma points come with several additional numerical advantages, including linear scaling of the numerical costs with respect to distributed parameters and derivative free estimation of nonlinearly mapped expectation and variance-covariance. The latter property allows applying a robust OED to dynamic models that have non-smooth right-hand side functions, e.g. cybernetic models of cellular metabolism ?.

Finally, a real life application was described, where a cyclic workflow between wet and dry labs has been established to analyze DNA damage sensing. In this application well-established experimental protocols regarding the readout  $\gamma$ H2AX were combined with robust dynamic stimulus experiments for generating a computational model. After careful identifiability analysis, the computational model could then be used as a surrogate of the experimental system to analyze the rapid dynamics and interplay of important sensor molecules post stimulation. Model predictions were in line with existing literature and gave rise to new verifiable experiments (biphasic response of  $\gamma$ H2AX) and allowed understanding the roles of the two PIKK family members (DNA-PK<sub>cs</sub> and ATM) in DNA damage signaling.

## 4

# Methods for identifying structural models of biochemical reaction systems

*The cause is hidden, but the result is well known.*

---

Publius Ovidius Naso  
Metamorphoses

The previous chapters demonstrated how challenging the identification of small signaling models can be, even though a moderate amount of time course data and highly sophisticated modeling methods are at hand. From these demonstrations it should be clear that a large-scale dynamic model identification of an entire biological system, say mammalian cell, organ or microorganism, including interaction quality and kinetic parameters is a challenging if not hopeless venture. In the last decade however, biochemical network reconstruction has become a very active field of research. Network reconstruction aims at identifying large-scale biological interactions structures only. Algorithms from network reconstruction allow analyzing the increasing amount of data generated by *omics* technologies (??). Whereas the structure of metabolic reaction networks could be reconstructed - mainly from genomic information - in great detail for many organisms (?) knowledge of the topology of regulatory and signal transduction networks is in many cases still incomplete and wiring diagrams even of canonical signaling pathways may differ in different cell lines (?). From a conceptual point of view, network reconstruction is equivalent to model and/or parameter identification as it is based on discriminating causal from correlation behavior between players in a biochemical network (gene, proteins, metabolites etc.) on the basis of perturbation experiments. Just as in model or parameter identification, network reconstruction can be regarded as a classification problem, where for given data one has to decide (=classify) whether a certain interaction is plausible or not.

## 4. METHODS FOR IDENTIFYING STRUCTURAL MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

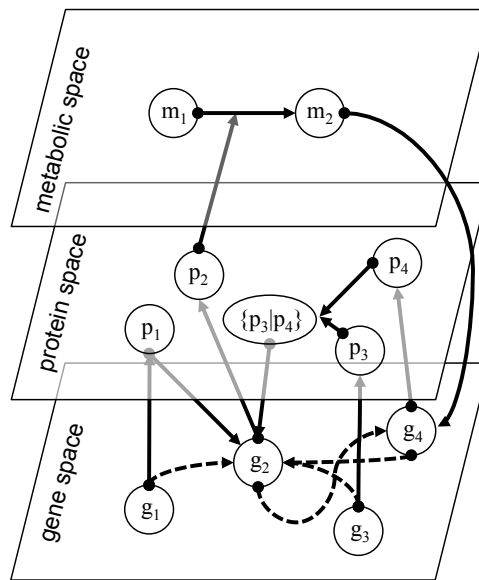
This second part of the thesis presents our recent work and contributions to algorithms that aim at reconstructing biochemical interaction networks based on high-throughput data from diverse types of microarrays (?), whereas the focus was put on the fundamental class of biological interaction networks, namely gene regulators networks. In principle the presented methods can also be applied to arbitrary interaction networks, e.g. metabolism or signaling networks. However, appropriate data need to be provided. In this chapter, a survey on (i) how to interpret reconstructed gene regulatory network, (ii) challenges for reconstruction and (iii) important methods for reconstruction is given. The following terms are used synonymously throughout the presentation: network model identification, network reconstruction, reverse engineering networks and network inference.

### 4.1 What are gene regulatory networks?

Gene regulatory networks provide the basis for systems-level understanding of interacting genes and phenotype formation in living systems. They condense different types of molecular interactions on the signaling, metabolic and genetic level to a network representation of causalities. Therefore, gene regulatory networks represent a causal projection of gene activities, neglecting detailed molecular mechanisms, Fig. 4.1 (?). This means that for given data (s. Sec. 4.2), reconstructed interactions may on the one hand represent direct gene-to-gene interactions. On the other hand, interactions can also represent influential interactions between two genes involving signal transduction, metabolism or epigenetic (?). Epigenetic refers to effects on gene expression levels, which result from mechanisms other than DNA sequence alteration or change in transcription factor activity. Examples of epigenetic regulations are histone modifications or DNA methylation. The fact that one cannot distinguish between projected or direct interaction is an inherent problem resulting from experimental limitations, as not all relevant input-output factors are observed. As a result, additional pseudo interactions may be derived during network identification owing to statistical dependencies, which have nothing to do with either projected or direct interactions and are referred to as indirect interactions. There exist several methods for removing such statistical dependencies, including TRANSWESD, which is presented in Ch. 5 for different kind of data sets.

#### 4.1.1 Definition of a gene regulatory network

A gene regulatory network can be represented as a graph  $G = (V, E)$  made up from a set of nodes or vertices  $V$  and a set of edges  $E$  interconnecting nodes. Nodes can represent states (phenotypes) of genes, gene regions, mRNA, proteins etc., whereas edges represent direct physical or influential interactions as discussed in the previous section. Targeted perturbation experiments allow identifying pairs of nodes via cause and effect reasoning. An edge  $e \in E$  is an ordered pair  $e = (i, j)$  indicating that node  $i$  affects or regulates



**Figure 4.1:** Gene regulatory network and what interactions in the gene space may represent. Dashed lines indicate the gene regulatory network, solid lines represent the actual mechanistic interaction. For instance: gene  $g_1$  up-regulates gene  $g_2$  via protein  $p_1$ , which may represent a transcription factor. The interaction between gene  $g_2$  and gene  $g_4$  is achieved via protein synthesis  $p_2$  and metabolite conversion  $m_1 \rightarrow m_2$ , which impacts on gene  $g_4$ .

## 4. METHODS FOR IDENTIFYING STRUCTURAL MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

by direct physical or influential causes node  $j$  and is thus called a directed edge, which is denoted as  $i \rightarrow j$ . The graph is then called a directed graph or digraph. Further, a signed digraph  $G = (V, E, \phi)$  has an additional sign function  $\phi : E \rightarrow \{-, +\}$ . The sign function maps the quality of the interconnection, i.e. promoting or inhibiting effect from the regulating node to the target node, which is indicated by the sign  $s$  via  $i \rightarrow_s j$ . Finally a weighted signed digraph has an additional weight mapping  $\Gamma : E \rightarrow \mathbb{R}^+ \setminus 0$ , assigning a non-zero, non-negative weight to each edge. Edge weights can be used to indicate the belief in a certain edge. Some algorithms deliver refined representations such as Boolean networks (??), reaction networks (?) or differential equations (?) but the main result is still the underlying network topology (s. Sec. 4.3.2).

### 4.1.2 Reconstructing a gene regulatory network

Reconstructing a gene regulatory network is the task of identifying interactions between known genes or gene regions, using experimental data, which represent the network of interest. Thus, as for parameter identifiability or model output distinguishability, a unique reconstruction of such an interaction structure is only possible when properly represented by the data. Besides identifiability restrictions owing to the specific experimental design and structural properties of a gene regulatory network, practical identifiability is also challenging. Typically one has the scenario  $n_{\text{genes}} \gg n_{\text{data}}$ , resulting in non-unique reconstruction solutions or non-identifiability issues. As in parameter estimation for ODE models, such ambiguities can partially be resolved by including prior knowledge (e.g. topological constraints or known node-node interactions), which reduces the space of possible solutions, ultimately yielding a unique, experimentally verifiable solution (??). Besides the curse of dimensionality, many of the reconstruction algorithms (s. Sec. 4.3.2) apply numerical optimization, resulting in computational complex and demanding problems. Therefore, in ? and ? methods based on simple correlation measures were developed especially tailored to (i) be applicable to genome scale reconstruction problems and (ii) perform well for the case of small sample sizes. Details on this method are given in Secs. 5.1, 5.2.

Even though the solution to the reconstruction problem may be non-unique, inferred gene regulatory networks can be used as a guide for further, detailed experimental analysis. GRN thus provide a tool for constrained hypothesizing reducing experimental efforts and costs especially for large scale reconstruction and refinement of gene regulatory networks. In this way, reconstructed GRNs can be used to narrow down genetic analysis by massively reducing the number of potential molecular interactions or locations of interaction sites. In the same way, GRNs can be used to identify putative intervention points by relating genetic spots to pathologic phenotypes (??). With regard to biotechnological production processes, GRNs can further be used to help optimize such, e.g. with respect to product spectrum and product yield.

## 4.2 Reconstruction data

As for the dynamic ODE model identification, network identification needs experimental data. Here non-targeted and targeted data classes can be distinguished. Each of these data types can be measured in a static or dynamic way. The term static is used instead of steady state, since a priori it is not clear, whether a specific snapshot of the system is in steady state or not. To the class of non-targeted experiments belong measurements that track changes in gene expression levels, without any information on the causes (non-targeted perturbation). From these data, only undirected networks can be reconstructed. *Undirected* refers to the lack of knowledge whether gene  $g_1$  activates  $g_2$  or vice versa. Targeted experiments keep track of changes in expression levels upon a known network perturbation (e.g. gene knock-out). Here, experimental designs comprise either one-perturbation at a time (or one-factorial, e.g. ?) or multi-factorial perturbations (??). One-perturbation at a time experiments may provide reliable identification for identifying GRNs (??). Large-scale reconstruction on one-perturbation at a time data is however not feasible given costs and technical difficulties. Further one-perturbation at a time experiments are typically based on gene knock-outs/ins, which can induce artificial biological effects (confounding effects) and are sometimes even unstable or lethal.

In contrast to one-factorial experiments, systems genetics builds on multi-factorial experiments using naturally occurring, multi-factorial genetic variations (e.g. single nucleotide polymorphisms (SNP)) as multi-factorial perturbations from which causalities can be unraveled more efficiently (??). Systems genetics methods use properly controlled genetic crosses (segregating populations) such as recombinant congenic strains (RCSs), recombinant inbred lines (RILs), advanced intercross lines (AILs) or chromosome substitution strains (CSSs) to causally link genetic or chromosomal regions to observed phenotypic trait data (??). In this way, systems genetics can reveal complex genetic interactions in biological systems by relating genetic variations to various phenotypic data from high-throughput measurements (??). The choice of strain design depends on the reconstruction goal and should support a rigorous network reconstruction.

## 4.3 Methodological approaches for network reconstruction

Gene regulatory network reconstruction methodologies can be grouped into (i) physical or (ii) influential interaction based identification approaches. Physical approaches are based on DNA motif analysis seeking for pairs of DNA motifs and transcription factors, which may potentially interact on pure molecular basis. It is thus restricted to direct, physical interaction. This restriction simplifies network identification, but does not allow reconstructing gene interactions based on mechanisms other than transcription factors. Here the influential approach has an advantage of reconstructing a more comprehensive interaction network, capturing gene-to-gene interactions independent on their interaction mechanism. Current research on reconstruction algorithms has focused on influential approaches owing to the fact that less prior knowledge is needed and less

## 4. METHODS FOR IDENTIFYING STRUCTURAL MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

specific data have to be generated. Depending on the data, influential approaches may also infer physical interactions. For instance, interactions reconstructed from data reflecting activity (e.g. expression levels) of transcription factors and genes most likely represent direct molecular interactions. Therefore, when it comes to interpreting the reconstructed networks one has to be aware of whether a specific reconstructed edge (=interaction) is based on physical or influential information. In the following, data preprocessing and influential reconstruction approaches are discussed.

### 4.3.1 Data preprocessing

Before any reconstruction method can yield meaningful results, data need to be preprocessed in order to account for technical errors and biological variance. This includes normalization (remove systematic variations to for instance compare different microarrays) and transformation (?). A transformation may be necessary for approaches that use binarized data or assume a specific distribution of measurement variations. Depending on the reconstruction method, feature reduction by means of filtering (e.g. noise) and clustering (e.g. gene regions) may be needed in order to obtain a numerically tractable problem (??). In Secs. 5.1 and 5.2 noise filtering (experimental and biological variance) as well as clustering is applied using z-score and Pearson correlation. The z-score only works well for normally distributed gene activity data, which is typically not the case for raw, unprocessed data. Here, a logarithmic transformation has proven to often yield approximate normal distributions.

### 4.3.2 Reconstruction methods

Identification of gene regulatory networks is a highly challenging task asking for sound knowledge in biological and technological aspects but also in data mining, statistical analysis, network modeling and graph theory. In the following, the most frequently used reconstruction methods are briefly discussed. The discussion is however not complete, since many methodological variations exist. The reason lies in the fact, that data sets are often very special depending on the biological question, and thus existing methods have been tailored to fit the problem demands. Therefore, in contrast to the detailed survey on recent dynamic model identification methods in chapter 2 this part gives a rough classification of methodological approaches for network reconstruction.

**System of linearized equations** Equation systems can be used to interrelate changes in gene activity levels, e.g. expression levels. Owing limited amount of data, linear equation systems of the form

$$x_i = \mathbf{a} \mathbf{x} \tag{4.1}$$

are being used, relating the activity level  $x_i$  of a gene  $i$  to the activity levels  $x_j$  of genes  $j = 1, \dots, n_{\text{genes}}$  (note that one may exclude self-regulation via  $j \neq i$ ). By doing so, linear



### 4.3 Methodological approaches for network reconstruction

---

and additive interaction mechanisms are assumed. The vector  $\mathbf{a}$  needs to be estimated from data for each gene and can be used to form an  $n_{\text{genes}} \times n_{\text{genes}}$  matrix representing the interaction or association weights (?). Structural equation modeling as an extension to multiple regression has been applied to reconstruct gene regulatory networks (??). Regression algorithms have further been improved with regard to covariate selection to address the problem of small sample size. Here, least angle regression (?) and regression shrinkage via the lasso (?) have been proven to be very useful. An extension of Eq. (4.1) to a dynamic description is achieved by linear ordinary differential equations (???). As pointed out in the introduction of this chapter, such models are however only applicable to small-scale networks, owing to the tremendous amount of experimental effort to generate dynamic expression data.

**Boolean network** Boolean networks are network interaction graphs that also describe the quality of interaction. Using Boolean logic, Boolean networks allow to incorporate regulatory mechanisms on a very crude level (??). One of the earliest methods for reconstructing biological Boolean networks has been proposed by ?. In ? temporal Boolean networks are developed describing temporal effects on a discrete time steps. A further extension to describe dynamic state transitions on a continuous basis was proposed by ?. Here, the logical gates are represented by continuous functions to derive a set of ordinary differential equations.

**Bayesian network** A Bayesian network is represented by a directed, acyclic graph and a set of (conditional) probability distributions. The probability distributions describe state probabilities of the nodes. Thus, in a Bayesian network, nodes represent stochastic variables, whereas edges represent conditional interdependencies. If two nodes are not interacting, their conditional probability is simply the product of the probability of each node. Bayesian network reconstruction methods allow to incorporate prior knowledge in an easy way, by specifying the corresponding prior distributions (??). They are also very flexible when it comes to incomplete data sets (?). A limitation of Bayesian networks is their restriction to acyclic graphs, e.g. ?.

**Association network** Association networks are derived from association measures as mutual information or correlation coefficients (??). The main advantage of such an approach is that even large-scale data can be used to derive an association measure with very little computational effort. Typically, no optimization algorithm for estimating network parameters is applied, approaches are solely based on analyzing the provided data alone. A disadvantage is, that often only one-by-one interactions are considered, which may potentially lead to problems when a gene is regulated by a combination of several genes (?). Conditional association measures may be used, e.g. conditional correlation (?), to overcome this problem. Then however, computational demands increase. Association networks are ideally suited for data containing differential information, e.

## 4. METHODS FOR IDENTIFYING STRUCTURAL MODELS OF BIOCHEMICAL REACTION SYSTEMS

---

g. knock-out vs. control conditions. Being easily scaleable, association based reconstruction methods can be used on genome scale, either to directly reconstruct a GRN or to filter potential interactions, which may be refined by methods that also account for higher order interactions. Also, methods from machine learning have been tailored to reconstruct gene regulatory networks. The random forest approach is used for classification and regression problems (??). As correlation based approaches, random forests do not assume a specific kind of interaction. One advantage of the random forest approach is its scaleability, owing to the fact that each tree in the random forest can be trained independently (?).

### 4.4 Summary

In this chapter a survey on identifying gene regulatory networks was given. Starting from a general discussion on gene regulatory network interpretation, required data, challenges and methodological approaches were outlined. As has been illustrated, there exist many different approaches for network inference, whereas the choice of method mostly depends on the specific scientific question and available data. The next chapter presents two methodological approaches based on association networks that have been developed for reconstructing large-scale gene regulatory networks.

## 5

# TRANSWESD: A reverse engineering algorithm for identifying gene regulatory networks

*Everything should be as simple as it is, but not simpler.*

---

Freely adapted from Albert Einstein

This chapter presents the original TRANSWESD (TRANSitive Reduction for WEighted Signed Digraphs) reconstruction algorithm, a generalized variant of transitive reduction designed for one-perturbation at a time data, which has been published in ?. In ?, we refined the original TRANSWESD in an international collaboration for one-perturbation at a time data, whereas in ? we developed a reconstruction framework based on TRANSWESD for multifactorial perturbation data. With TRANSWESD we combined and extended existing variants (??) to handle weighted perturbation graphs with negative cycles. Despite its conceptual simplicity, TRANSWESD is highly competitive with other reverse engineering methods and especially useful for the scenario  $n_{\text{genes}} \gg n_{\text{data}}$ . The presentation in this chapter is based on ?, ?, ? and ?.

## 5.1 TRANSWESD

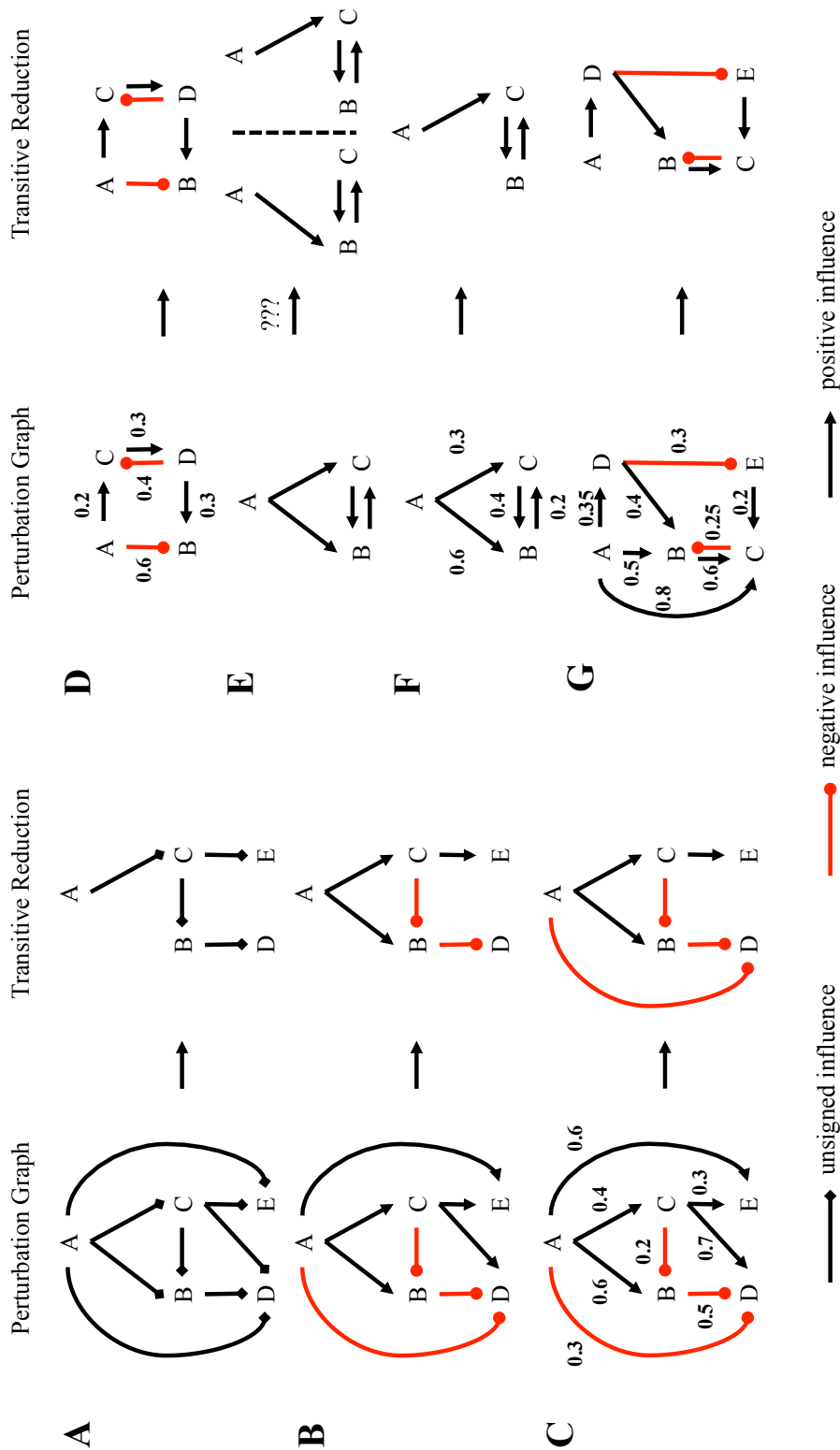
This section is based on results that we have generated and published in ?.

A simple yet smart method for reconstructing a regulatory network with  $n$  nodes comprises the following steps: the node states of a biochemical reaction network are measured in a control scenario, e.g. unperturbed wild-type. Then at least  $n$  different perturbation experiments are conducted: in experiment  $i$  node  $i$  is perturbed, whereas

## 5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS

---

all other  $n - 1$  nodes are screened for potential state changes compared to their control states. If a perturbation in node  $i$  affected  $j$  according to some statistical measure (s. below), a directed edge from node  $i$  to  $j$ ,  $i \rightarrow j$  is initially assumed. The complete set of observed effects in all perturbation experiments leads to an initial *perturbation graph*. As pointed out in Ch. 4, each edge in the perturbation graph reflects either a direct (including direct projection) or an indirect (statistical dependency) effect of one node upon another. The next step deals with a central issue in network reconstruction, namely identification and removal of edges that represent indirect effects. Here, transitive reduction as used by ? can be used to find the minimal (most parsimonious) subgraph that can explain all effects seen in the experiments. Transitive reduction in its most general form allows removal and addition of edges to find the minimum graph (?). However, in the context of network reconstruction, one usually focuses on the special case where edges may only be removed, i.e. where one searches for a minimal subgraph explaining the perturbation graph. This is also known as minimum equivalent graph problem (??). Therefore, TRANSWESD uses transitive reduction restricted to edge removal. ? determined the minimal subgraph from the perturbation graph by removing all edges  $i \rightarrow j$  for which a (simple) path starting in  $i$  and ending in  $j$  (not using  $i \rightarrow j$ ) can be found, assuming the effect of  $i$  on  $j$  to be indirect, thus explainable by the path. The resulting graph is the transitive reduction of the perturbation graph. A simple example is depicted in Fig. 5.1A. The method proposed by ? has some drawbacks. First, transitive reduction as described above does not consider the full amount of information that is available from perturbation experiments, even when considering only qualitative observations. If a node shows a significant response to a perturbation, one can at least classify the measured effect as *up* or *down*. This information can be taken into account by adding a sign label to each edge in the perturbation graph, which becomes then a signed directed graph (see Fig. 5.1B, a signed version of Fig. 5.1A). Transitive reduction can then be performed in a similar way: an edge  $i \rightarrow j$  is deleted only if there is a path from  $i$  to  $j$  whose overall sign (product of the signs of the involved edges) corresponds to the sign of this edge. As can be seen in the example in Fig. 5.1B, this may save edges that were mistakenly deleted in the unsigned version. A second drawback of the original approach of transitive reduction is the risk to remove true edges, even in signed perturbation graph. The radical pruning strategy of transitive reduction aims at minimizing false positive (FP, type I error) edges in the reconstructed network but it may lead to a high number of false negatives (FNs, type II error). This effect becomes visible in networks comprising many (coherent) feed-forward loops where a node may affect another node via direct (edge) and indirect (path) links of the same sign. Since feed-forward loops have been shown to occur frequently in gene regulatory networks (?), this property can become a serious limitation of the method. A third shortcoming is the prerequisite that the perturbation graph is acyclic - a condition that is often not fulfilled in realistic biological networks. If the perturbation graph is cyclic, the solution of transitive reduction is, in general, not unique. Further, negative cycles in signed perturbation graphs may bring about even more complications for transitive reduction.



**Figure 5.1:** Example for transitive reduction and its extension to TRANSWESD. Details are discussed in the text. Adopted from ?.

## 5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS

---

These drawbacks motivated the development of TRANSWESD in ?. Major changes and improvements concern: (i) new statistical approaches for generating high-quality perturbation graphs from systematic perturbation experiments, (ii) the use of edge weights (association strengths) for recognizing true redundant structures, (iii) causal interpretation of cycles, (iv) relaxed definition of transitive reduction, and (v) approximation algorithms for large networks. The following sections introduce the method in detail using the example of one-perturbation at a time data. In Sec. 5.2, modification of TRANSWESD to work on multi-factorial perturbation data are described. Standardized benchmark tests are used to demonstrate the potential of TRANSWESD. As the results highlight, despite its conceptual simplicity, TRANSWESD outperforms existing variants of transitive reduction and is highly competitive with other reverse engineering methods (s. Sec. 5.1.6, 5.2.2.1,5.2.2.2).

### 5.1.1 General workflow of TRANSWESD for one-perturbation at a time data

- Step 1 (Sec. 5.1.2): as explained in the introduction, starting point is a wild-type experiment plus  $n$  perturbation experiments in each of which one of the  $n$  nodes is perturbed and the resulting response of the other nodes is measured, either in transient phase or in steady state. The control or wild-type states are denoted with  $\mathbf{x}_0$  ( $x_{0,i}$  denotes the control state of the  $i$ -th node). It is further assumed, that appropriated normalization and transformation have been performed to have a normally distributed measurement signals. The vector of normalized state activity (e.g. gene expression) measured in the  $i$ -th perturbation experiment (where node  $i$  is perturbed) is denoted by  $x_{ij}$ , i.e.  $x_{ij}$  is the state of the  $j$ -th node in perturbation experiment  $i$ .
- Step 2 (Sec. 5.1.3): for each node, the unperturbed state ( $x_{0,j}$ ) is compared to the measured states in the perturbation experiments ( $x_{ij}$ ). Using an appropriate classification strategy, significant changes are identified and included as signed edges  $i \rightarrow j$  in the resulting signed, cyclic perturbation digraph G1.
- Step 3 (Sec. 5.1.4): each identified edge in the perturbation graph is endowed with a weight extracted from the measurements indicating the association/interaction strength between the two connected nodes, extending G1 to the signed, cyclic and weighted perturbation digraph G2.
- Step 4 (Sec. 5.1.5): the final step is the computation of the transitive reduction graph G3 using TRANSWESD, which can handle weighted, cyclic and signed digraphs. Note that, in principle, TRANSWESD may accept any perturbation graph, even if the way to generate the graph is different from Steps 1-3. This is illustrated in Sec. 5.2.

## 5.1.2 Perturbation graph

### 5.1.2.1 Single perturbation data

To decide whether a perturbation of  $i$  induces a significant effect on node  $j$  (and is thus integrated as edge  $i \rightarrow j$  in the perturbation graph) one can either use correlation analysis of the entire data or only direct variation measures quantifying the change of  $x_j$  when perturbing  $x_i$ . The correlation measure of the entire data is beneficial for determining the strength of association between nodes (see Sec. 5.1.4). Following the simple idea of relating state changes with respect to the control, one might completely ignore the presence of noise and define the variation measure for node pair  $(i, j)$  as

$$\Delta_{ij} \equiv (x_{ij} - x_{0,j})\pi^i, \quad (5.1)$$

with  $\pi^i$  as an indicator of perturbation direction (up  $\pi^i = +1$  by e.g. knockout or knockdown, down  $\pi^i = -1$  by e.g. over-expression). Initially, edges may be introduced according to

$$i \rightarrow_- j \quad \text{if} \quad \Delta_{ij} < 0 \quad (5.2)$$

$$i \rightarrow_+ j \quad \text{if} \quad \Delta_{ij} > 0, \quad (5.3)$$

leading to an initial, signed perturbation graph. Apparently, this graph will capture direct (edge) and indirect (path) effects.

Owing to measurement and intrinsic noise, many spurious direct and indirect interactions will be identified as well. A naive use of Eq. (5.1) will therefore result in a very dense - most likely a fully connected - perturbation graph, containing many false positive interaction predictions. A simple extension to Eq. (5.1) can be obtained as follows: If the data set is large enough, meaning that either several hundreds of nodes have been measured and/or several replicates are available, it may be advisable to (i) replace  $x_{0,j}$  by a robust control level  $E[x_j]$  derived from averaging over all available data of node  $j$  and (ii) by weighting  $\Delta_{ij}$  with the state sample standard deviation  $STD_j$  of node  $j$ :

$$z_{ij} = \frac{x_{ij} - E[x_j]}{STD_j}. \quad (5.4)$$

By doing so, deviations from the control level are robustified against fluctuations of the control level itself. Further, deviations are weighted according to the overall variance of the target node, which may heavily depend on the position and connectivity of the target node. Such a robustification is justified if the target node  $j$  is only directly affected by a small number of perturbations. In addition, the above z-score in Eq. (5.4) may be refined iteratively by excluding node state measurements from sample mean and standard deviation calculation that correspond to identified edges. This approach has been proven to be very powerful when measuring perturbations on hundreds of nodes. It has been successfully applied in ? and ?.

## 5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS

---

### 5.1.3 Identifying significant edges

The z-score from Eq. (5.4) can be used to identify significant deviations from the control  $E[x_j]$ , which result from direct or indirect interactions and are thus interpreted as potential regulations. If the distribution of measurement fluctuations (including noise, inherent and also higher order influence from *further away perturbations*) is known, significant deviation from the mean (=outliers) can be derived for a given p-value, which functions as a threshold parameter for significant edges. The p-value choice can be optimized, since a specific significance level of say  $\alpha = 0.05$  must not correspond to a biological significant effect. Here, there are several approaches for deriving an optimal value. Either training against known networks or a selection based on the minimal/maximal number of expected edges to be found. Further, an optimal threshold value can be derive based on the rate of edge inclusion when increasing/decreasing the threshold. At a threshold value where the rate of edge inclusion increases significantly, one can assume that this is the critical level, where interactions cannot be distinguished from noise. Note that thresholds may be defined based on p-values but also directly in terms of standard deviations, provided that the sample based z-score follows a standardized normal distribution (which is the case for properly processed data following a Normal distribution). A suitable threshold strategy for obtaining a high quality perturbation graph from noisy data is an important step, since there is a critical edge density for each graph up to which transitive reduction related algorithms work well in terms of pruning result and computational time. Whereas edges reflecting indirect effects may be filtered by TRANSWESD at a later stage (see below), edges indicating neither direct nor indirect (thus noise) effects cannot be corrected and will lead to reconstruction errors. On the other hand, the number of FNs is also to be minimized as they cannot be recovered by transitive reduction. An example illustrating the z-score based thresholding strategy is given in the supplementary material of ?. After defining a threshold value, significant directed signed edges are selected, whereas the sign is derived from Eqs. (5.2,5.3) applied to Eq. (5.4). The resulting initial perturbation graph  $G_0$  represents a signed digraph, which may contain cycles.

Initially, in ? we have introduced two thresholds based on (i) the overall sample standard deviation and (ii) the individual standard deviation of each node. An overall standard deviation was thought to reflect the magnitude of the variation measure Eq. (5.1) in contrast to the overall variation when searching for edges. The node specific standard deviation was introduced to account for individual dynamic nature of each node. Recently, we demonstrated in ? that the individual standard deviation in combination with the correlation coefficient as an overall variation measure and a sign consistency check (positive or negative z-score deviation vs. sign of the corresponding Pearson correlation coefficient) performs much better than our original approach in ?. In the original approach, we also ignored sign inconsistencies between correlation coefficient and z-score.



### 5.1.4 Weight association to the edges

Signed, weighted perturbation digraphs enable TRANSWESD to also work on cyclic network structures. In the original TRANSWESD (?), weights are derived from a conditional Pearson correlation  $r_{ij}$ , where perturbation data on node  $j$  are excluded to not bias the correlation. Accordingly,  $r_{ij}$  is not symmetric with respect to index permutation. For each potential edge in the perturbation graph  $G_0$  a weight

$$w_{ij} = 1 - |r_{ij}| \quad (5.5)$$

is assigned, reflecting behavioral distance or association uncertainty leading to the signed, weighted, perturbation digraph  $G = (V, E, \phi, \Gamma)$ . The higher the weight, the weaker is the association of nodes  $i$  and  $j$ , thus the more unlikely is the identified direct edge between them. This weight representation may seem contrary to other work, where edge weights typically represent confidence or likelihood. However, this weight representation can directly be used by TRANSWESD when calculating shortest paths for identifying paths with lowest overall weight (=highest overall association).

### 5.1.5 Removing indirect edges: TRANSWESD

The reconstruction may stop at  $G$ , however, as described in the introduction of this chapter, it is very likely that edges display direct (edges) or indirect relationships (paths). Edges that represent indirect effects result in false positive predictions, which transitive reduction seeks to remove to obtain true negatives (TNs). However, this removal is at the risk to also remove true positive (TP) predictions resulting in false negatives. In the following, the description of the developed algorithm TRANSWESD is started with simple signed acyclic perturbation graphs. Then the algorithm is extended to signed, weighted acyclic graphs to finally obtain the full TRANSWESD algorithm for signed, weighted cyclic perturbation graphs. In this way, the idea of transitive reduction is generalized step by step and extensions of TRANSWESD to minimize shortcomings of previous variants can easily be followed.

#### 5.1.5.1 Transitive reduction in signed acyclic graphs

? used transitive reduction to prune unsigned acyclic perturbation graphs. It is straightforward to generalize this procedure to signed acyclic perturbation graphs  $G$  (at this point weights are neglected). The basic idea is to check for each edge  $i \rightarrow_s j$  in  $G$  whether there is an elementary path  $i \Rightarrow_s j$  (a sequence of edges between nodes, where no node occurs twice) not involving edge  $i \rightarrow_s j$ , which can then be seen as an explanation for the observed influence  $i \rightarrow_s j$  allowing one to remove this particular edge. For this purpose, in a first step for each pair of nodes  $(i, j)$  the shortest positive (positive overall sign) and shortest negative (negative overall sign) path is calculated. Shortest paths can be used to test whether a positive and/or negative path from  $i$  to  $j$  exists at all. As one is only interested in the existence of paths one may use arbitrary edge

## 5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS

---

weights, e.g. setting all to one, and arbitrary metric. The double label algorithm is employed, which is a generalized version of the Dijkstra algorithm, for computing shortest positive/negative paths. The double label algorithm delivers exact results in polynomial time if the signed graph is acyclic (??). The lengths of the shortest positive and negative paths are stored in a matrix  $\mathbf{S}^+$  and  $\mathbf{S}^-$ , respectively. For example,  $\mathbf{S}^+(i, j)$  stores the length of the shortest positive path from  $i$  to  $j$ . An infinite length ( $\infty$ ) is stored if no path exists. After this preparatory step,  $G$  is pruned to the minimal graph  $G_{\text{TR}}$ . Here minimal refers to the number of edges. The minimal graph satisfies

$$\mathbf{S}_{\text{TR}}^+(i, j) < \infty \quad \forall \text{ removed positive edges } i \rightarrow_+ j \text{ in } G \quad (5.6)$$

and

$$\mathbf{S}_{\text{TR}}^-(i, j) < \infty \quad \forall \text{ removed negative edges } i \rightarrow_- j \text{ in } G. \quad (5.7)$$

In acyclic signed graphs, the unique solution can easily be found with the help of  $\mathbf{S}^{+/-}$ . For each edge  $i \rightarrow_s j$  one simply checks the existence of a successor  $k \neq j$  with  $i \rightarrow_q k$  and  $k \Rightarrow_t j$  which fulfills the sign condition  $qt = s$ . Such a path  $k \Rightarrow_t j$  exists if  $\mathbf{S}^t(k, j) < \infty$ . If such successor  $k$  exists, one can conclude that  $i \rightarrow_s j$  is explainable by the augmented path  $i \rightarrow_q k \Rightarrow_t j$ . Then, one can remove the direct edge  $i \rightarrow_s j$ . This line of reasoning only holds for acyclic graphs. Note that it is not necessary to re-compute the shortest paths lengths  $\mathbf{S}^{+/-}$  after removal of the edge  $i \rightarrow_s j$ . In all paths where this particular edge is contained one can replace the latter with  $i \rightarrow_q k \Rightarrow_t j$ . Here again, acyclicity of the graphs ensures that the resulting path is still elementary and thus a valid explanation. Eliminating all such removable edges, one obtains the unique minimal equivalent graph  $G_{\text{TR}}$  which yields the same perturbation effects as the original graph  $G$ . Transitive reduction in unsigned graph uses the same algorithm but neglects the sign condition. This definition of transitive reduction differs in some aspects from the version used in ?. First, only elementary paths (not involving cycles) are considered as possible explanations for edges. Second, instead of condition in Eqs. (5.6,5.7), ? follow the original definition of transitive reduction, which is

$$\mathbf{S}_{\text{TR}}^+(i, j) < \infty \quad \text{wherever } \mathbf{S}^+(i, j) < \infty \quad (5.8)$$

and

$$\mathbf{S}_{\text{TR}}^-(i, j) < \infty \quad \text{wherever } \mathbf{S}^-(i, j) < \infty. \quad (5.9)$$

This condition can be relaxed to condition in Eqs. (5.6,5.7), since there is no necessity to preserve a path  $i \Rightarrow_s j$  between two nodes  $i$  and  $j$  if no edge  $i \rightarrow_s j$  (i.e. neither a direct nor an indirect effect of  $i$  on  $j$ ) could be deduced from the experiments. However, as long as acyclic graphs are considered, both conditions will lead to the same result because then, condition in Eqs. (5.8,5.9) follows from condition Eqs. (5.6,5.7). The

example in Fig 5.1B illustrates that accounting for edge signs avoids edge removals that cannot be explained: in contrast to Fig. 5.1A (unsigned perturbation graph) the edge  $A \rightarrow_- B$  is kept because the path  $A \rightarrow_+ C \rightarrow_+ B$  cannot explain the negative sign of this edge.

### 5.1.5.2 Transitive reduction in signed and weighted acyclic graphs

Transitive reduction cannot detect redundant structures such as coherent feed-forward loops implying a possibly large number of FNs. A relaxed pruning strategy could be achieved by considering also edge weights quantifying the overall strength of the associations. In this step of developing TRANSWESD, it is now allowed to remove an edge (then considered as an indirect influence) only if its sign and also its weight can be explained by another path. The condition in Eqs. (5.6,5.7) is thus generalized to demanding that the pruned graph  $G_{\text{TR}}$  should be minimal and satisfy

$$\mathbf{S}_{\text{TR}}^+(i, j) < \psi w_{ij} \quad \forall \text{ removed positive edges } i \rightarrow_{+,w_{ij}} j \text{ in } G \quad (5.10)$$

and

$$\mathbf{S}_{\text{TR}}^-(i, j) < \psi w_{ij} \quad \forall \text{ removed negative edges } i \rightarrow_{-,w_{ij}} j \text{ in } G, \quad (5.11)$$

with positive confidence factor  $\psi$ . In order to fulfill the condition Eqs. (5.10,5.11), the transitive reduction step has to be modified as follows: remove an edge  $i \rightarrow_{+,w} j$  if a successor  $k \neq j$  of  $i$  can be found such that  $i \rightarrow_{q,c} k$  and  $k \Rightarrow_{t,d} j$  exist which fulfill (i) sign condition  $qt = s$  and (ii) weight condition  $\max(c, d) < \psi w$ . Notice that for quantifying the overall weight (=length) of a path, the MAX-metric is used (maximal weight along the path). This reflects the property that an influence path is as good as its weakest edge having the largest weight and thus lowest degree of association. For path calculation, the double label algorithm adapted for the MAX-metric is used. The factor  $\psi$  controls the overall association strength a path must have in order to explain a given edge. It represents another tuning factor of TRANSWESD (besides the threshold for edge detection, Sec. 5.1.3). For results produced in ?, ? and ? we used  $\psi = 0.95$ , i.e. a value close to one. Smaller values of  $\psi$  require larger associations in all edges of a path to explain an edge. For  $\psi = 0$  one has  $G_{\text{TR}} = G$ . If  $\psi > 1$  one would accept paths to explain an edge, despite the fact of the overall weight being smaller than the weight of the edge, which is to be explained by the path. For  $\psi = \infty$  the condition Eqs. (5.10,5.11) coincide with Eqs. (5.8,5.9), thus the original transitive reduction in unweighted graphs. Importantly to note, an acyclic graph ensures that the augmented path  $i \rightarrow_{q,c} k \Rightarrow_{t,d} j$  resulting in  $i \Rightarrow_{s,\max(c,d)} j$  is elementary, i.e.  $k \Rightarrow_{t,d} j$  does not contain edge  $i \rightarrow_{q,c} k$  and is thus a valid explanation for  $i \rightarrow_{s,w} j$ . Additionally, acyclicity reduces computational demands, since  $\mathbf{S}^{+/-}$  do not have to be re-computed after edge removal. In Fig. 5.1C the impact of the additional weight consideration is shown. Here an edge is kept if alternative paths cannot explain its high

## 5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS

---

association strength. In contrast to Fig. 5.1B,  $A \rightarrow_{-,0.3} D$  is retained because the path  $A \rightarrow_{-,0.6} B \rightarrow_{+,0.5} D$  has weight 0.6 and is thus not a valid explanation for  $\psi < 1$ . ? use an analogous scheme, however limited to triangle motifs, i.e. an edge  $i \rightarrow j$  is removed only if two consecutive edges  $i \rightarrow k \rightarrow j$  can explain the effect from  $i$  on  $j$ .

### 5.1.5.3 Transitive reduction in signed and weighted cyclic graph

Finally, signed weighted digraphs with cycles are considered, which are of course present in most cellular networks. Although cyclic structures are of vital importance to cellular systems, e.g. to damp a specific response via a feedback leading to complex dynamics within the network, they hamper network reconstruction. As in the acyclic case, TRANSWESD starts with the computation of shortest path lengths  $\mathbf{S}^{+/-}$ . Here, one faces an intrinsic algorithmic problem: in graphs containing negative cycles this problem is known to be NP-complete for elementary paths (?). Fortunately, one can check with low computational demand whether a negative cycle exists or not. If not, one may again use the double label algorithm computing exact results in polynomial time. Even if negative cycles exist, it turns out that exact shortest path computation is often possible in realistic cellular networks with several hundreds of nodes by using a depth first search or special variants thereof (?). The latter article also describes a polynomial algorithm that produces reasonable approximations in large-scale networks. A second technical issue concerns the interpretation of causality in negative cycles. In Fig. 5.1D, a small example of a perturbation graph containing the negative cycle  $C \rightarrow_{+,0.3} D \rightarrow_{-,0.4} C$  is shown. The key question is whether the negative non-elementary path (walk)  $A \rightarrow_{+,0.2} C \rightarrow_{+,0.3} D \rightarrow_{-,0.4} C \rightarrow_{+,0.3} D \rightarrow_{+,0.3} B$  is considered as a valid explanation for the negative influence  $A \rightarrow_{-,0.6} B$  observed when perturbing  $A$ . With  $\psi < 1$ , sign and length of this walk would support such a reasoning. ? considered walks as possible explanations and although ? did not consider weights, their approach is also based on this interpretation. This brings the advantage that one only needs to compute the shortest positive/negative walks, which is computationally easy, e.g. by an adapted Floyd-Warshall algorithm (??), in contrast to shortest elementary paths. For the following reason, the negative edge between  $A$  and  $B$  is kept by TRANSWESD: As described in Sec. 4.2, provided data mostly reflect the network in steady state after perturbation of  $A$  (without loss of generality an over-expression in  $A$  is assumed). The negative edge in the perturbation graph in Fig. 5.1D indicates that a decreased activation level of  $B$  was measured. From system theory (?), one can prove that the graph without this edge cannot show a decrease in  $B$  upon constitutive over-expression of  $A$  when the response of  $B$  is measured in transient or steady-state phase. The initial response in a network is governed by the sign of the elementary paths. Since removal of edge  $A \rightarrow_{-,0.6} B$  would imply that only a positive elementary path from  $A$  to  $B$  remains, the initial response would be positive in  $B$  (simply speaking, the effect of the positive path cannot be overtaken by the effect of the negative feedback induced by this path when looking at the initial response in  $B$ ). Also in steady state,  $B$

cannot exhibit a decreased activity (compared to unperturbed wild-type) if the negative edge from  $A$  to  $B$  is removed. If only positive elementary paths from  $A$  to  $B$  exist, a negative feedback can induce an opposite effect in steady state only in conjunction with other structural requirements including positive feedbacks (?). Albeit a negative effect in  $B$  might be observed transiently, one generally considers non-elementary paths containing a negative cycle as not sufficient for explaining an edge; only elementary paths with appropriate sign and weight are accepted. The negative edge from  $A$  to  $B$  is, therefore, kept in Fig. 5.1D. A third problem that may arise in cyclic graphs is non-uniqueness. An advantage of TRANSWESD is that edge weights eliminate many possible sources of non-uniqueness, in particular those related to positive cycles. Figure 5.1E depicts an unweighted perturbation graph containing a positive cycle. The positive edge from  $A$  to  $B$  could be explained by the positive path  $A \rightarrow_+ C \rightarrow_+ B$ . On the other hand, the positive edge from  $A$  to  $C$  could be explained by the positive path  $A \rightarrow_+ B \rightarrow_+ C$ . Methods based on unweighted perturbation graphs as in ? will thus remove one of both edges and keep the other. The choice depends on the edge processing order. With additional information on association strengths (edge weights) a unique solution can often be found with  $\psi < 1$  as shown in Fig. 5.1F: one would remove the edge from  $A$  to  $B$  as it can be explained by the positive path from  $A$  to  $C$  via  $B$  whose overall length (in MAX-metric) is shorter than that of the edge whereas the edge from  $A$  to  $C$  would be kept. However, even with edge weights non-uniqueness may occur as illustrated in Fig. 5.1G. In a first step, one may remove edge  $A \rightarrow_{+,0.8} C$  (with  $\psi = 0.95$  explainable by path  $A \rightarrow_{+,0.5} B \rightarrow_{+,0.6} C$  or, alternatively, by  $A \rightarrow_{+,0.35} D \rightarrow_{+,0.4} B \rightarrow_{+,0.6} C$ ). In a second step, one may either remove edge  $A \rightarrow_{+,0.5} B$  (explainable by  $A \rightarrow_{+,0.35} D \rightarrow_{+,0.4} B$ ) or edge  $D \rightarrow_{+,0.4} B$  (explainable by  $D \rightarrow_{-,0.3} E \rightarrow_{+,0.2} C \rightarrow_{-,0.25} B$ ). One can only remove one of both and then have to stop pruning because otherwise no explanation for the removed edge  $A \rightarrow_{+,0.8} C$  would remain in the network and thus violate condition Eqs. (5.10,5.11). Hence one may end up with two possible minimal solutions for the reconstructed graph. In general, such case can only occur if for a given edge at least two explaining paths exist and, again, if the network contains negative cycles. In TRANSWESD, a greedy strategy is used, i.e. in each iteration the explainable edge with largest weight (lowest association strength) is removed obeying condition Eqs. (5.10,5.11).

### 5.1.6 *In silico* application: DREAM4 challenge

This subsection, presents part of the results published in ?.

The fourth challenge of the Dialogue of Reverse Engineering Assessments and Methods (DREAM4) on *in silico* gene network reconstruction was used as a testbed for TRANSWESD in the form described above (Sec. 5.1.1-5.1.5.3). The DREAM project offers a platform for objective assessment of rivaling methods based on *in silico* data providing a realistic scenario for high-throughput gene expression profiling and reconstruction of gene regulation networks (???). The dataset used for testing TRANSWESD

## 5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS

---

belongs to the Insilico-Size-100 subchallenge and can be downloaded from the DREAM website ([www.the-dream-project.org](http://www.the-dream-project.org)). It comprises 5 sub-networks of 100 nodes sampled from gene networks of *Escherichia coli* and yeast, realistic kinetic models with randomly selected parameters were generated and simulated with GeneNetWeaver (?) using stochastic differential equations. For reconstructing these networks, *in silico* data were provided containing noisy steady-state mRNA expression levels of wild-type and single-gene knockout and knockdown experiments as well as time-series data. The gold standards of the five networks were provided after announcing the results of all submissions. Thus, an independent assessment of participating groups with different reconstruction methodologies was possible.

For each network, the perturbation graph was generated as described in Sec. 5.1.2 using the wild-type and knockout steady-state data. Edge weights were computed as conditional correlation coefficients from knockout and knockdown data. The results were very similar when using the knockout data only. The provided time series data were not used at all. Then TRANSWESD was applied to the generated perturbation graphs yielding the final reconstructed graph.

For benchmarking based on AUROC (area under the receiver operator characteristics curve) and AUPR (area under the precision-recall curve) values, potential edges were sorted according to their weights. Receiver operator characteristics and precision-recall curves are used in machine learning for characterizing binary classifiers. In the case of network reconstruction, a binary classification is given by predicting an edge or no-edge. Although both characterization curves are related, they provide different perspectives of the prediction (?). Precision-recall curves can indicate how precise the most confident detected edges are. On the other hand, ROC curves indicate the overall trade-off between TP and FP detection rate. Precision as a measure of fidelity is defined as

$$prec = TP/(TP + FP) \quad (5.12)$$

and recall as a measure of completeness as

$$rec = TP/(TP + FN). \quad (5.13)$$

The ROC curve represents false positive rate ( $FPR=FP/(FP+TN)$ ) vs. true positive rate (equivalent to recall). Figure 5.2 gives an illustrative example on how ROC and PR curves can look like for different noise settings in the data. Here, it was assumed that the distributions of an edge score describing the truth are known, i.e. one distribution associated to true edges and one distribution associated to true no-edges. In the lower part of Fig. 5.2 the impact of the data quality (red, green, orange) on PR and ROC curves for a given threshold level can be seen. The threshold level is used for edge detection based on the edge score (e.g. z-score of gene expression levels). Threshold variations along the curves are also indicated. An ideal PR curve has 100% precision for all recall values. In practice, this is however not achieved. In Fig. 5.2 it is seen that

when decreasing the threshold, TPs are gained but on the cost of precision, since also more FPs are included. A comparison of different classifiers - in our case reconstruction algorithms - based on curves is a rather daunting task, especially when comparing to many different reconstruction settings. Therefore the areas under the respective curve are being used (AUROC and AUPR). Further details can be found in (??). For the DREAM4 challenge, TRANSWESD obtained an overall score of 64.715, being ranked 3<sup>rd</sup> place out of 19 submissions. The winning team had an overall score of 71.5889. The overall score represents a log-transformed average of AUROC and AUPR p-values over the 5 networks and is computed as  $\text{score} = -0.5 \log_{10} (\langle P_{\text{AUROC}} \rangle \langle P_{\text{AUPR}} \rangle)$ . Further details on the scoring metrics for this specific challenge can be found on the DREAM website ([www.the-dream-project.org](http://www.the-dream-project.org)). After a correction of a minor implementation bug and improved procedures for perturbation graph generation and transitive reductions (?), the score could be increased up to 88.7594 for TRANSWESD.

### 5.1.7 Summary TRANSWESD on one-perturbation at a time data

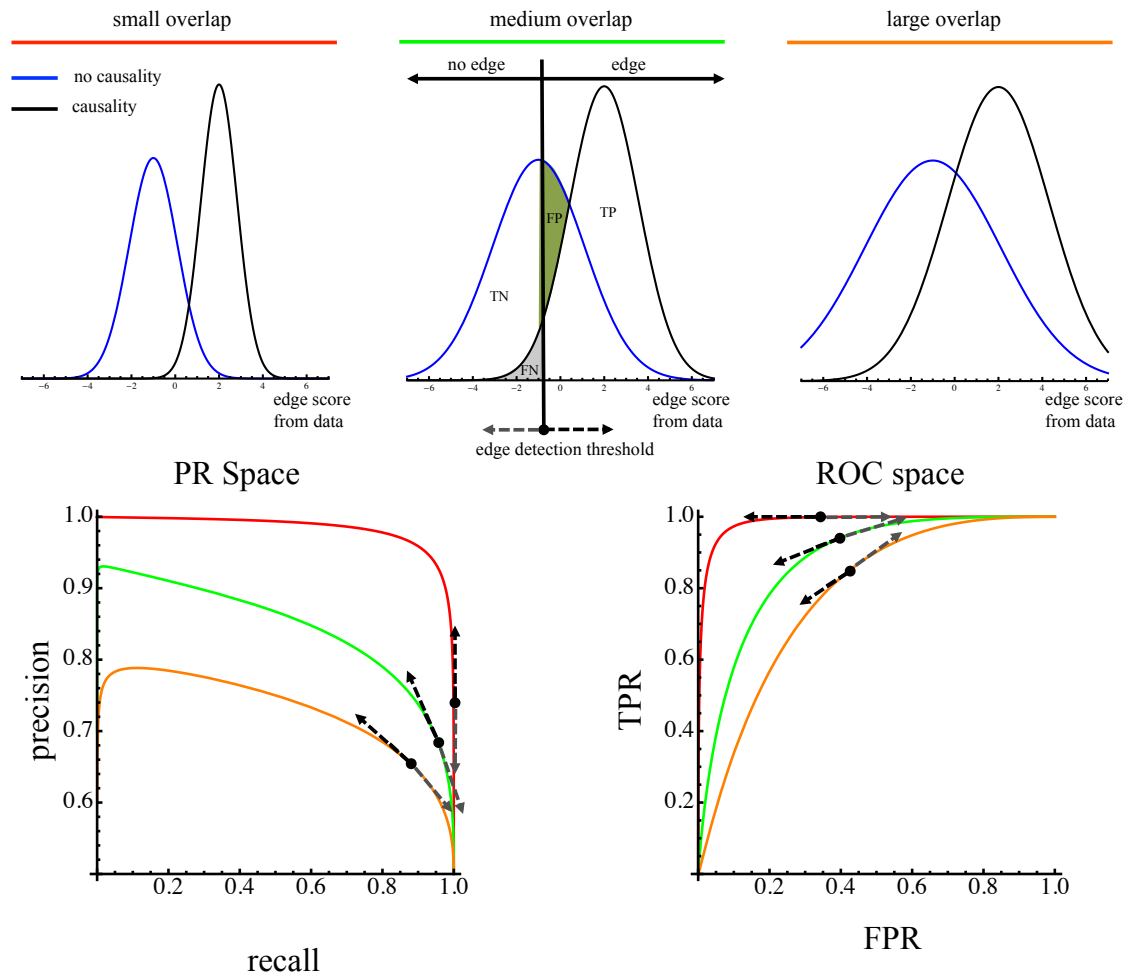
In this section 5.1 TRANSWESD was presented as a modular procedure for reconstructing gene regulatory networks. Starting from

1. data analysis and processing,
2. the perturbation graph is obtained providing the basis for
3. TRANSWESD to reduce FP predictions

to ultimately obtain the final reconstructed GRN. Step 2 for generating the perturbation graph consists of three sequential steps: (i) planning and conducting perturbation experiments, (ii) generation of signed perturbation graph from experimental data and (iii) edge weight (reflecting association strengths) association derived from correlation measures. Each of the above modules might be exchanged or adapted, e.g. if other types of data are available as is done in Sec. 5.2. Certain interactions may not be deducible from single perturbations or/and steady-state data and may require special perturbation strategies. For instance, only multiple knockouts, will detect a positive influence of one node upon another if this influence is combined with others via Boolean OR-logic. It is straightforward to integrate information of single and multiple perturbations when deriving the perturbation graph. Furthermore, data of the transient response combined with suitable data analysis could also be considered when generating the perturbation graph. Notice that depending on the specific perturbation data (transient, steady-state, time-courses) other, possibly nonlinear correlation measures such as mutual information might be better suited to quantify strengths of associations (?), though linear measures appear to be appropriate if monotone dependencies (unique edge signs) can be assumed.

In ?, TRANSWESD was optimized with regard to perturbation graph generation and weight derivation used to quantify path lengths. In the optimized versions of TRANSWESD, the perturbation graph is derived by a combination of z-score and correlation

## 5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS



**Figure 5.2:** Illustration of precision-recall and receiver-operating characteristics curves. Different settings (small, medium, large noise) of overlapping edge score (e.g. z-score of gene expression levels) distributions associated to a true edge (black) and true no-edge (blue) illustrate the impact on PR and ROC curves. For clearance vertical axes have been omitted on the distribution plots. Further indicated is the qualitative behavior when varying the edge detection threshold parameter (=discriminating edge score: for edge scores above/below an edge/no-edge is introduced).



measure, whereas different edge weights are used for the transitive reduction and edge sorting. A further improvement is obtained by restricting potential explanation paths for indirect effects to a maximal length of two edges. Since the optimized version of TRANSWESD follows the same principle as described in Sec. 5.1.5 we refer to ? for further details and in depth discussion on algorithmic improvements. Generally, the success of transitive reduction depends to a large extent on the quality of the perturbation graph, and thus on the properties of the available data. This includes signal to noise ratio, the type of data (e.g. gene expression, protein level, protein phosphorylation level, etc.) and the contribution of biologic variance itself, which all profoundly affect the observable perturbation effects. For an in depth analysis on an extensive data set, comprising different combinations of experimental and biological noise ratios as well as network topologies see ?.

## 5.2 Systems genetics

This section presents a framework for reconstructing gene regulatory networks from genetical genomics data where genotype and phenotype correlation measures are used to derive an initial graph, which is subsequently reduced by pruning strategies, including the core algorithm of TRANSWESD to minimize false positive predictions. Applied to realistic simulated genetic data from a DREAM challenge (DREAM5, subchallenge 3A), it is demonstrated that this simple approach is effective and outperforms more complex methods (including the best performer) with respect to (i) reconstruction quality (especially for small sample sizes) and (ii) applicability to large data sets due to relatively low computational costs. The presentation of the results as well as the results themselves are adopted from ? and ?.

Systems genetics approaches, in particular those relying on genetical genomics data, put forward a new paradigm of large-scale genome and network analysis. These methods use naturally occurring multifactorial perturbations (e.g. polymorphisms) in properly controlled and screened genetic crosses to elucidate causal relationships in biological networks. However, although genetical genomics data contain rich information, a clear dissection of causes and effects as required for reconstructing gene regulatory networks is not easily possible. In genetical genomics, a particular subclass of systems genetics, gene-expression levels are considered as phenotypic traits (called etraits) and identified QTLs (quantitative trait loci, comprising single genes or gene regions) are referred to as expression-QTLs (eQTLs). One application of eQTL maps obtained from genetical genomics approaches is the reconstruction of GRNs. According to ?, a GRN reconstruction pipeline for genetical genomics data consists of three major steps: (i) eQTL mapping, (ii) candidate regulator selection, and (iii) network refinement. Step (i) is used to identify chromosomal regions (eQTLs) that impact on expression levels (=traits) of genes. A detailed review on eQTL mapping is, for instance, given by ?. In step (ii), the eQTL map in combination with a genetic map is used to select single candidate

## 5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS

---

(regulator) genes from the eQTLs. Frequently used methods include conditional correlation (??), local regression (?), or analysis of between-strains SNPs (?). In the third step (iii), network refinement methods are employed to the topology obtained in step (ii), e.g. with the goal to identify and eliminate FP edges arising from indirect effects. Here, Bayesian network approaches (?) and structural equation modeling, SEM, (?) have been used.

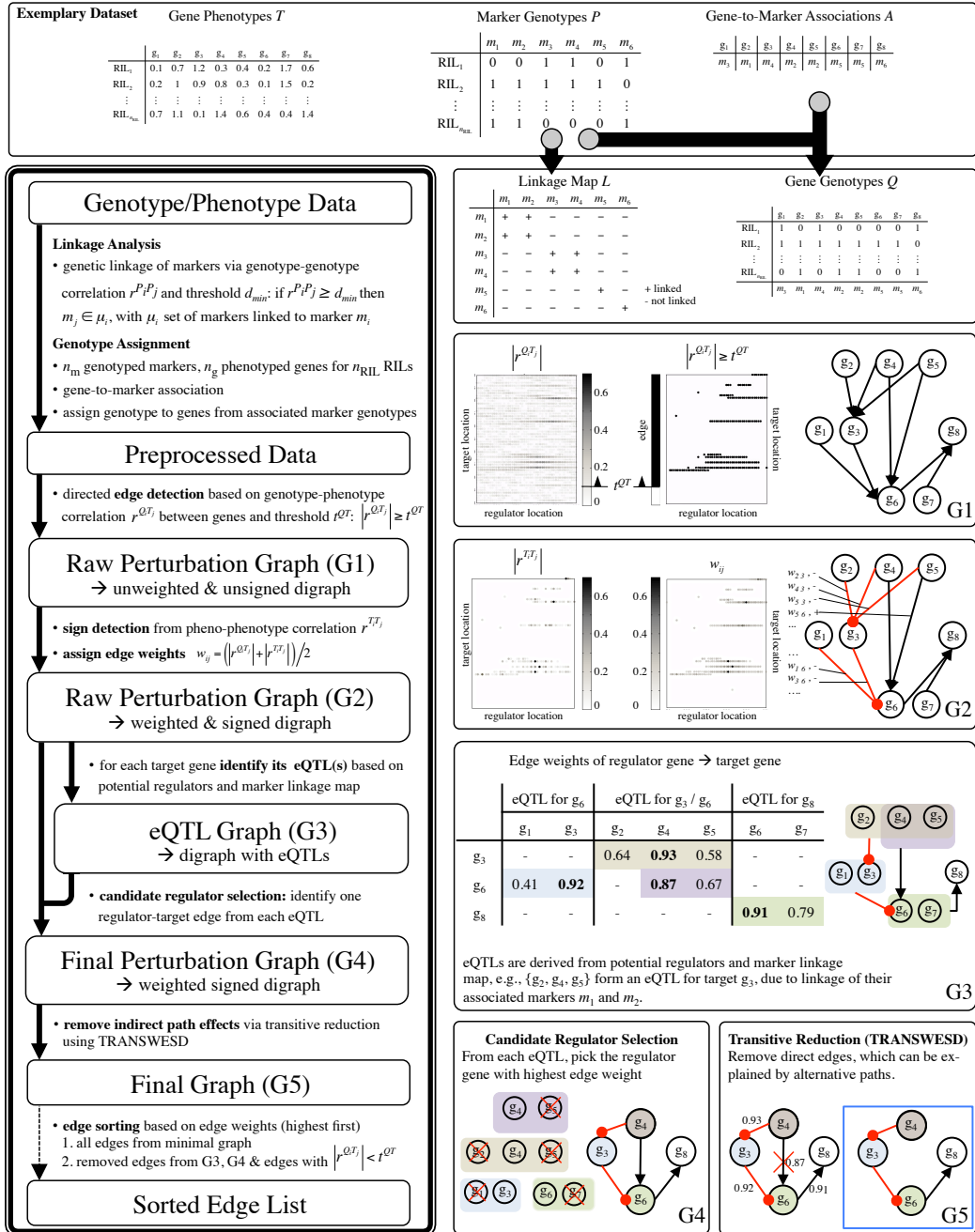
The here presented GRN reconstruction framework is tailored to genetical genomics data, which incorporates the three major reconstruction steps mentioned above in a modular fashion (?). The framework follows a simple-yet-effective paradigm: it is based on simple correlation measures, without the need for computational demanding optimization steps. This approach is therefore suited for small- and large-scale networks and performs well in the case of little samples but many genes, as illustrated in ? and ? using simulated and biological data.

### 5.2.1 GRN reconstruction on genetical genomics data

The workflow of the framework with a simple illustrative example is shown in Fig. 5.3. Starting from a typical set of genetical genomics data that include genotyped markers, phenotyped genes and gene-to-marker association, marker linkage analysis and genotype assignment for each gene is performed in a preprocessing step. From these data an unweighted and unsigned perturbation graph  $G_1$  is derived using genotype-phenotype correlation in combination with an appropriated thresholding strategy. The nodes in the graph directly correspond to genes while linkage information is kept to allow later eQTL assignment for each gene. The perturbation graph  $G_1$  is refined to  $G_2$  by quantifying each identified edge with respect to edge sign and weight, which indicate activation/repression and interaction strength, respectively. Due to genetic linkage true regulators may be masked by other genes, e.g. on adjacent positions on the genetic map, leading to eQTLs. EQTLs of a given target gene  $t_i$  can be identified on the basis of all potential regulator genes of  $t_i$  and the marker linkage map. These relationships are captured in graph  $G_3$ . Graph  $G_4$  is subsequently obtained by selecting one candidate regulator per eQTL based on the maximum of the edge weights.  $G_4$  is referred to as the final perturbation graph, whose edges reflect direct and indirect effects between genes induced by genetic variations. To remove indirect edges that can be explained by the operation of sequences of edges (paths) TRANSWESD is applied resulting in the final graph  $G_5$ . Optionally, if one is left to verify the interactions experimentally, it is desirable to have a list of edges sorted with respect to edge confidences. Such a list is also used by the DREAM5 evaluation procedures to assess the quality of a reconstructed network.

#### 5.2.1.1 Preprocessing genetical genomics data

A genetical genomics dataset typically consists of the following information (see Fig. 5.3): From a segregating population such as RILs, each segregant is genotyped for



**Figure 5.3:** Workflow of the proposed framework for reconstructing gene regulatory networks from genetical genomics data (left) with an illustrative example (top panel and right). For detailed explanations see text. Reproduced with permission of Oxford University Press from ?.

## 5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS

---

a set of polymorphic genetic markers that cover the genome or at least part of it. The genotype of each marker in each RIL is captured in a matrix  $P$  (e.g. two-valued (0/1) in the case of haploidic genomes). Additionally, genes are expression-profiled in each RIL (stored in a matrix  $T$ ). In the typical case that several genes are associated to one marker, genes can be associated to a specific marker based on their position on the genome map (yielding the list  $A$  in Fig. 5.3). From this information, one extracts by simple preprocessing steps two additional matrices needed before the actual reconstruction process is started. First, the gene-to-marker association  $A$  is used to assign, for each RIL, an (approximated) genotype  $Q$  to each gene, which is taken from its associated marker genotype  $P$ . This genotype assignment is based on the assumption of genetic linkage between markers and genes. Further, genetic linkage of the markers needs to be taken into account to identify potential eQTLs in G3 at a later reconstruction step. If genetic linkage of the markers is unknown, a linkage analysis can be performed based on genotype-genotype Pearson correlation of the markers  $m_i$  and  $m_j$  ( $P_i$  and  $P_j$  denoting their genotype). With a given threshold  $d_{\min} \in [0, 1]$ , then  $m_j \in \mu_i$  with  $\mu_i$  being the set of markers linked to marker  $m_i$ . By this procedure one obtains a linkage map  $L$ . The parameter  $d_{\min}$  represents the minimal genotypic correlation at which two markers are considered to be linked. The threshold can be derived from (i) testing for significance of deviation from zero by a t-test (e.g. appendix of ?), whereas empirical significant levels can be derived from permutation tests (??), (ii) the typical separation of candidate regulators in the eQTL map based on  $r^{Q_i T_j}$  (see Fig. 5.3 right, panel of G1). Specifically, one may analyze the average number of eQTLs over the genome as a function of  $d_{\min}$ . Regions of  $d_{\min}$  where the average number of eQTLs does not change much, indicate an optimal value. A similar thresholding strategy could be applied if genetic distances (given in centiMorgan) between the markers are known a priori.

### 5.2.1.2 Generating the raw perturbation graph from genetical genomics data

The next step is the generation of the perturbation graphs G1 and G2 from the (pre-processed) genetical genomics data. The idea for detecting a potential regulator-target interaction is - as in the case for one-perturbation at a time data - that a variation in a regulator gene's genotype causes a variation in the phenotype of the target gene.  $T_j$  is used to indicate the expression phenotype (etrait) of a gene  $j$  and  $Q_i$  for the genotype of a gene  $i$  (obtained from the marker genotype as described above). Based on the genotype-phenotype Pearson correlation coefficient  $r^{Q_i T_j}$ , an edge  $i \rightarrow j$  is assumed to exist, if it exceeds a threshold value  $t^{QT}$ :

$$|r^{Q_i T_j}| \geq t^{QT}. \quad (5.14)$$

The derived candidate edges reflect regulation of gene  $j$  by gene  $i$  by either *cis*, *cis-trans* or *trans* effects. In the case of  $i = j$  it is most likely a *cis* effect, otherwise one has to condition on  $i$ : if gene  $i$  has a *cis* effect then gene  $j$  is *cis-trans* regulated else it is

*trans* regulated. All three effects will result in increased correlations and Eq. (5.14) can thus be used to derive candidate edges for the GRN. The threshold  $t^{QT} \in [0, 1]$  can be selected based on a combination of several criteria, including (i) similar to the marker linkage analysis by p-values for rejecting  $|r^{Q_i T_j}| > 0$  based on a t-test, (ii) minimal/maximal edge numbers one expects to find in the GRN and (iii) existing data. In the case of small sample size, Spearman correlation might be more appropriate. For diploidic genomes, where  $Q_i$  is three-valued, one may apply the same procedure for each pairwise combination of genotypes and merge the resulting networks to G1.

Importantly, the nodes in the obtained graph G1 directly correspond to genes (as required to eventually reconstruct gene regulatory networks); the eQTL (regions) will be assigned later in graph G3 based on the linkage map  $L$ . Beforehand, edge signs and edge weights are assigned to each candidate edge  $i \rightarrow j$  in G1 resulting in G2 (Fig. 5.3). The edge sign  $s_{ij}$  is derived from via  $\phi(r^{T_i T_j})$ , i.e. the sign mapping applied on the correlation coefficient of expression levels of genes  $i$  and  $j$ . The strength  $w_{ij}$  of an edge is quantified by

$$w_{ij} = (|r^{Q_i T_j}| + |r^{T_i T_j}|) / 2. \quad (5.15)$$

The edge weight accounts for genotype-phenotype ( $QT$ -) and phenotype-phenotype ( $TT$ -) correlations by averaging both. This is especially important for (*cis*-)*trans*-regulated targets, as these are affected by both, geno- and phenotype of the potential regulator. We have also tested either  $QT$ - or  $TT$ -correlation alone, which in both cases led to reconstructed GRN of significant lower quality (at least when applied to the DREAM5/3A data). We also found that substituting the  $QT$  correlation coefficient in Eq. (5.14) by the average  $TT$ - and  $QT$ -correlations as used in Eq. (5.15) is not favorable, probably because then many high  $TT$ -correlation wrongly indicate a directed relationship, e.g. owing to common upstream regulators.

### 5.2.1.3 Identification of eQTLs and candidate regulator selection

Due to genetic linkage (correlated genotypes), a gene  $j$  that is found to be targeted by a gene  $i$  (i.e. an edge from  $i$  to  $j$  exists in G2) is typically also targeted by several other genes genetically adjacent to  $i$  resulting in an eQTL. An eQTL with respect to a given target gene  $j$  is identified by the set of all those genes that are potential regulators of  $j$  in G2 and that are genetically linked via their markers (see Fig. 5.3, G3: target gene  $g_3$  has one eQTL formed by  $\{g_2, g_4, g_5\}$ ). Importantly, two potential regulators  $g_i$  and  $g_j$  can be in the same eQTL, even in the case of their associated markers  $m_i$  and  $m_j$  being not linked in the linkage map  $L$ . This happens if there is another candidate regulator  $g_k$  whose marker  $m_k$  is linked to  $m_i$  and  $m_j$  in  $L$ . Note also that for each target gene, there may exist several eQTLs: in Fig. 5.3, gene  $g_6$  has two eQTLs formed by genes  $\{g_4, g_5\}$  and  $\{g_1, g_3\}$ . Once all eQTL(s) are identified for each target gene one arrives at the eQTL graph G3 (Fig. 5.3) in which the edges connect eQTLs with their target genes. G3 would represent the final result of classical eQTL mapping. If eQTL mapping was the envisioned goal, one could stop the procedure at this point. However, if the

## 5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS

---

reconstruction of a gene regulatory network is the ultimate goal then single candidate genes from each eQTL need to be selected. Since the probability is quite high that only a few or even only one of all the potential regulators of an eQTL are truly connected with the target gene, keeping all interactions in G2 that emanate from one and the same eQTL (the eQTLs being captured in G3) would result in many false positive predictions in the reconstructed network. Therefore we suggested selecting the candidate regulator from each eQTL with the maximal edge weight to be the true regulator of the target gene, i.e. for each eQTL identify  $i \rightarrow j$  with  $w_{ij} = \max(w_{kj})$ ,  $k \in \text{eQTL}$ , as the true edge and all other edges are removed from the eQTL. One then has the final perturbation graph G4 in Fig. 5.3 in which the nodes represent again genes.

### 5.2.1.4 Identifying and removing indirect effects (TRANSWESD)

Candidate regulator selection in the previous section leads to the reduced graph G4 where genetic linkage effects have been removed. One can now assume that edges in G4 reflect true causalities. However, an edge may still represent an indirect effect induced by a chain of interactions. For instance the effect of gene  $g_4$  on gene  $g_6$  in G5 of Fig. 5.3 is likely to be induced by the double-negative (thus positive) path  $g_4 \rightarrow -g_3 \rightarrow -g_6$ . The goal of this final step is therefore to identify and eliminate edges that arise from indirect effects. To this end TRANSWESD is used (?). TRANSWESD needs association weights  $\tilde{w}_{ij}$  between nodes of the graph, which can be directly derived from the edge weight via  $\tilde{w}_{ij} = 1 - w_{ij}$ , i.e. a low  $\tilde{w}_{ij}$  indicates a high association (small distance) between  $i$  and  $j$ , as in Eq. (5.5). After applying TRANSWESD the final, reconstructed graph G5 is obtained.

### 5.2.1.5 Sorted edge list

Optionally, a sorted list (ranking) of regulator-target interactions can be generated from the final graph G5, e.g. for prioritizing edges for experimental validation. One possible sorting is made up of two parts. The first part of the sorted edge list contains all edges from the final graph, sorted according to edge weights  $w_{ij}$  with highest weights (=most significant) first. In order to also account for edges that potentially have been wrongly dropped during thresholding (not contained in G1), cluster removal, or by TRANSWESD, the second part contains all of these removed edges, also sorted according to their edge weights Eq. (5.15) in descending order.

## 5.2.2 Applications

The following subsections illustrate the application of the developed framework to (i) synthetic genetical genomics data that were provided for a systems genetics challenge of the DREAM project (DREAM 5, subchallenge 3A), and (ii) real genetical genomics data from yeast, which were originally published in ?.

### 5.2.2.1 *In silico* application: DREAM5 challenge

The task of the systems genetics challenge of DREAM (DREAM5, subchallenge 3A) was to infer causal gene regulatory networks from phenotype expression data of a genotyped, segregated population. Due to lack of reliable experimental data sets for benchmarking different reconstruction algorithms, participants were given realistic *in silico* data which were generated by the SysGenSIM software (?). The provided simulated data represent (noisy) data from homozygous recombinant inbred lines (RILs), whereas the genome of each individual consists of 1000 genes and is made up of 20 chromosomes with 50 genes each. Five different networks of modular scale free topology had to be reconstructed for three different sample sizes (populations of 100, 300 and 999 RILs) eventually resulting in 15 reconstructed GRN. The haploidic genotype for all genes in all RILs was given as a binary vector (simulating the ideal situation of one marker per gene). The genotypes of adjacent genes were correlated mimicking genetic linkage. Each gene was assumed to have one functional genetic variant, either in the promotor (*cis* effect on the gene's expression rate) or coding region (*trans* effect on the target gene) of the gene. One motivation of the challenge was to analyze the reconstruction quality of participating methods when the population size becomes very small in comparison to the number of genes (e.g., 100 RILs / 1000 genes). For each sample size, the reconstructed networks in form of a sorted edge list (last step in Fig. 5.3) are passed to the evaluation script of DREAM (for details see ? and [www.the-dream-project.org](http://www.the-dream-project.org)). Since self-regulation was excluded by the challenge, candidate edges  $i \rightarrow j$  where  $i = j$  were removed. The evaluation of the reconstruction quality is based on AUROC and AUPR values derived from comparing the reconstructed GRN to the gold standard. The resulting overall score is based on empirical p-values computed from all submitted reconstructed GRNs (?). Applying the developed systems genetic reconstruction framework to the described DREAM5/3A data was straightforward and led the results presented in Table 5.1. The preprocessing step (see Fig. 5.3) was reduced to generating the marker/gene linkage map  $L$ , since each gene had its own associated marker. When applying the framework parameters  $t^{QT}$  and  $d_{\min}$  need to be specified. Two scenarios were considered for the threshold  $t^{QT}$ . First, based on the gold standards, optimal values for each of the three RIL population sizes were determined (delivering the highest overall score for all networks of this size), which were then used for all five networks. It turns out that smaller population sizes require larger threshold values: 0.23 for 100 RILs; 0.15 for 300 and 0.09 for 999. These optimal values correspond to p-values smaller than 0.01, when assuming t-statistic. In addition, for an unbiased scenario,  $t^{QT}$  was sampled uniformly in the range of 0.05...0.6, which defines a plausible range of maximal and minimal edge numbers contained in G1. Further average result over all networks (column G5\* in Table 1) are computed to have an estimate of an the average reconstruction performance. Regarding the parameter  $d_{\min}$  required for identifying eQTLs  $d_{\min} = 0.5$  was chosen after inspecting geno-phenotype relationships in the data (see Fig. 5.3, panel G2). As it turned out, results were extremely robust with respect to changes of  $d_{\min}$ . For example, the overall scores varied less than 1% when varying  $d_{\min}$  in a large range of 0.3-0.8.

## 5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS

---

**Table 5.1:** Reconstruction results for the DREAM5/3A: From G2 to G5 with optimal threshold parameter  $t_{QT}$ . Column G5\* shows averaged results when sampling parameter  $t_{QT}$  equally distributed on the interval  $[0.05 \dots 0.6]$ . Reproduced from ? with permission of Oxford University Press.

DREAM5	G2	G4	G5	G5*	best performer DREAM5/3A
averaged over 5 different networks					
100/AUPR AUROC	0.140 0.802	0.186 0.806	0.190 0.807	0.179 0.807	0.061 0.703
300/AUPR AUROC	0.215 0.883	0.342 0.887	0.345 0.887	0.286 0.887	0.148 0.786
999/AUPR AUROC	0.243 0.924	0.446 0.930	0.458 0.930	0.341 0.928	0.234 0.859
100/TP/FP	1138/35651	836/8310	769/3974		
300/TP/FP	1682/34153	1328/4699	1270/3657		
999/TP/FP	2371/51644	1832/3555	1721/2837		
100/score	193.77	231.24	236.11	214.84	81.87
300/score	170.54	237.3	238.51	188.63	89.4
999/score	172.67	250.04	251.69	191.71	140.56



In contrast, disregarding genetic linkage between markers by setting  $d_{\min} = 1$  lead to much lower reconstruction quality (see Tab. 5.1 overall scores of G2 vs. G4). This also holds for the other extreme case  $d_{\min} = 0$ , which would result in one large eQTL for all identified regulator candidates.

Several key observations can be made. Moving from the initial perturbation graph G2 to G4 by the candidate regulator selection approach, one can see a clear improvement with respect to FP reduction at minimal loss of TPs by one order of magnitude. For example, for the 999 RIL individual scenario, when transforming G2 to G4 the number of FPs reduces on average from 51644 down to 3368; whereas the number of TPs reduces only from 2371 to 1844. Undesired removal of TPs may occur by selecting the wrong regulator gene of an eQTL or because several genes from an eQTL target the same gene concurrently. In the second pruning step from G4 to G5, TRANSWESD removes many indirect edges due to alternative paths found in G4 improving in almost all cases the AUPR value. As the precision ( $TP/(TP+FP)$ ) of the reconstructed network increases significantly in all cases upon applying TRANSWESD one could expect an even better relative improvement of the AUPR value. However, there is only a moderate increase of the AUPR because TRANSWESD removes mainly edges with lower edge weight and thus with lower confidence (and ranking position) in the edge list. Generally, TRANSWESD works better for networks with lower connectivity (in DREAM5/3A, the edge density increases with increasing network index from approximately 2000 up to 5000 edges) and with larger sample size.

The made observations also hold for averaged scores from uniformly sampled threshold parameters  $t^{QT}$  (see G5\* in Tab. 5.1), i.e. the method is robust against threshold selection. Comparing the results of the proposed framework to the best performer of the DREAM5/3A challenge for each RIL sample size and different network topologies (last column in Tab. 5.1), one can see a clear improvement also for randomly chosen threshold parameters (G5\* in Tab. 5.1). Even without applying any FP reduction, G2 is almost always better than the best performer, although it was constructed based on a simple eQTL mapping approach alone. G4/G5 obtained after pruning are always better than the best performer. This also holds for an improved version of the best performer method (?). Further, moving from large to small sample sizes one can see a clear relative improvement, i.e. increase of the overall score (e.g. of G5\* averaged over the 5 different networks) with respect to all DREAM participant submissions. Consequently, the developed method performs especially well for small sample sizes. At a given sample size, the method has averaged AUPR values which are up to 3 times larger than the best performer in the case of 100 samples (G5\* vs. best performer averaged over the 5 networks). This shows, that for small sample sizes, a rather simple method based on pure correlation measures in combination with FP reduction methods seems to be the best choice, keeping in mind that many different methods have been used by the 16 participants in this specific reconstruction challenge of DREAM (?). However, even for the largest RIL populations provided in the DREAM5/3A challenge, the developed method still achieves significantly higher scores.

## 5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS

---

### 5.2.2.2 *In vitro* application: genetical genomics data of yeast

As a real life test case for the developed genetical genomics data based reconstruction method, genotypic and expression data from 112 segregants obtained from a yeast cross between BY and RM strains of *S. cerevisiae* (?) have been used. Only 1573 of all 2956 markers were associated to at least one of the 5736 expression-profiled genes. Further, a gene-to-marker association list  $A$  was available. In contrast to the DREAM5/3A data, there were thus much fewer markers than genes. After preprocessing the data by computing the linkage map  $L$  from the marker genotypes in the RILs with  $d_{\min} = 0.5$  (same value as for the *in silico* data), the reconstruction framework was applied to the matrices  $T$  and  $Q$  to obtain  $G2$ . The correlation threshold was set to  $t^{QT} = 0.23$ , corresponding to the determined optimal threshold for the *in silico* data with 100 RILs (as this is closest to the 112 RILs available in this study). The two parameters were thus unbiased and not specifically optimized for this dataset. Using the linkage map  $L$ , the eQTL graph  $G3$  was obtained, which was used to derive the final perturbation graph  $G4$  by selecting from each eQTL the gene-target interaction in  $G2$  that has the highest edge weight.

In the literature and in some databases one can find published (most likely sub-networks) of the yeast gene regulatory network, whose interactions have been identified from different sources, including ChIP-chip and motif finding studies (??). However, a rigorous evaluation of the inferred network is not trivial, as the reliability of the gold standards from the sources mentioned above is unclear. Therefore the following strategy for evaluating the quality of the reconstructed network was followed: In another challenge of DREAM5 (subchallenge 4.4; not to be confused with the systems genetics subchallenge 3A described in the previous Sec. 5.2.2.1) the goal was to infer a subpart of the yeast GRN focusing on 333 candidate transcription factors (TF) and their interactions with 5950 (potential target) genes based on 536 microarrays each containing expression profiles for a given perturbation (e.g. specified gene knock-out or over expression, including partial replicates). The evaluation in this challenge was based on a given gold standard containing interactions considered to exist between the genes of the given transcriptions factors and all other genes (by the time of the analysis, it was not known to me nor my collaborators how this gold standard was compiled), whereas self-regulation has been excluded. Most genes (5451) of the data from ? are present in the DREAM5/challenge 4.4 yeast gold standard. To compare the absolute performance with respect to this gold standard and relatively to the other 29 participants of this challenge, a subgraph of the reconstructed yeast GRN was evaluated, which was restricted to potential interactions between the 333 TFs as regulators and the 5451 target genes present in DREAM5/4.4. The specific DREAM5 evaluation script was used, which computes the AUPR and AUROC (and their p-values) of an inferred network with respect to the provided gold standard. The results are presented in Table 5.2. First observe that even though the reconstruction was based on only 112 RILS samples (compared to the large number of 536 microarray experiments available to the participants of this challenge), the reconstruction belongs to the very best of the submissions (rank 1 for

**Table 5.2:** Reconstruction results for the yeast genetical genomics data set ? compared to the yeast gold standard of DREAM5/4.4. The first row shows (separately for G2, G3 and G5) the AUPR, the p-value of AUPR and the (virtual) rank within the DREAM5/4.4 AUPR performance ranking (total number of participants: 29). The second row gives the same values with respect to AUROC. Reproduced from ? with permission of Oxford University Press.

	G2	G4	G5
AUPR/ $p_{\text{AUPR}}$ /rank	0.0274/5.7E-11/4	0.0293/2.34E-14/3	0.0293/1.89E-14/3
AUROC/ $p_{\text{AUROC}}$ /rank	0.5396/6.7E-28/1	0.5407/6.14E-30/1	0.5407/6.4E-30/1

AUROC, rank 3 for AUPR). Therefore, as observed also for the *in silico* data in Sec. 5.2.2.1, the performance of the reconstruction framework proves again its suitability for small sample sizes. In Table 5.2 one can also see that the FP pruning strategies always improves the precision-recall and associated p-values when moving from G2 via G4 to G5. The same holds for the AUROC score, except that in G5 a minor reduction of the p-value can be observed due to the loss of some TPs during FP reduction.

Regarding the absolute quality of the reconstructed network, notice that it does not meet the high scores of the *in silico* challenge DREAM5/3A, Sec. 5.2.2.1. There are several possible reasons for this behavior. First, the amount and resolution of the *in silico* data quality is much better, in particular, there was one marker per gene (whereas only 1573 markers for 5736 genes are available in the yeast data). Although noise has been added to the *in silico* data, it might be much higher under realistic conditions or/and other sources of uncertainty might also hamper the visibility of true interactions as illustrated in ?. Furthermore, the given gold standard for the yeast transcriptional network cannot automatically be considered to be the full truth. The similar low or even worse quality of reconstructed networks submitted by the other participants for DREAM5/4.4, may, at least partially, point to missing or false edges in the gold standard itself. To test the relevance of the inferred networks, it would therefore be interesting whether the top-ranked interactions of the reconstructions (not present in the gold standard) could be validated in experiments.

### 5.2.3 Summary TRANSWESD on systems genetics data

In this section a simple yet effective modular framework for gene regulatory network reconstruction from genetical genomics data was presented. In the case of the DREAM5/3A *in silico* data, the methodological framework was shown to outperform the best performer (even in the non-tuned case), who applied a combination of Bayesian network analysis, LASSO, and the Dantzig selector (?). In the case of real data, the DREAM5/4.4 yeast GRN gold standard has been used to assess the quality of the yeast GRN inferred by the framework applied to genetical genomics data relatively to networks inferred by classical perturbation experiments and microarray data. The performance of the reconstructed network compares to the very best of the submissions, although here only 112 RILs have been used, in contrast to DREAM5/4.4 submissions,

## 5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS

---

which were based on 536 microarrays with well-defined perturbations. Consequently, these results indicate that simple correlation methods paired with subsequent FP pruning strategies outperform complex methods, especially for small sample sizes (experimentally still most relevant). This is most likely due to a larger noise sensitivity of multi-locus method in contrast to univariate correlation analysis. Since correlation-based eQTL mapping yields many true positive, but also many false positive interactions, a local pruning based on linkage information and a global pruning based on path knowledge is important.

The proposed framework performs best on data with one marker per gene, which might be realistic for future ultra-high-throughput sequencing methods. As illustrated in Sec. 5.2.2.2, the framework can be readily applied to the general case where markers cover several genes and it can also be adapted easily to cases with more than two different genotypes. The presented local pruning approach for genetic linkage assumes that only one regulator gene is selected per eQTL, which therefore cannot account for the case that a target has several regulators within one eQTL. A relaxed selection strategy can be used based on partial correlation to potentially select more than one candidate regulator per eQTL (?). In the case of complex traits, i.e. a trait is influenced by several eQTLs, the univariate approach to generate G1 based on correlating one  $Q_i$  with one  $T_j$  at a time possibly misses combinatorial effects and could potentially result in a higher number of false-negative regulator-target interactions. Alternatively, a multi-locus method as the Lasso (?), the elastic net (?) or the random forests (?) might be used. However, as has been pointed out by ?, for very large data sets (millions of dense markers and phenotyped genes), multi-locus methods cannot be used due to computational overload. Here, the presented approach provides a computationally feasible and effective framework for filtering the most important interaction sites (on which multi-locus approaches can be applied) as it does not use any optimization algorithm.

In order to apply the developed framework, one needs to specify threshold parameters. Several ways for determining optimal values based on an explorative data analysis have been discussed. As was further shown, the reconstruction framework allows an adjustment to other reconstruction methods, i.e. modules in the workflow (Fig. 5.3) may be replaced. For instance the candidate regulator selection based on the eQTL map G3 may be exchanged by other approaches, e.g. partial correlation or local regression. Although the framework is based on a univariate analysis it can provide reconstructed GRN at higher precision-recall level than advanced multi-locus methods, even for smaller sample size. This might not hold when combinatorial or epigenetic effects (?) are present, where multi-locus approaches may become advantageous (?). Therefore, in line with the key result of the DREAM initiative stating that community efforts based on many different reconstruction methods produce best results (?), meta-methods, e.g. as proposed by ?, should combine both simple and complex methods.

### 5.3 Summary

This chapter introduced and analyzed TRANSWESD, a FP reduction methodology in a modular framework for reconstruction gene regulatory networks. Overall, the framework constitutes simple, exchangeable modules, as has been illustrated when adapting the original TRANSWESD for one-factorial data to multi-factorial data. As was shown by several examples, TRANSWESD including its preprocessing procedures is a powerful approach despite its simplicity and especially suited for large-scale reconstruction or filtering. Finally, note that we have successfully applied the z-score approach of TRANSWESD in a species translation network reconstruction challenge provided by sbv IMPROVER platform (??). Here we were ranked 1<sup>st</sup> place (?). Just like the DREAM project, sbv IMPROVER provides biological challenges focused on industrial applications of systems biology (??) and enables an objective evaluation of methods.

## **5. TRANSWESD: A REVERSE ENGINEERING ALGORITHM FOR IDENTIFYING GENE REGULATORY NETWORKS**

---

## 6

# Concluding Remarks

*It is perfectly true, as philosophers say, that life must be understood backwards. But they forget the other proposition, that it must be lived forwards. And if one thinks over that proposition it becomes more and more evident that life can never really be understood in time simply because at no particular moment can I find the necessary resting place from which to understand it backwards.*

---

Søren Kierkegaard  
Journals and Papers, 1843

## 6.1 Summary

This thesis presents methodological solutions to model identification problems within the application field of systems biology, more generally, biochemical systems. The solutions are (i) a robust experimental stimulus design methodology for supporting computational model identification of dynamical processes in biochemical reaction systems and (ii) a modular framework for structural network identification based on high-throughput data. The developments were driven by challenges that come along with this specific class of systems, which are not as often found in physical, nonliving systems analysis. Most of the nonliving systems are very well characterized and understood, whereas biochemical systems are often only understood on a qualitative basis owing to their inherent complexity paired with strong biological variability. Further, experimental probing of biochemical systems is not as straightforward as is the case for physical systems.

In this challenging setting of strong inherent variability, difficult experimental probing, vague mechanistic knowledge and complexity (with regards to the number of players involved, but also with regard to the emergent properties), we are in the need of sophisticated methods, that help using our existing skills in an optimal way, to ultimately follow the engineering paradigm of model-based understanding and design.

## 6. CONCLUDING REMARKS

---

In this work, a *new experimental design approach* with the focus on robust stimulus design aimed at model discrimination was created (Ch. 3). The approach has been benchmarked with several *in silico* test cases. As was shown, it is ideally suited for nonlinear models that have highly uncertain parameters enabling researchers to robustly design complex stimulus profiles with little computational effort. In this way, model-based predictions are robustified against parametric uncertainties, which represent experimental and inherent biological variabilities.

The developed stimulus design method was applied within an interdisciplinary research project (Prof. Dr. rer. nat. Michael Naumann, adj. Prof. Dr.-Ing. Michael Mangold, Dr. rer. nat. Michael Wulkow, Prof. Dr.-Ing. Kai Sundmacher). Here, we could *demonstrate a cyclic workflow, including experiments, experimental design and computational modeling*. In this way, a predictive, dynamic signaling model for DNA damage detection could be identified, that allowed predicting individual contributions to a specific protein modification, which is essential for DNA repair pathway initiation (Ch. 3). Within this research project we thus exemplified how difficult experimental tasks can be transferred to computational modeling challenges by using the model as a crutch for understanding biological signaling processes. By applying statistical methodologies for data and model analysis, including the developed optimal experimental design methodology, a predictive model could be used to generate verifiable predictions.

Finally, in collaboration with the group of Dr.-Ing. Steffen Klamt, *TRANSWESD was developed as a framework for large-scale network reconstruction based on one- or multi-factorial perturbation data* (Ch. 5). The framework has been benchmarked on international platforms for *in silico* network reconstruction problems and proven to be one of the leading approaches for structural network identification. Further, the performance on a real life application underpinned the strength and applicability of TRANSWESD, although it was shown, that there is still a discrepancy between *in silico* and *in vitro* performance. Despite this discrepancy, *in silico* benchmarking is an important tool for identifying strengths and weaknesses of reconstruction methodologies.

In total, the most important achievements of this work are methodological advances in mathematical model identification tailored to handle the challenges of biochemical systems with a high degree of variability and complexity. The achievements may be summarized as follows:

- A new experimental design methodology for robustly designed stimulus profiles, aimed at providing data for best model discrimination in the presence of parameter uncertainties
- An illustration of an iterative model identification procedure of a dynamic cell signaling model, including experimental design and model analysis
- A new dynamic model for DNA damage signaling upon ionizing irradiation
- TRANSWESD, a framework for large-scale reconstruction of biochemical reaction networks including benchmark tests



- An illustration of TRANSWESD to a real life test case

Also important to note that TRANSWESD has been acknowledged as one of the state-of-the-art network reconstruction algorithms. In international reconstruction challenges it has been ranked 3rd and 1st place (??).

## 6.2 Conclusion and Outlook

From a high level point of view, identification of dynamic models or large-scale network structures use the same methodological principle, namely statistical classification, which is most likely one of the key approaches to the identification of reliable, biochemical models. Although both contributions were either driven by computational model identification of dynamic signaling processes or by reconstructing gene regulatory networks, they are - taken cum grano salis - not restricted to these specific applications. This is simply the consequence of the fact that one can reinterpret the entities described by the mathematical formulas. Therefore, depending on the specific application, the developed methodologies help distinguishing (i) plausible from unlikely dynamic ODE models and (ii) edges from no-edges in large-scale interaction networks.

Even though methodological advances have been quite impressive over the last years within the field of computational modeling of biological processes, one should keep in mind that only the inclusion of experimental advances will help reaching the next level of systems understanding. This work has also given a flavor in this direction by illustrating an interactive, highly integrative working mode that combines experimental and modeling strengths to increase biochemical systems understanding. Systems understanding includes new knowledge about how molecules interact and respond to ex- or internal stimuli and how this impacts on the phenotype of the cell, tissue, organ and ultimately the living organism. This knowledge allows designing strategies to influence and modify living organisms and bears huge potential in the area of biomedical applications by, for instance, reducing the time of clinical trials for vaccine candidates or by optimizing cancer therapy with individualized protocols based on the patient's bio-marker profile, which can serve as a bio-fingerprint (?).

Additionally, with emerging knowledge, understanding and LEGO<sup>®</sup>-like biological building bricks (?), synthetic biology should pave the way to a rational design of biochemical devices and reaction systems either from scratch or via modification of existing biological building blocks. Examples for such applications include diagnostic tools for patients' care or optimized strains for biological productions processes (e.g. yeast or fungal based production of proteins for industry or medical application (????)). Finally, one should note that such a technology should not miss a sound dispute about ethical questions and responsibilities, in order to explore implications of manipulating life by means of adaption or even creation (?).

## 6. CONCLUDING REMARKS

---

# Appendix A

## Supplementary methodological information

### A.1 NLP problem formulation for optimal experimental stimulus design

For completeness of the thesis, the NLP problem formulation is included. It is taken from the supplementary material of ?.

#### A.1.1 Direct sequential approach

Within the direct sequential approach, the stimulus is parameterized, whereas the system dynamics is solved by numeric integration. The basic structure of the resulting NLP problem is given by

$$\mathbf{U}_{\dagger}^{\mathcal{E}} = \arg \min_{\mathbf{U} \in \mathbf{U}_{\text{CD}}} O^{\mathcal{E}}(\mathbf{U}) = \mathbb{E} \left[ \Phi^{\mathcal{N}} \left( \mathbf{E}_{t_{A|B}}^{\mathcal{E}}[\mathbf{Y}], \mathbf{C}_{t_{A|B}}^{\mathcal{E}}[\mathbf{Y}] \right) \right]_{\mathbf{t}} \quad (\text{A.1})$$

subject to:

$$\frac{d}{dt} \mathbf{x}_m(t) = \mathbf{f}_m(\mathbf{x}_m(t), u(\mathbf{U}, t), \boldsymbol{\theta}_{\mathbf{x}_m}) \quad (\text{A.2})$$

$$\mathbf{x}_m(t_0) = \mathbf{x}_{0_m} \quad (\text{A.3})$$

$$\mathbf{y}_m(t) = [x_{2m}^*(t), x_{3m}^*(t)]^{\text{T}} \quad (\text{A.4})$$

$$x_{im}(t) \geq 0 \quad ; \quad y_{jm}(t) \geq 0 \quad (\text{A.5})$$

$$u_{\min} \leq u(\mathbf{U}, t) \leq u_{\max} \quad (\text{A.6})$$

$$dt_{\min} \leq dt_k \leq dt_{\max} \quad (\text{A.7})$$

$$\sum_{k=0}^{(n_t-1)} dt_k \leq (t_f - t_0) \quad (\text{A.8})$$

$$+\text{method to estimate } \mathbf{E}_{t_{A|B}}^{\mathcal{E}}[\mathbf{Y}] \quad \text{and} \quad \mathbf{C}_{t_{A|B}}^{\mathcal{E}}[\mathbf{Y}], \quad (\text{A.9})$$

## A. SUPPLEMENTARY METHODOLOGICAL INFORMATION

---

where  $i \in \{1, 2, 3, (4)_B\}$  index the species,  $j \in \{1, 2\}$  the response measurement signals and  $m \in \{A, B\}$  the models. The measurement time points are being fixed and placed according to  $\mathbf{t} = [t_0, t_1, \dots, t_{(n_t-1)} = t_f]^T$  with constant  $dt = \frac{t_f - t_0}{n_t}$ .

### A.1.2 Direct simultaneous approach

A full discretization of the dynamic system equations by means of orthogonal collocation results into nonlinear constraints in the form of nonlinear algebraic equations, which represent an implicit Runge-Kutta scheme (?). The advantage of this approach is the straight forward implementation of additional nonlinear path constraints and the availability of powerful NLP solvers. The basic structure of the resulting NLP problem is given by

$$\mathbf{U}_\dagger^\varepsilon = \arg \min_{\mathbf{U} \in \mathbb{U}_{\text{CD}}} O^\varepsilon(\mathbf{U}) = \mathbb{E}[\Phi^N(\mathbf{E}_{t_{A|B}}^\varepsilon[\mathbf{Y}], \mathbf{C}_{t_{A|B}}^\varepsilon[\mathbf{Y}])]_{\mathbf{t}} \quad (\text{A.10})$$

subject to:

$$x_{lk_{A|B}} = x_{l0_{A|B}} + dt_{l_{A|B}} \sum_{j=0}^{n_c-1} \mathbf{W}_{kj} f_{A|B}(x_{lj_{A|B}}, \mathbf{U}_l, \boldsymbol{\theta}_{\mathbf{x}}) \quad (\text{A.11})$$

$$x_{l-1n_c-1_{A|B}} = x_{l0_{A|B}} \quad \forall l = 2 \dots n_{FE} \quad (\text{A.12})$$

$$x_{10_{A|B}} = x_{0_{A|B}} \quad (\text{A.13})$$

$$y_{lk_{A|B}} = x_{lk_{A|B}} \quad (\text{A.14})$$

$$u_l = u_{l+1} = \dots = u_{l+b} \quad \forall l = 1, (2+b), \dots \leq (n_{FE} - b) \quad (\text{A.15})$$

$$+ \text{additional box constraints on all variables,} \quad (\text{A.16})$$

with  $n_{FE} = 100$  representing the number of finite elements, indexed by  $l = 1 \dots n_{FE}$ ,  $n_c = 3$  being the number of collocations points in each finite element, indexed by  $k; j = 0 \dots (n_c - 1)$ . Index  $b \leq n_{FE}$  represents the number of subsequent design variables that cannot be varied, i.e. the minimal time window of stimulus change.  $\mathbf{W}_{kj}$  is the collocation point weighting matrix and  $f_{A|B}$  the model dependent right hand side in the ODE system. When using the linear design strategy, the sensitivity equation and corresponding constraints (analogous to the system state discretization  $x(t)$ ) were implemented. For the sigma point design, Eqs. (A.11-A.16) have to simultaneously hold for the  $(2n_\theta + 1)$  sigma points. For the presented results, box constraints were chosen in such a way, that switching between two stable states might be induced ( $0.5 \leq U_l \leq 2.5$ ).

## A.2 Transformation of log-normal to normal PDF

This section concerns the log-transformations applied in Sec. 3.4.1, in order to account for log-normality. In the case of uncorrelated parameters one can start directly from the log-normal moments of the parameter PDF  $\mathbf{E}_{\log}[\boldsymbol{\Theta}]$  and  $\mathbf{C}_{\log}[\boldsymbol{\Theta}]$  to derive the corresponding normal moments, i.e. the moments of the normal PDF that characterizes the

## A.2 Transformation of log-normal to normal PDF

---

parameter on the logarithmic scale. In detail, if a parameter  $\Theta$  has a log-normal PDF, then  $\tilde{\Theta} = \log(\Theta)$  is normally distributed. For each model parameter  $\Theta_i$ , the transformation of the log-normal moments (expectation and variance) in one-dimensional form to the corresponding normal moments of  $\tilde{\Theta}$  is given by (?)

$$E_{\text{norm}}[\Theta_i] = \log(E_{\log}[\Theta_i]) - \frac{1}{2} \log \left( 1 + \frac{\sigma_{\Theta_i}^2}{E_{\log}^2[\Theta_i]} \right) \quad (\text{A.17})$$

$$\sigma_{\text{norm}}^2 = \log \left( 1 + \frac{\sigma_{\Theta_i}^2}{E_{\log}^2[\Theta_i]} \right), \quad (\text{A.18})$$

whereas  $\sigma_{\Theta_i}^2$  represents a diagonal element of  $\mathbf{C}_{\log}[\Theta]$ . These moments can be used to derive the sigma points  $\text{SP}_{\text{norm}}$  for the normal parameter PDF. The next step consists of exponentiating the normal sigma points to obtain the corresponding sigma points on the original logarithmic scale in the parameter space

$$\text{SP}_{\log} = \text{Exp}(\text{SP}_{\text{norm}}).$$

Then, the ODE system is solved for the resulting  $\text{SP}_{\log}$  to yield the propagated sigma points, i.e. model response trajectories on the normal scale. These model response trajectories can then be used to calculate the estimates for response expectations and variance-covariances.

If the parameter distribution can be represented by samples, which show clear correlations, one can simply log-transform the samples and derive estimates for the normal moments from these. Then again, calculate the sigma points, exponentiate and propagate these through the ODE system. Note that this discussion applies to any transformation moving an arbitrary, multivariate skewed PDF to normality (or at least close to normality). Within the linearization approach, one simply linearizes the response in logarithmic form with respect to the parameters.

### A.3 Supplementary *in vitro* application information

#### A.3.1 Model equations

The model equations are scaled to the total concentration of  $[\text{Ku7080}]_{\text{tot}}$  to make use of the intrinsic scale invariance of ODE in dimensional form to improve parameter estimation in terms of efficiency, see for instance supplement of ?. Therefore, brackets - usually indicating a protein in concentration units - have been dropped, as the states of the ODE then represent relative concentration levels and are thus dimensionless.

$$\text{initially damaged DNA:} \quad \frac{d}{dt} \text{DDNA1} = R_1 - R_2 \quad (\text{A.19})$$

$$\text{complex \{Ku7080:DDNA1\}:} \quad \frac{d}{dt} \text{RC11} = R_2 - R_3 \quad (\text{A.20})$$

$$\text{complex \{DNA-PK}_{cs}\text{:RC11\}:} \quad \frac{d}{dt} \text{RC12} = R_3 - R_4 - R_{6M} \quad (\text{A.21})$$

$$1^{\text{st}} \text{ phosphorylation step RC12:} \quad \frac{d}{dt} \text{RC12}^p = R_4 - R_5 \quad (\text{A.22})$$

$$2^{\text{nd}} \text{ phosphorylation step RC12:} \quad \frac{d}{dt} \text{RC12}^{pp} = R_5 - R_7 \quad (\text{A.23})$$

$$\text{complex damaged DNA:} \quad \frac{d}{dt} \text{DDNA2} = R_{6M} - R_{9M} \quad (\text{A.24})$$

$$\text{complex \{MRN:DDNA1\}:} \quad \frac{d}{dt} \text{RC20} = R_{10} - R_{11} \quad (\text{A.25})$$

$$\text{complex \{ATM:RC20\}:} \quad \frac{d}{dt} \text{RC21} = R_{11} - R_{12} \quad (\text{A.26})$$

$$\text{double phosphorylated ATM:} \quad \frac{d}{dt} \text{RC21}^{pp} = R_{12} - R_{15} \quad (\text{A.27})$$

$$\text{complex repair step:} \quad \frac{d}{dt} \text{RC22}^{pp} = R_{9M} - R_8 \quad (\text{A.28})$$

$$\text{repaired DNA:} \quad \frac{d}{dt} \text{RDNA1} = R_7 \quad (\text{A.29})$$

$$\text{repaired DNA:} \quad \frac{d}{dt} \text{RDNA2} = R_8 \quad (\text{A.30})$$

$$\gamma\text{H2AX:} \quad \frac{d}{dt} \gamma = R_{13} - R_{14} \quad (\text{A.31})$$

$$\text{total damaged DNA:} \quad \frac{d}{dt} \text{tDSB} = R_1 \quad (\text{A.32})$$

$$\text{phosphorylated p53:} \quad \frac{d}{dt} \text{p53}^p = R_{16} - R_{17} \quad (\text{A.33})$$

Corresponding rates

$$R_1 = \alpha_0 \frac{dD}{dt} u(t) \quad (\text{A.34})$$

$$R_2 = \alpha_{11} \text{DDNA1} \quad (\text{A.35})$$

$$R_3 = \alpha_{12} \text{RC11} \quad (\text{A.36})$$

$$R_4 = \alpha_{13} \text{RC12} \quad (\text{A.37})$$

$$R_5 = \alpha_{141} (1 + \alpha_{142} \text{RC21}^{pp}) \text{RC12}^p \quad (\text{A.38})$$

$$R_{6A1} = \alpha_{15} \text{tDSB RC21}^{pp} \text{RC12} \quad (\text{A.39})$$

$$R_{6B1} = \alpha_{15} \text{tDSB RC12} \quad (\text{A.40})$$

$$R_{6A2} = \alpha_{15} \text{RC21}^{pp} \text{RC12} \quad (\text{A.41})$$

$$R_{6B2} = \alpha_{15} \text{RC12} \quad (\text{A.42})$$

$$R_7 = \delta_{16} \text{RC12}^{pp} \quad (\text{A.43})$$

$$R_8 = \delta_{16} \text{RC22}^{pp} \quad (\text{A.44})$$

$$R_{9A12} = \alpha_{17} \text{DDNA2} \quad (\text{A.45})$$

$$R_{9B12} = \alpha_{17} \text{RC21}^{pp} \text{DDNA2} \quad (\text{A.46})$$

$$R_{10} = \alpha_{21} \text{DDNA1} \quad (\text{A.47})$$

$$R_{11} = \alpha_{22} \text{RC20} \quad (\text{A.48})$$

$$R_{12} = \alpha_{231} (1 + \alpha_{232} \text{RC21}^{pp}) \text{RC21} \quad (\text{A.49})$$

$$R_{13} = \frac{a_{25} (\text{RC12}^p + \text{RC12}^{pp} + \text{RC21}^{pp})}{a_{25M} + \text{RC12}^p + \text{RC12}^{pp} + \text{RC21}^{pp}} (\xi - \gamma) \quad (\text{A.50})$$

$$R_{14} = \alpha_{26} \gamma \quad (\text{A.51})$$

$$R_{15} = \alpha_{23} \text{RC21}^{pp} \quad (\text{A.52})$$

$$R_{16} = \alpha_{24} \text{RC21}^{pp} \quad (\text{A.53})$$

$$R_{17} = \alpha_{25} p53^p. \quad (\text{A.54})$$

Here,  $u(t)$  represents the stimulus in form of a switching function, i.e. if the system is irradiated at dose rate  $\frac{dD}{dt}$ ,  $u(t) = 1$ . If the system is not irradiated,  $u(t) = 0$ .

### A.3.2 Parameter Inference

The parameters are estimated based on the maximum likelihood principle. Owing data processing, log-transform, noise model and ANOVA analysis, standard conditions can be assumed to hold. In fact, this assumption was verified after obtaining a fit by using Anderson-Darling statistics (s. Sec. 3.5, Tab. 3.3). By this we also tested model adequacy.

The objective function Eq. (2.13) itself was minimized using a hybrid optimization strategy, combining a genetic algorithm and interior-point/active-set optimization, which are implemented in MATLAB, to find a nearly global optimum. The models were

## A. SUPPLEMENTARY METHODOLOGICAL INFORMATION

---

also implemented in MATLAB and solved using the CVODES solver from (?). Rate constants and scaling parameters are positive and typically distributed on a logarithmic scale (??). Therefore, the parameter estimation was performed on a logarithmic scale. Further, possible realizations of the kinetic parameters were constrained to the interval  $[10^{-2} \dots 10^{+2}]$ , whereas upper bounds of scaling parameters have been adjusted up to  $10^4$ . Overall, 19 kinetic parameters and 8 scaling parameters per model were estimated. As already mention above, initial conditions of the proteins where assumed to be zero, reflecting zero activity of the unperturbed states. The inactive proteins Ku7080, MRN, DNA-PK<sub>cs</sub>, ATM and H2AX have large abundances, which allowed to reduce the number of parameters by assuming a constant supply of inactive to active protein forms. In the case of  $\gamma$ H2AX, the conservation relation

$$H2AX_{\text{tot}} = H2AX + \gamma H2AX \quad (\text{A.55})$$

has been used to simplify the back reaction. The final parameter for the final identified model A2 are given in Table A.1 in logarithmic representation. The lower and upper 95% point-wise confidence bounds are derived from the profile likelihood (see Sec. A.3.3). Bounds with  $\pm\infty$  indicate that the profile likelihood did not reach the critical value for significance. Notice that we have restricted the optimization effort for each model by constraining the parameter bounds on a range of 4 orders of magnitude in logarithmic space.

### A.3.3 Profile Likelihood Analysis

For model A2, we calculated the profile likelihood  $\chi_{PL}^2$  as for instance described in (?), which we have implemented in MATLAB in combination with the fast CVODES ODE integration package (?). Absolute and relative tolerances have been set to  $10^{-7}$  and  $10^{-6}$ , respectively. The MATLAB implementation of the profile likelihood algorithm has been parallelized and is based on a template from the first author of (?). In Figures A.1-A.19, we show the profile likelihoods for the kinetic parameters and the parameter dependencies in terms of relative parameter change for each kinetic parameter, when moving along the profile likelihood of each specific parameter in log-space. The relative parameter change of a parameter  $\theta_m$  for in- or decreasing parameter  $\theta_n$  from its maximum likelihood estimate and  $n \neq m$  is defined as

$$\delta\theta_{i,m} = \frac{\theta_{i,m} - \theta_m}{\theta_m}, \quad (\text{A.56})$$

with index  $i$  representing a position along the profile likelihood of  $\theta_n$  and  $\theta_m$  being the maximum likelihood estimate of model A2,  $m \in \{1, \dots, 19\} \setminus n$ .

As a rough interpretation guide, flat profile likelihoods indicate non-identifiable parameters, whereas profile likelihood that pass the critical  $\chi_{\alpha=0.05,df=1}^2$  value on both sides of the maximum-likelihood estimate of each parameter indicate an identifiable parameter. Profile likelihoods that hit the critical  $\chi_{\alpha=0.05,df=1}^2$  value (in the Figures indicated



### **A.3 Supplementary *in vitro* application information**

---

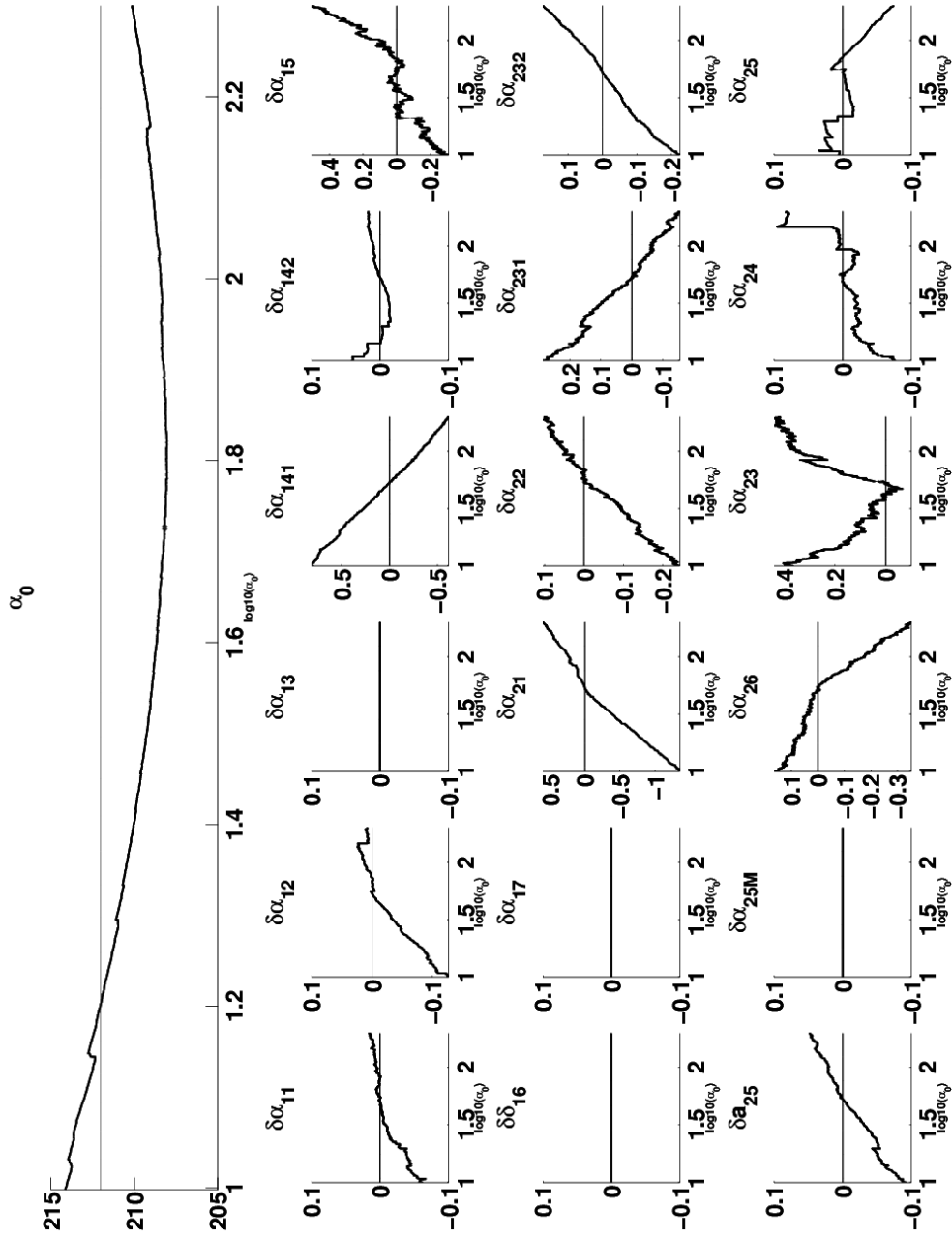
by the red line) only on one side indicate practically non-identifiable parameters. In this case, at least the lower or upper bound of the parameter are bounded.

## A. SUPPLEMENTARY METHODOLOGICAL INFORMATION

---

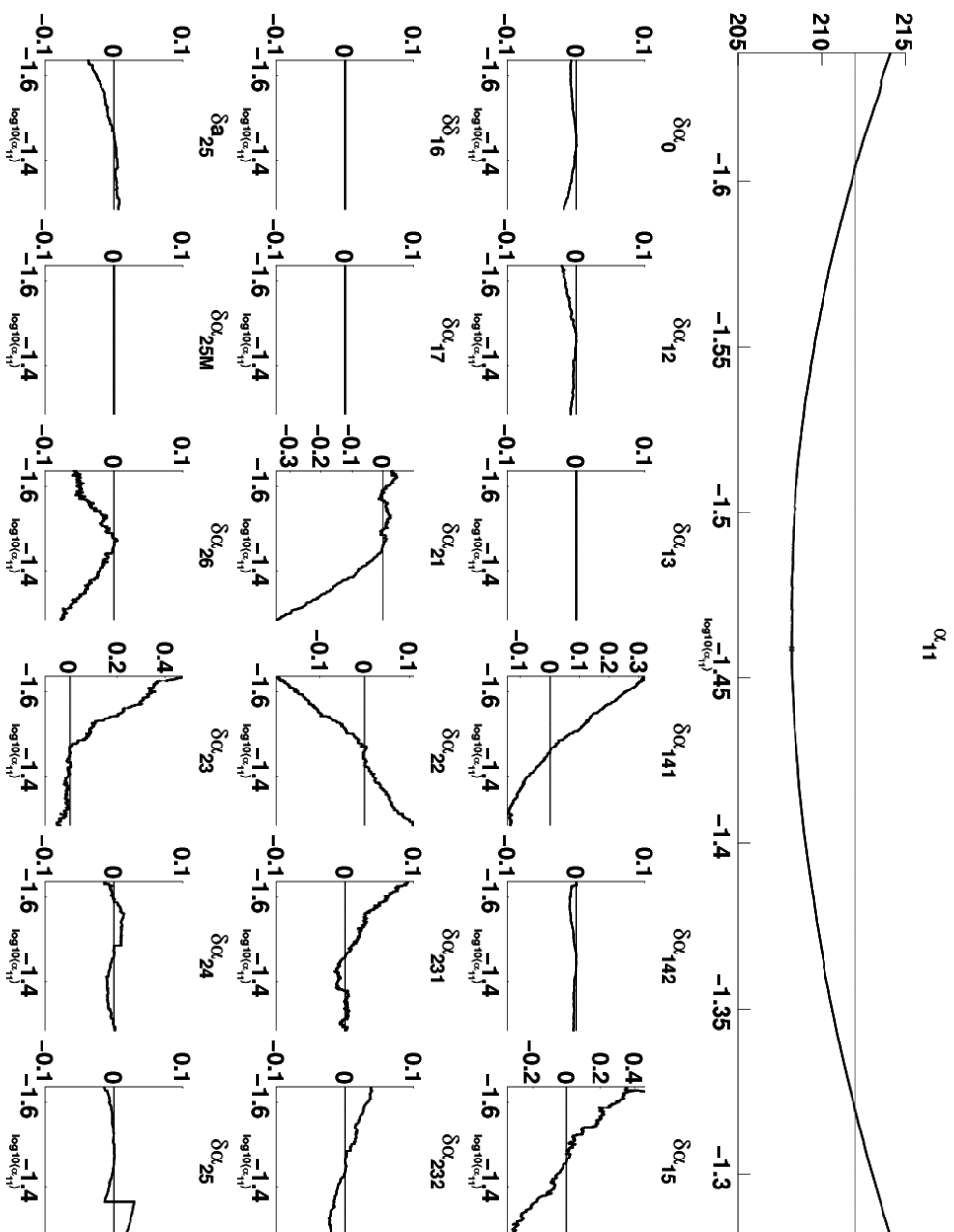
**Table A.1:** Final parameter set for model A2 and profile likelihood base lower and upper (LB,UB) 95% point-wise confidence bounds in log-space. Scaling parameters are represented as  $\xi = \frac{[H2AX_{tot}]}{[Ku7080_{tot}]}$  and  $s_i$  and are in principle non-identifiable owing relative measurement data.

Parameter	Units	LB	$\log_{10}(\theta)$	UB
$\alpha_0 = \frac{a_0}{[Ku7080_{tot}]}$	Gy <sup>-1</sup>	1.2024	1.7262	$\infty$
$\alpha_{11} = a_{11}[Ku7080_{tot}]$	min <sup>-1</sup>	-1.6041	-1.4588	-1.3195
$\alpha_{12} = a_{12}[Ku7080_{tot}]$	min <sup>-1</sup>	-0.2517	1.5123	$\infty$
$\alpha_{13} = a_{13}[Ku7080_{tot}]$	min <sup>-1</sup>	1.8086	2.0000	$\infty$
$\alpha_{141} = a_{141}[Ku7080_{tot}]$	min <sup>-1</sup>	-0.9869	-0.5246	-0.2279
$\alpha_{142} = \frac{a_{142}}{a_{141}}[Ku7080_{tot}]$	1	1.2977	1.7342	$\infty$
$\alpha_{15} = a_{15}[Ku7080_{tot}]$	min <sup>-1</sup>	-0.7768	-0.2492	0.1913
$\delta_{16} = d_{16}[Ku7080_{tot}]$	1	1.4718	1.9601	$\infty$
$\alpha_{17} = a_{17}[Ku7080_{tot}]$	min <sup>-1</sup>	$-\infty$	0.5089	$\infty$
$\alpha_{21} = a_{21}[Ku7080_{tot}]$	min <sup>-1</sup>	-0.7612	-0.4635	0.2067
$\alpha_{22} = a_{22}[Ku7080_{tot}]$	min <sup>-1</sup>	-0.9253	-0.6773	-0.3882
$\alpha_{231} = a_{231}[Ku7080_{tot}]$	min <sup>-1</sup>	-1.7834	-0.3972	0.2888
$\alpha_{232} = a_{232}[Ku7080_{tot}]$	min <sup>-1</sup>	0.7257	1.2354	1.5524
$a_{25}$	min <sup>-1</sup>	0.2562	1.355	$\infty$
$\alpha_{25M} = a_{25M}[Ku7080_{tot}]$	M <sup>2</sup>	$-\infty$	-2	-1.8033
$\alpha_{26} = a_{26}[Ku7080_{tot}]$	min <sup>-1</sup>	-0.1947	0.6083	1.0618
$\alpha_{23} = a_{23}[Ku7080_{tot}]$	min <sup>-1</sup>	-0.0538	0.2526	1.1834
$\alpha_{24} = a_{24}[Ku7080_{tot}]$	min <sup>-1</sup>	-1.6565	-1.2240	-0.8093
$\alpha_{25} = a_{25}[Ku7080_{tot}]$	min <sup>-1</sup>	$-\infty$	-1.7197	-0.8152
$\xi = \frac{[H2AX_{tot}]}{[Ku7080_{tot}]}$	1	-	-0.6832	-
$s_0$	1	-	2.7705	-
$s_1$	1	-	2.4559	-
$s_2$	1	-	2.6787	-
$s_3$	1	-	2.7727	-
$s_4$	1	-	3.0366	-
$s_5$	1	-	1.9483	-
$s_6$	1	-	-1.0169	-

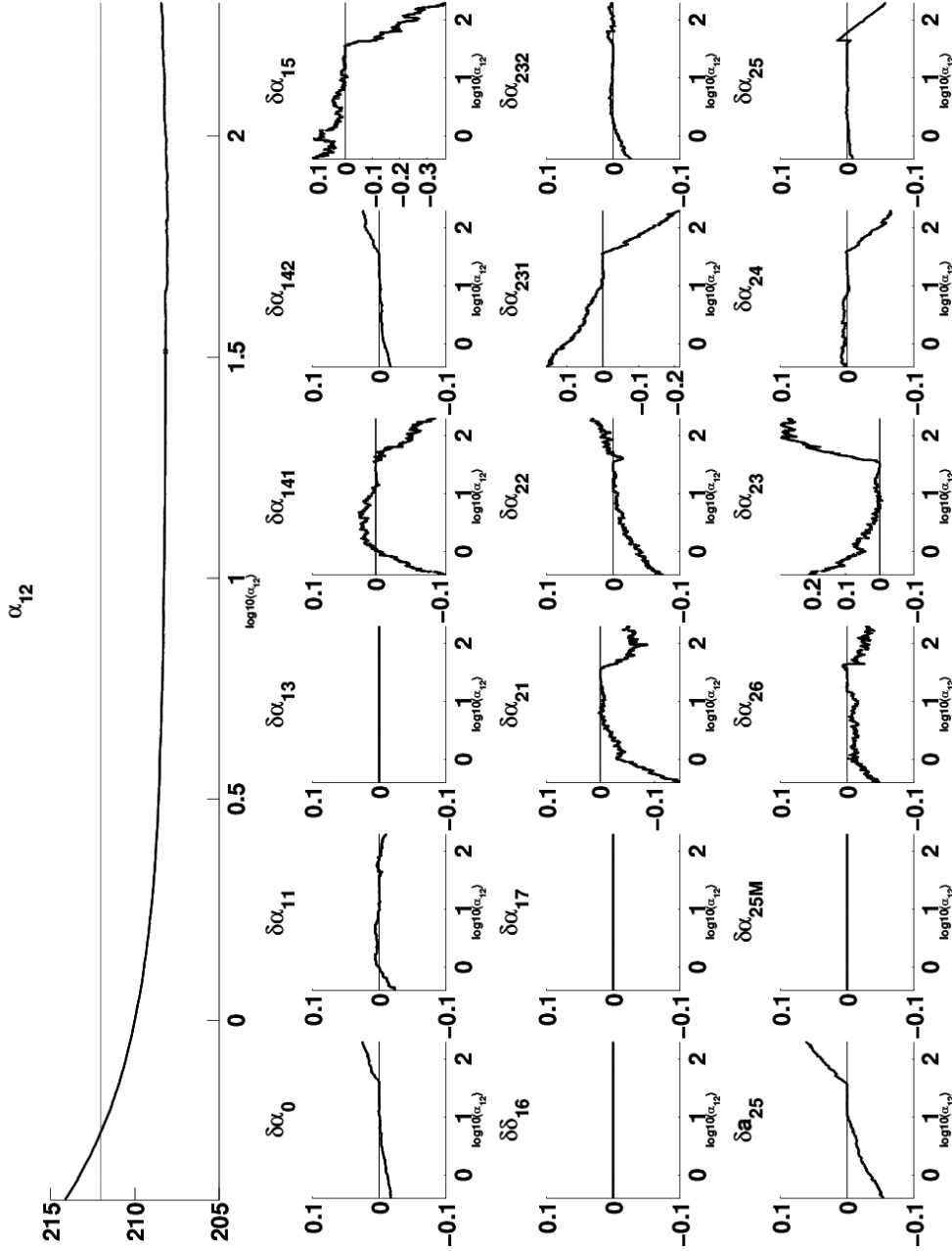


**Figure A.1:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2\log(l_{PL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.

## A. SUPPLEMENTARY METHODOLOGICAL INFORMATION

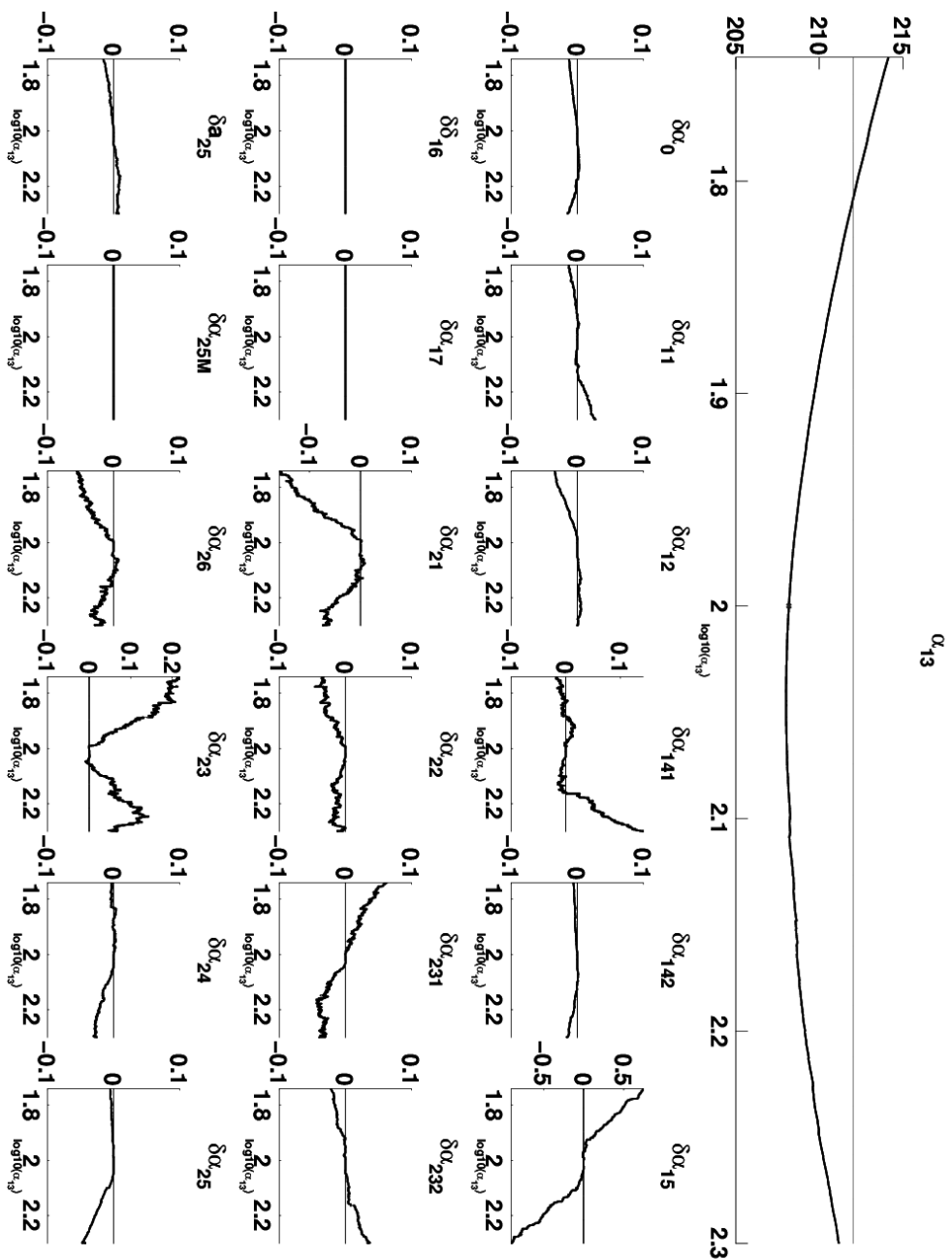


**Figure A.2:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2\log(\text{PL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.

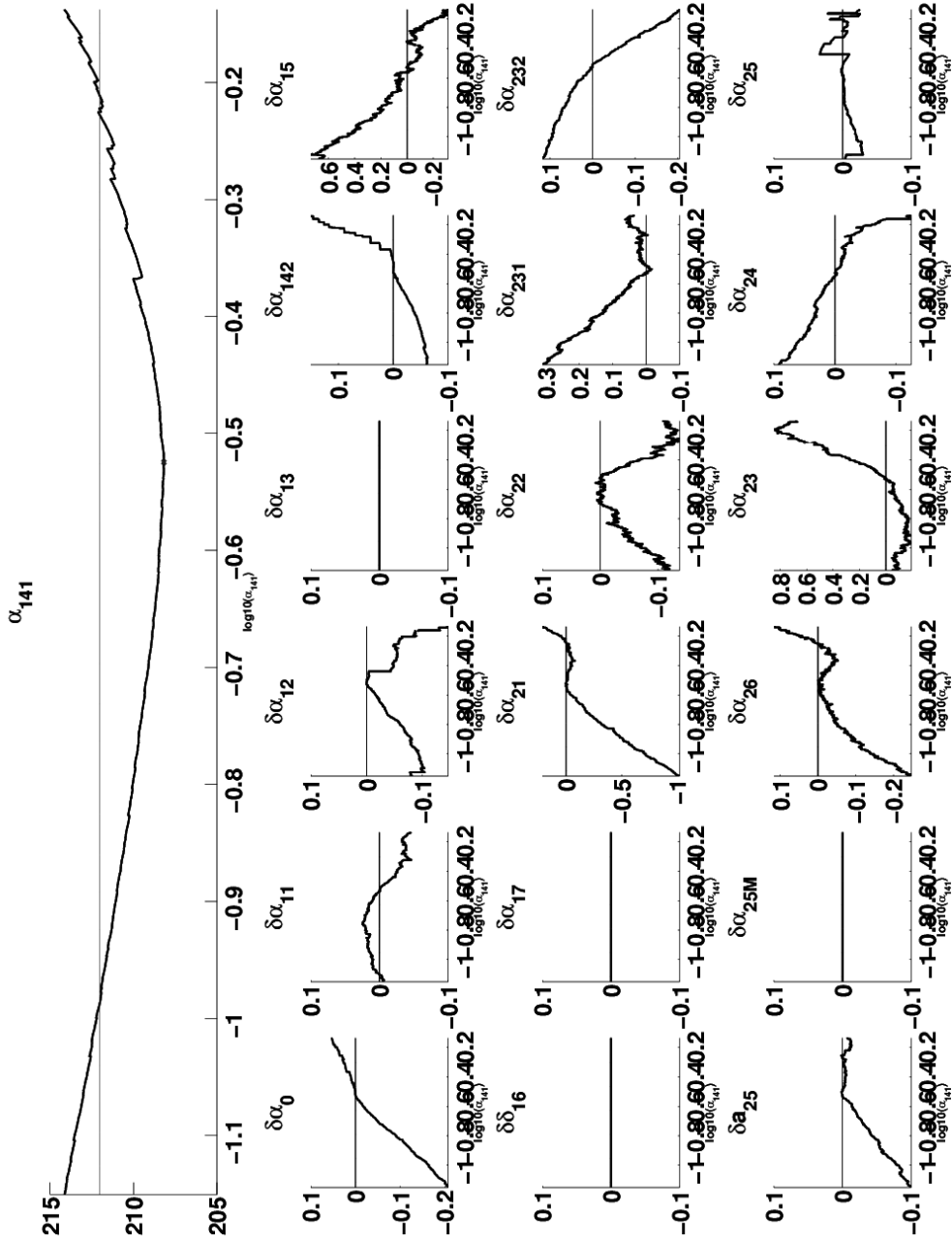


**Figure A.3:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(l_{PL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.

## A. SUPPLEMENTARY METHODOLOGICAL INFORMATION

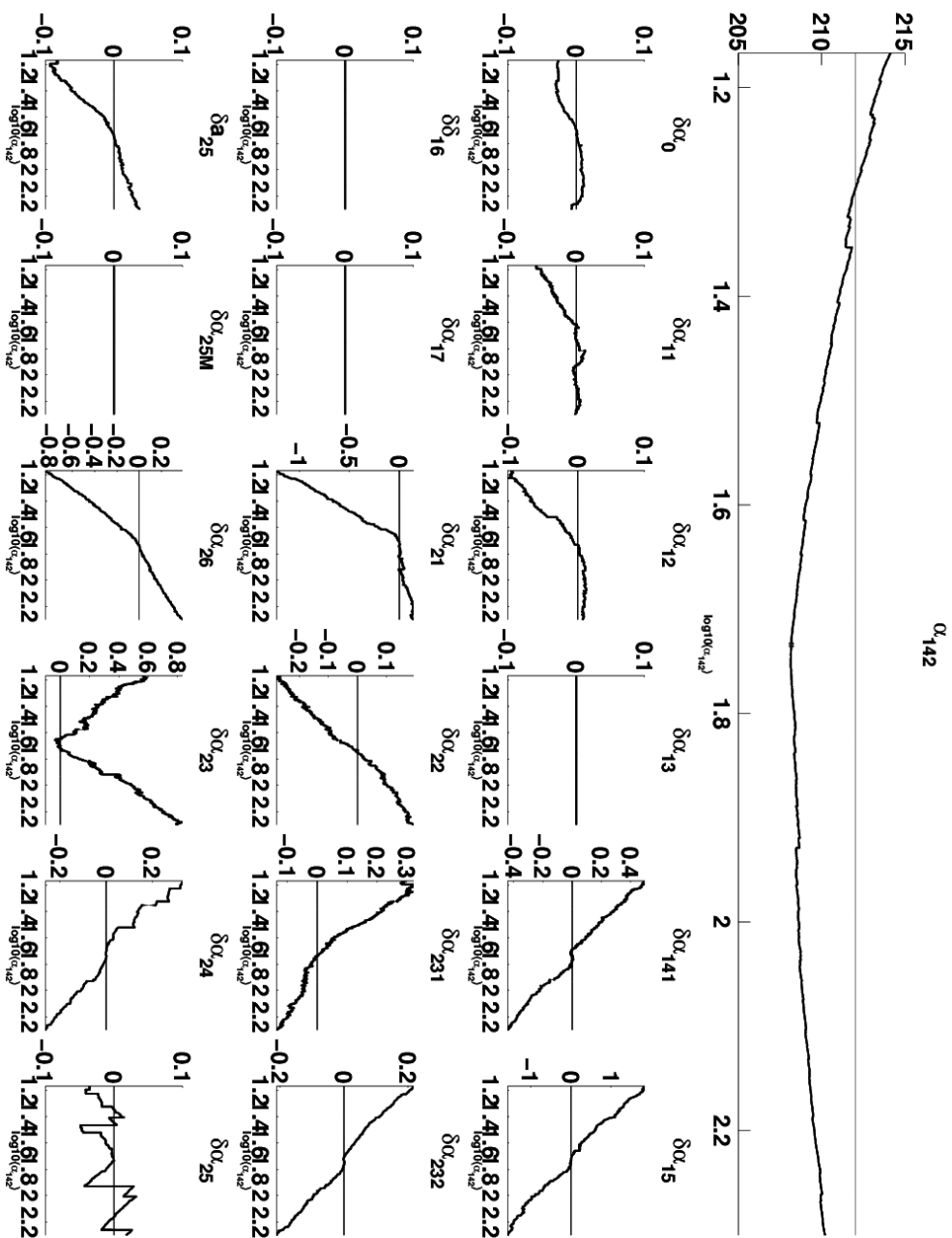


**Figure A.4:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(\text{PL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.



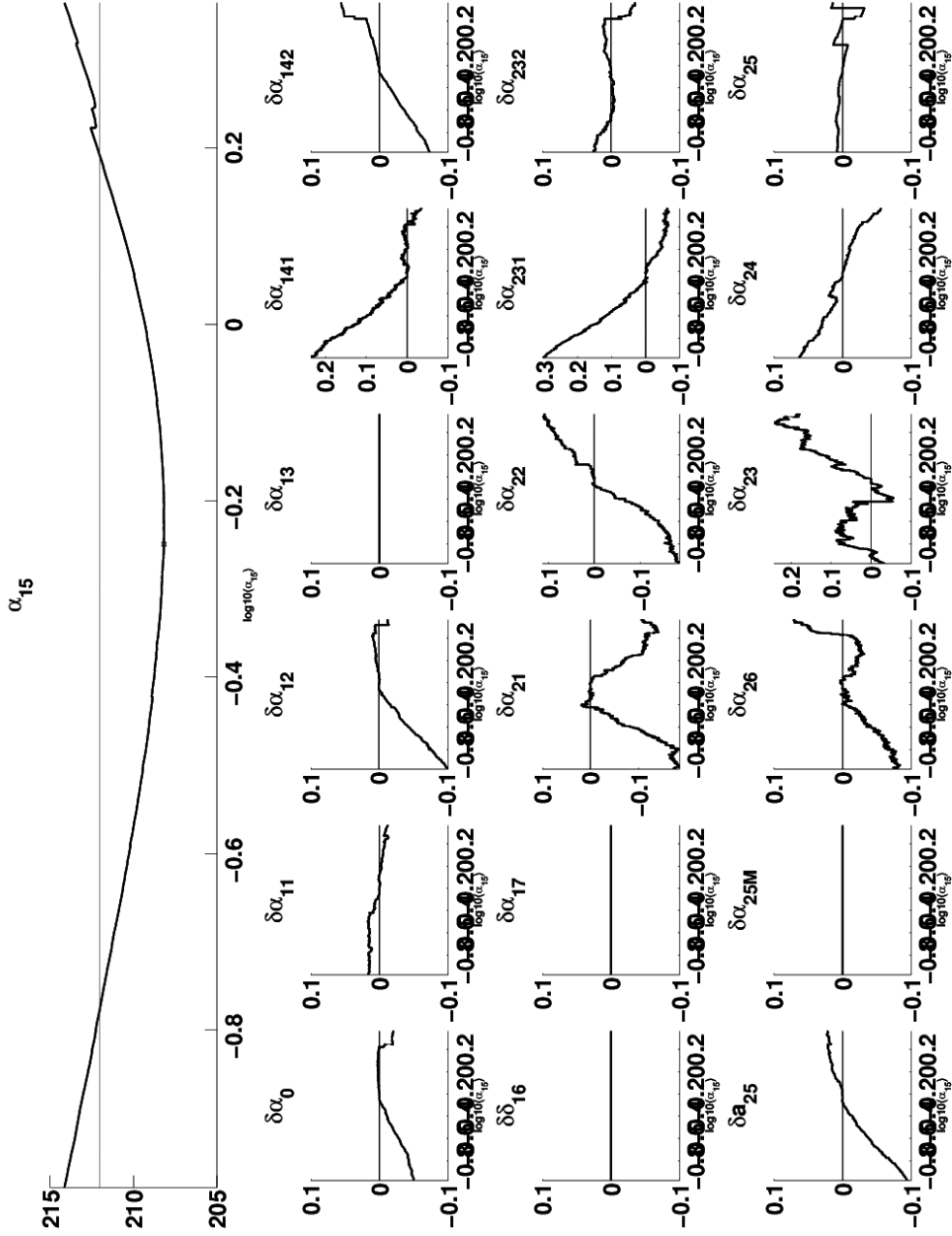
**Figure A.5:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(l_{PL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.

## A. SUPPLEMENTARY METHODOLOGICAL INFORMATION



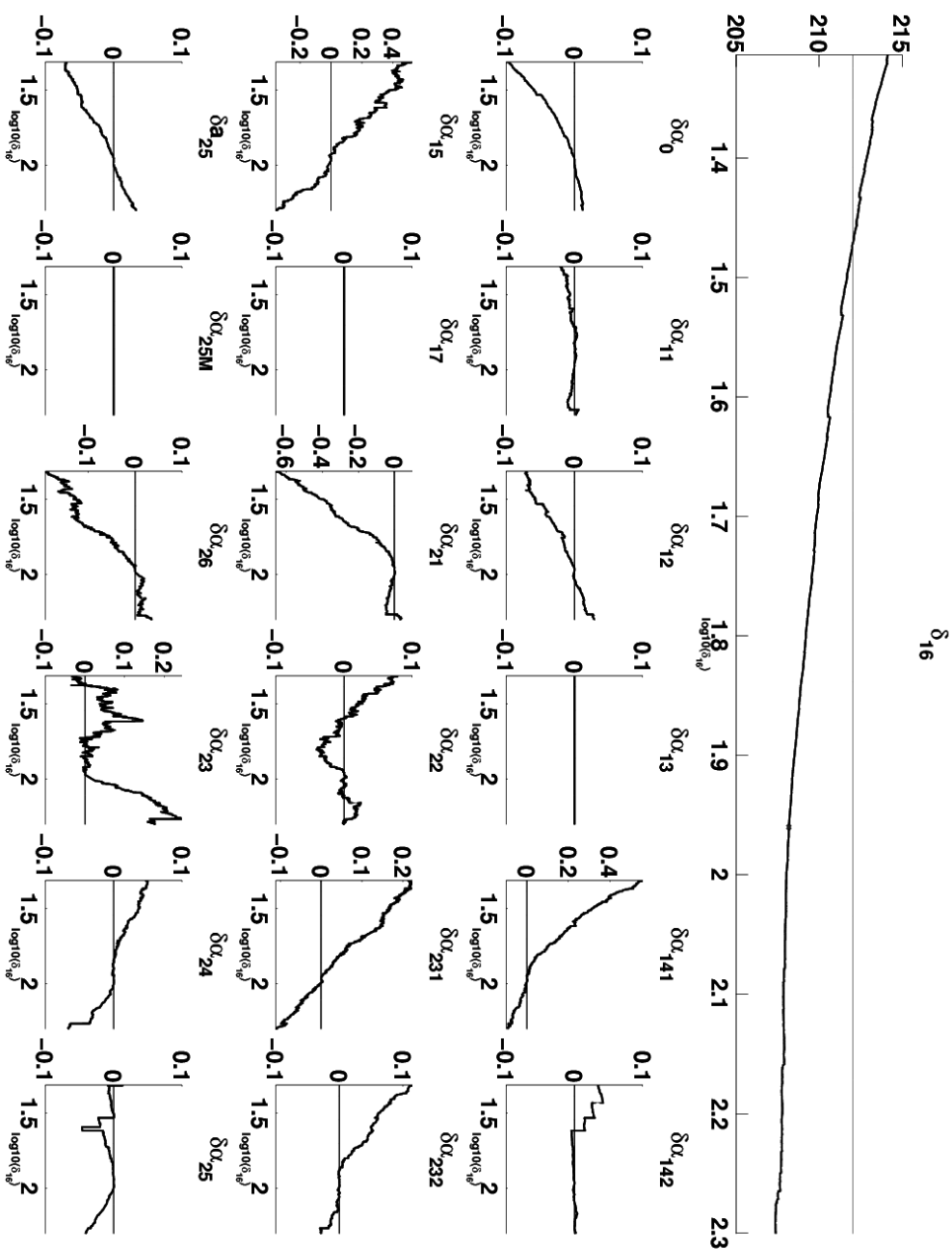
**Figure A.6:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(LPL)$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.



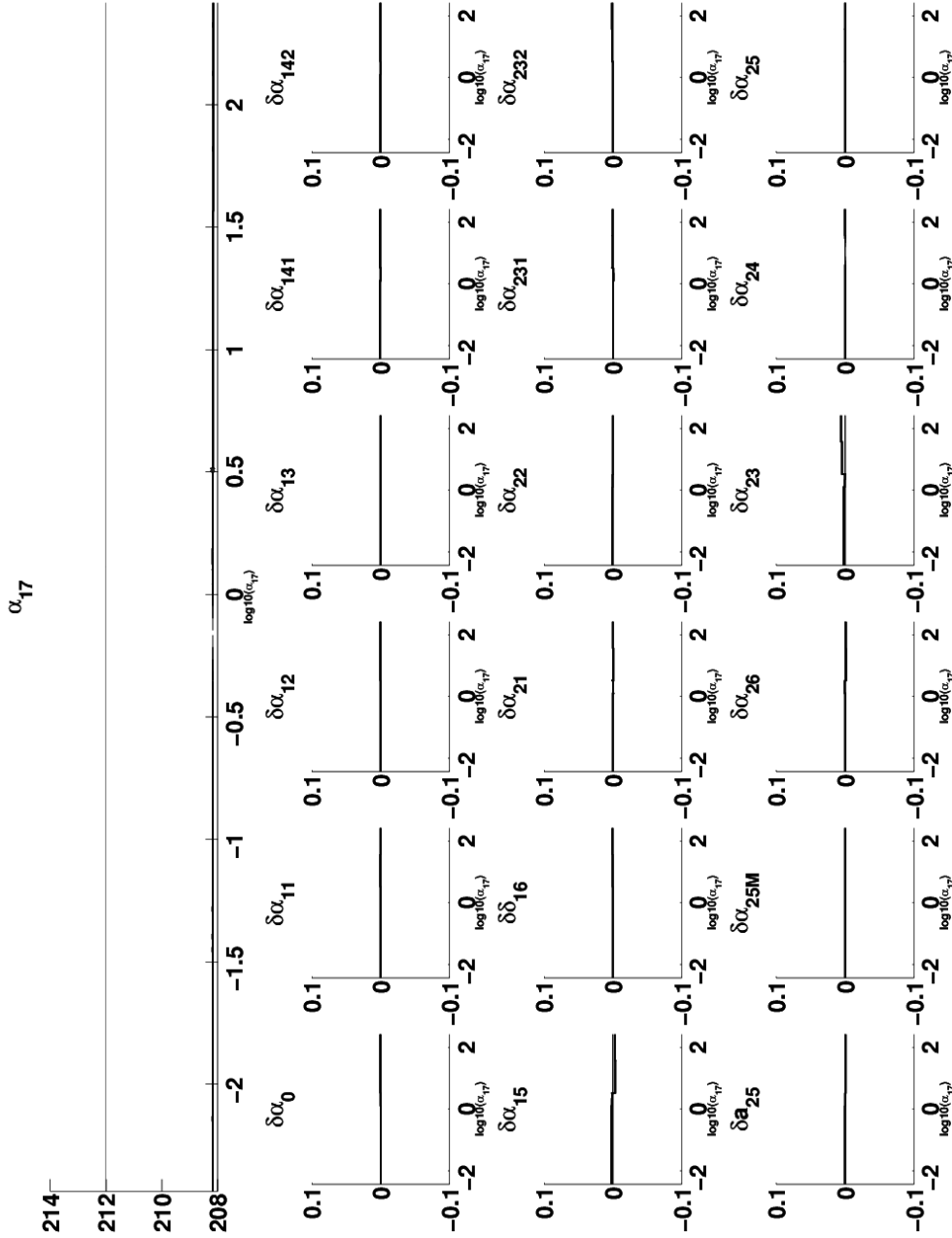


**Figure A.7:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(l_{PL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.

## A. SUPPLEMENTARY METHODOLOGICAL INFORMATION

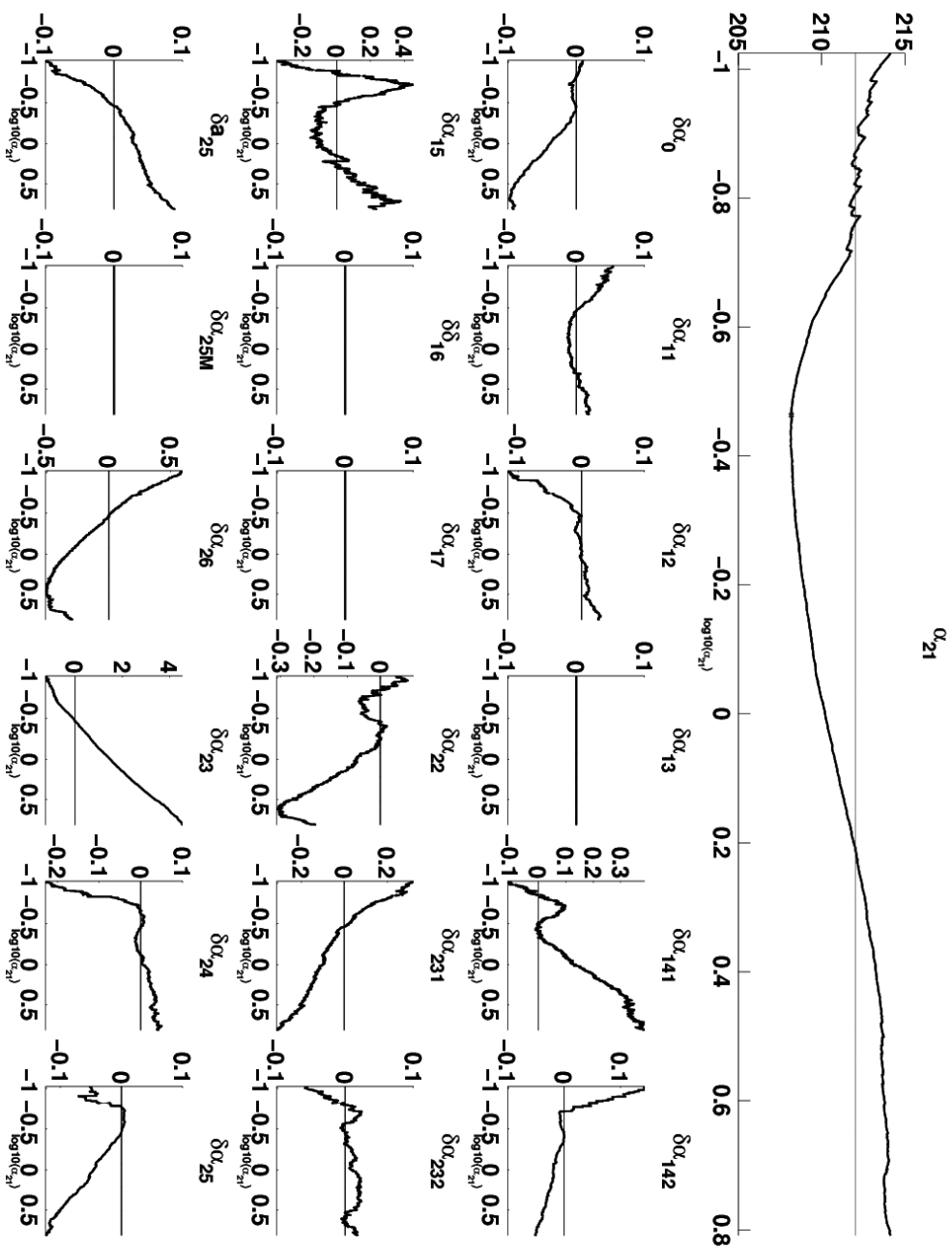


**Figure A.8:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(\text{pL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.

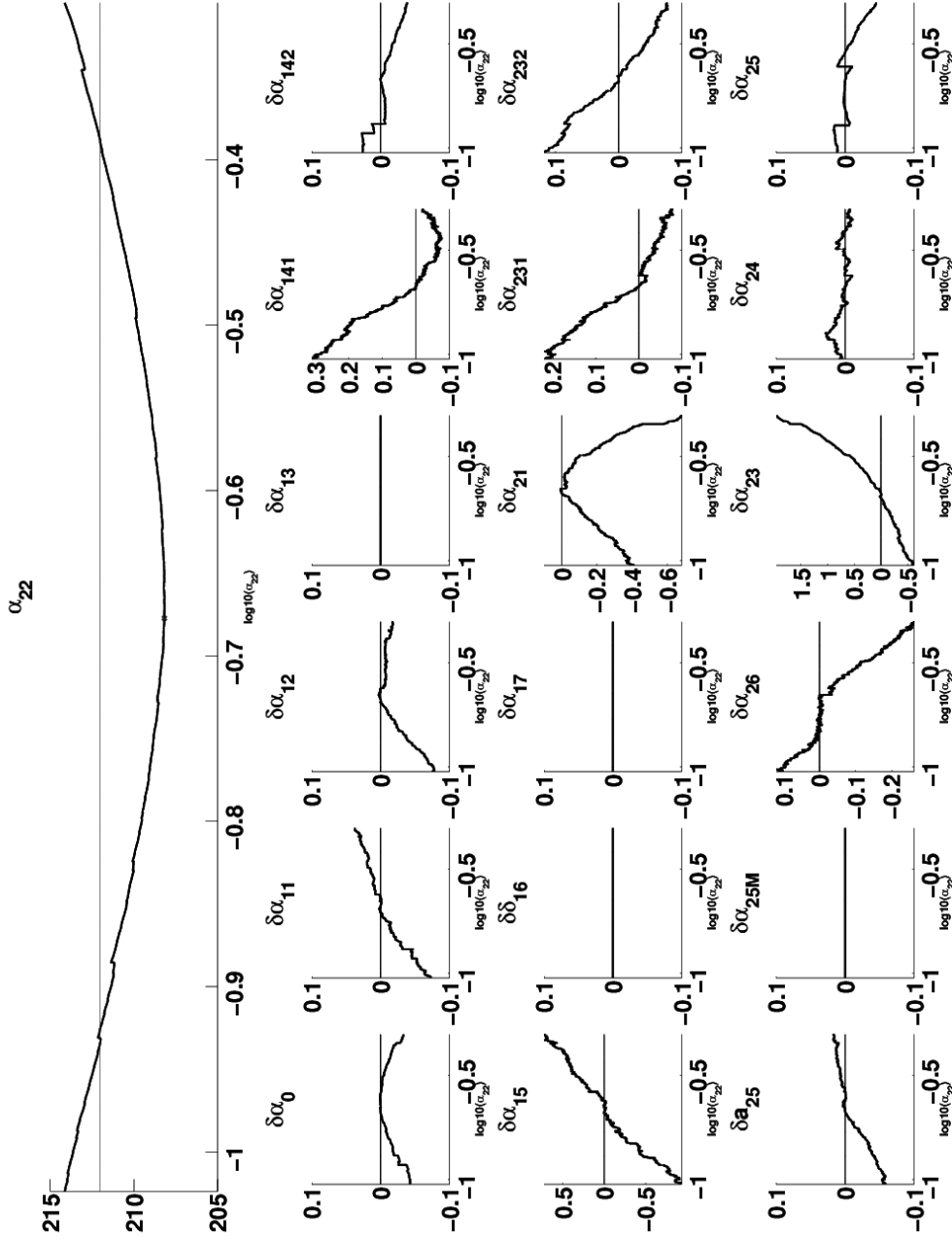


**Figure A.9:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2\log(l_{PL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.

## A. SUPPLEMENTARY METHODOLOGICAL INFORMATION

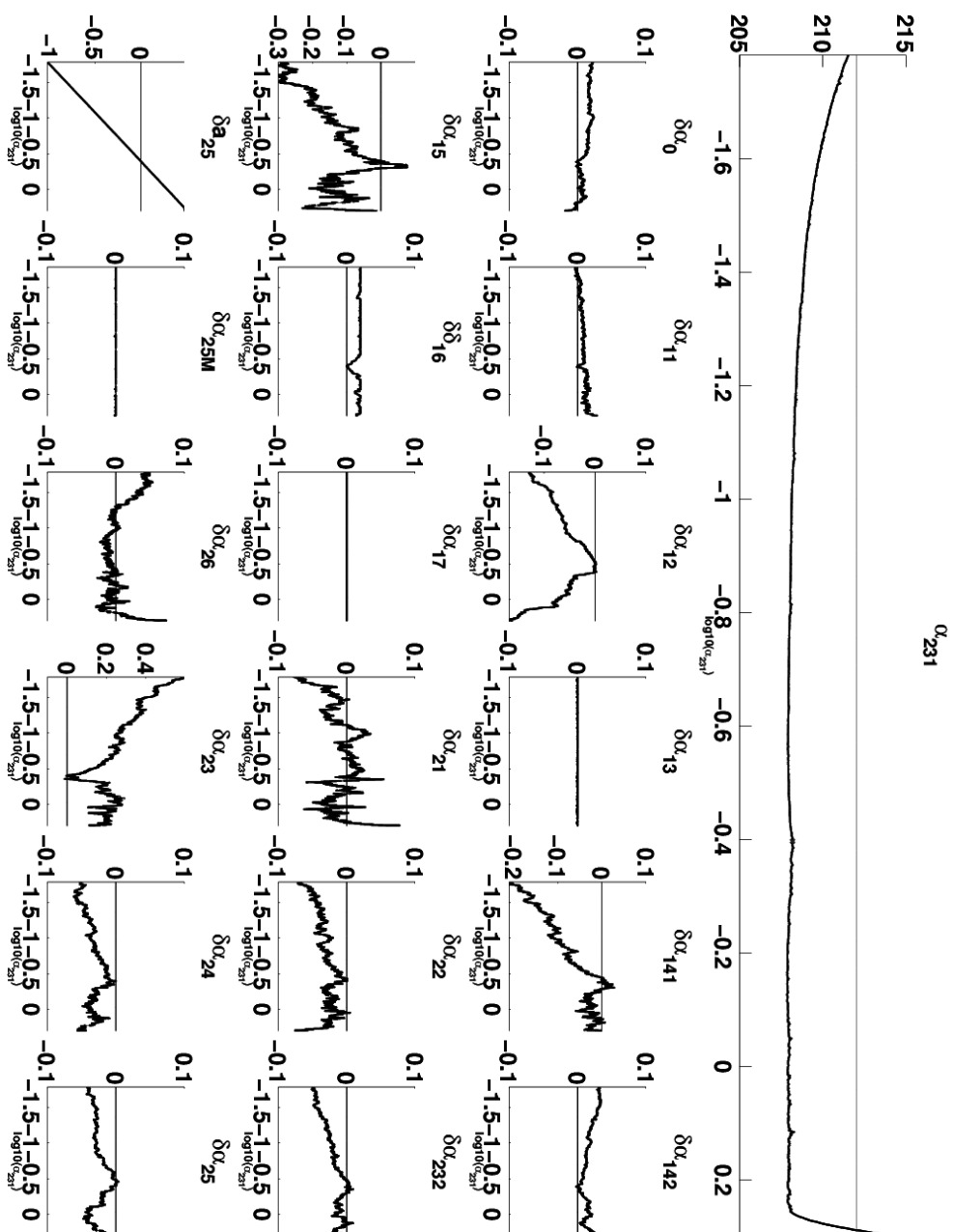


**Figure A.10:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(\text{pL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.

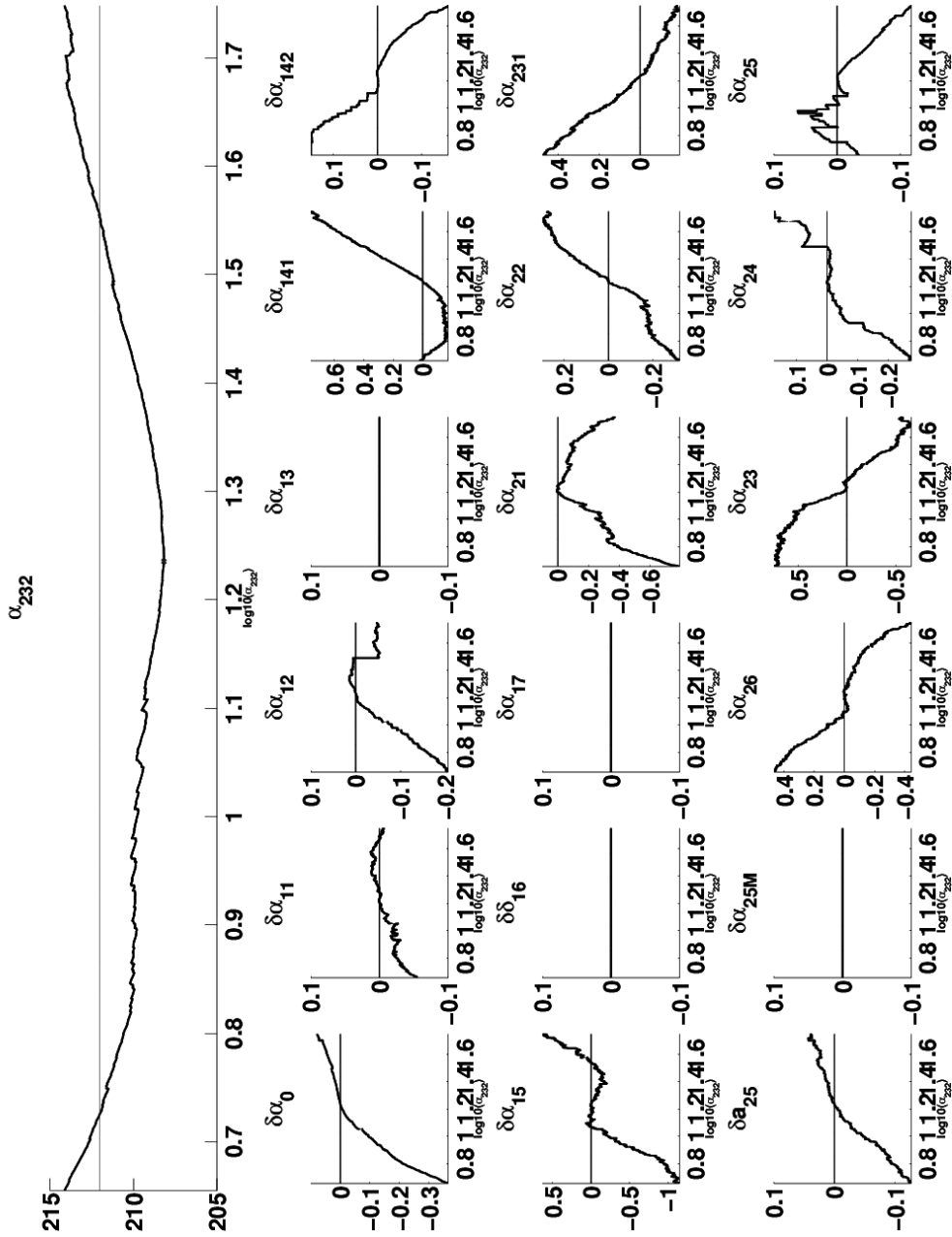


**Figure A.11:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(l_{PL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.

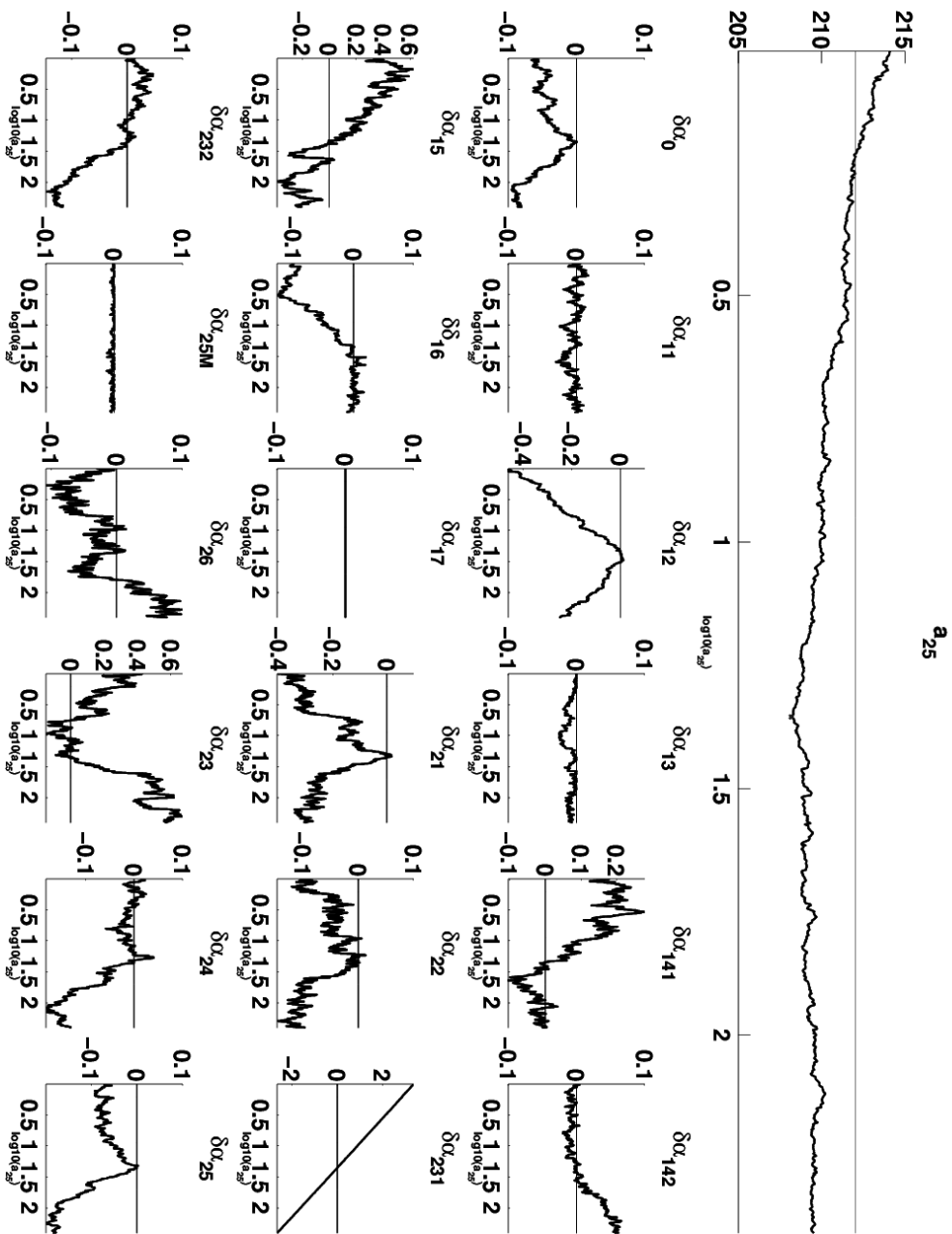
## A. SUPPLEMENTARY METHODOLOGICAL INFORMATION



**Figure A.12:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(\text{pL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.



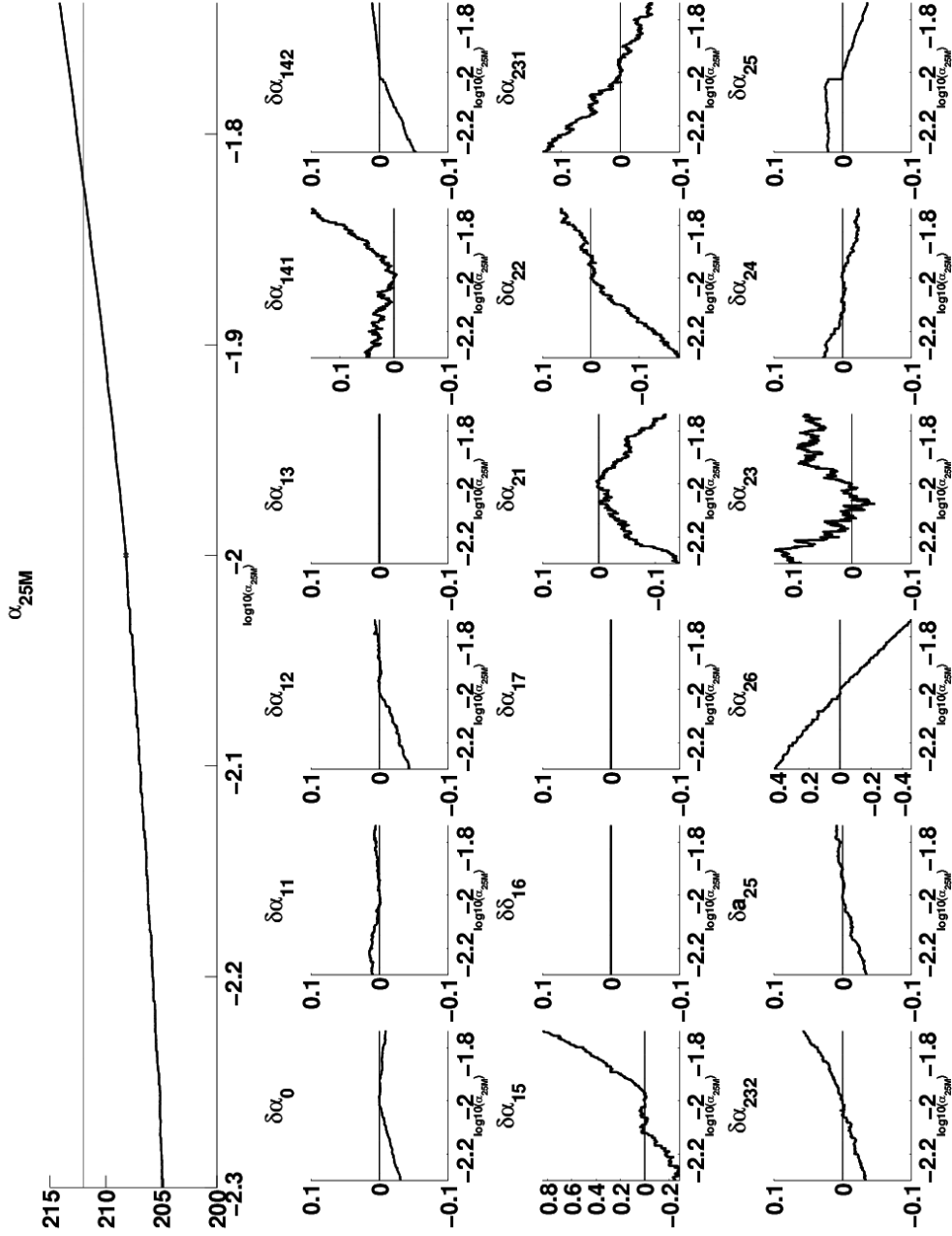
**Figure A.13:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(I_{PL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.



a<sub>25</sub>

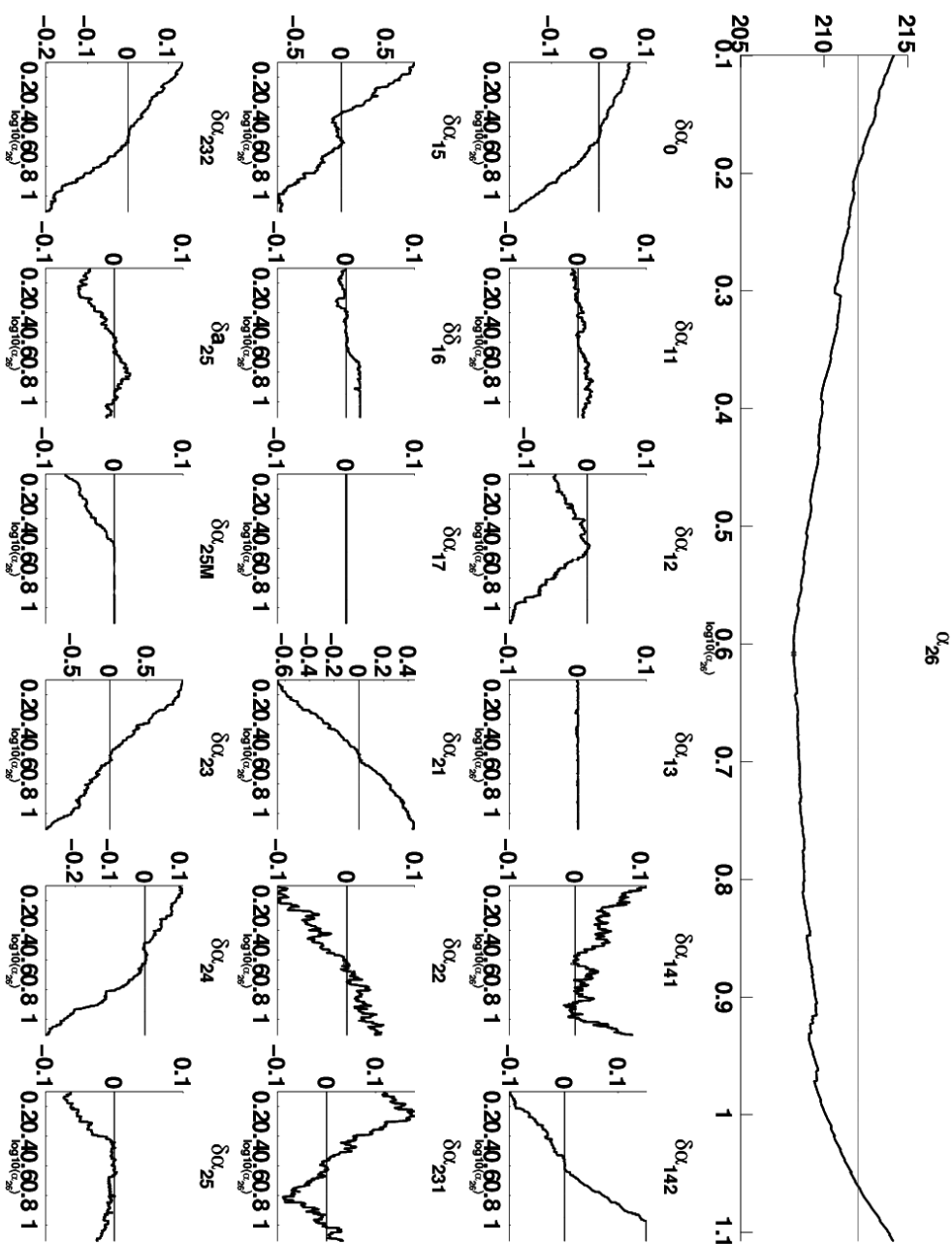
**Figure A.14:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(\text{pL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.



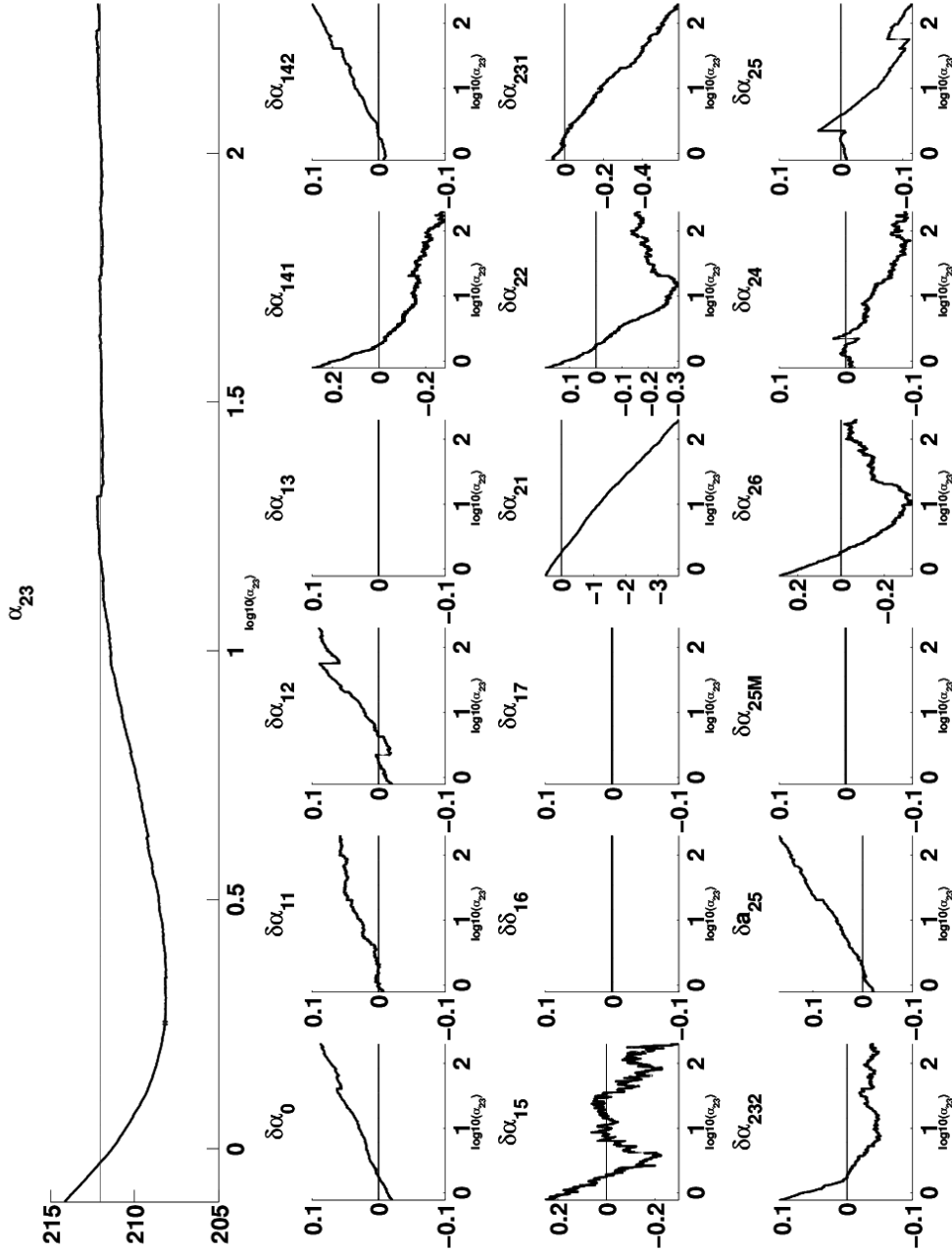


**Figure A.15:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(l_{PL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.

## A. SUPPLEMENTARY METHODOLOGICAL INFORMATION

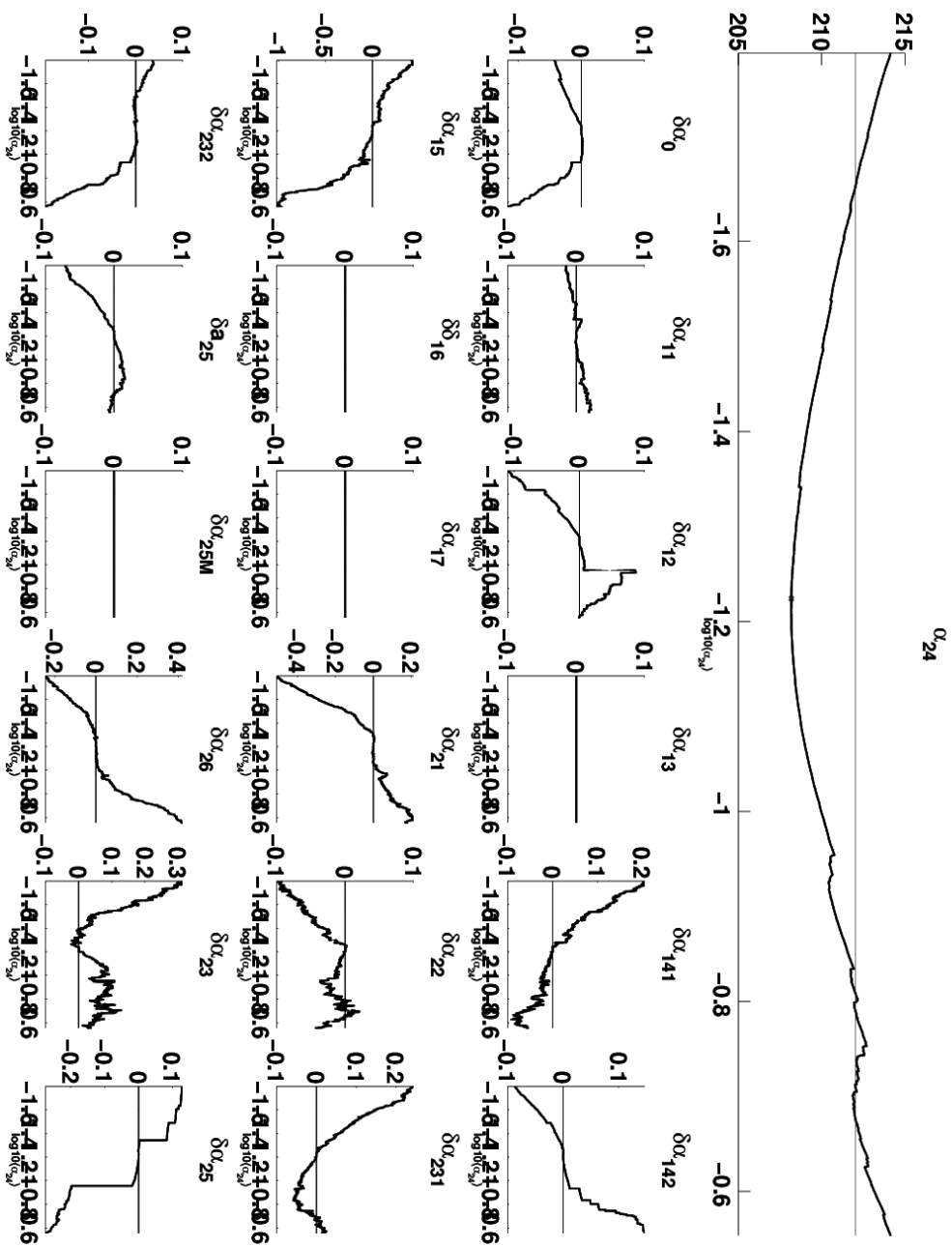


**Figure A.16:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(\text{pL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.

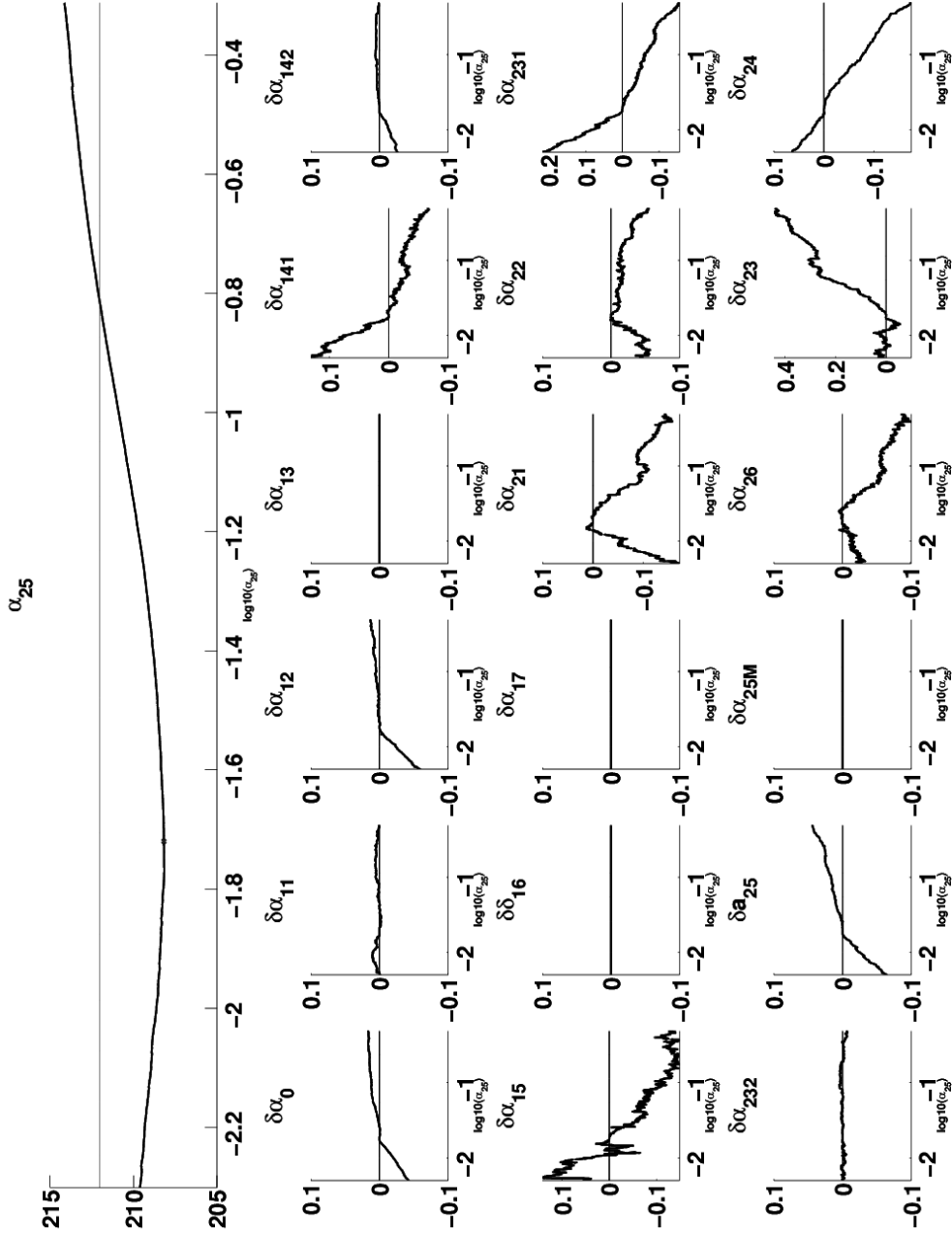


**Figure A.17:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(I_{PL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.

## A. SUPPLEMENTARY METHODOLOGICAL INFORMATION



**Figure A.18:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(LPL)$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.



**Figure A.19:** Profile likelihood of model A2 for specified kinetic parameter (upper panel, black line) and its dependencies on the remaining kinetic parameters in terms of relative change of each kinetic parameter in logspace (vertical axes of the small 3 x 6 subplots). The vertical axis in the upper panel indicates  $-2 \log(l_{PL})$ , i.e. logged profile likelihood value, whereas the horizontal axis represents the parameter value in logspace. The thin black line in the upper panel indicates the critical value for significance. The parameter value of the point estimate is indicated by a small black cross in the upper panel.

## A. SUPPLEMENTARY METHODOLOGICAL INFORMATION

---

# References

- AHERNE, F.J., THACKER, N.A. & ROCKETT, P.I. (1998). The bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, **34**, 363–368.
- AHO, A.V., GAREY, M.R. & ULLMAN, J.D. (1972). The transitive reduction of a directed graph. *SIAM Journal on Computing*, **1**, 131–137.
- AITCHISON, J. & BROWN, J.A.C. (1969). *The lognormal distribution, with special reference to its uses in economics*, vol. 5. CUP Archive.
- AKUTSU, T., KUHARA, S., MARUYAMA, O. & MIYANO, S. (2003). Identification of genetic networks by strategic gene disruptions and gene overexpressions under a boolean model. *Theoretical Computer Science*, **298**, 235–251.
- ALBERT, R., DASGUPTA, B., DONDI, R., KACHALO, S., SONTAG, E., ZELIKOVSKY, A. & WESTBROOKS, K. (2007). A novel method for signal transduction network inference from indirect experimental evidence. *J Comput Biol*, **14**, 927–949.
- ANDRAE, R., SCHULZE-HARTUNG, T. & MELCHIOR, P. (2010). Dos and Don'ts of reduced Chi-squared. *arXiv preprint arXiv:1012.3754*.
- ANGUELOVA, M. (2007). *Observability and identifiability of nonlinear systems with applications in biology*. Ph.D. thesis, Chalmers University of Technology.
- APGAR, J.F., TOETTCHE, J.E., ENDY, D., WHITE, F.M. & TIDOR, B. (2008). Stimulus design for model selection and validation in cell signaling. *PLoS computational biology*, **4**, e30.
- ATKINSON, A.C. & FEDOROV, V.V. (1975). The design of experiments for discriminating between two rival models. *Biometrika*, **62**, 57–70.
- AU, C. & TAM, J. (1999). Transforming variables using the dirac generalized function. *The American Statistician*, **53**, 270–272.
- BACHMANN, J., RAUE, A., SCHILLING, M., BÖHM, M.E., KREUTZ, C., KASCHEK, D., BUSCH, H., GRETZ, N., LEHMANN, W.D., TIMMER, J. & KLINGMÜLLER, U. (2011). Division of labor by dual feedback regulators controls jak2/stat5 signaling over broad ligand range. *Molecular Systems Biology*, **7**.
- BARTLETT, M.S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **160**, 268–282.

## REFERENCES

---

- BEHAR, M., HAO, N., DOHLMAN, H.G. & ELSTON, T.C. (2007). Mathematical and computational analysis of adaptation via feedback inhibition in signal transduction pathways. *Biophysical journal*, **93**, 806–821.
- BELLMAN, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ.
- BELLMAN, R. & ASTROEM, K.J. (1970). On structural identifiability. *Mathematical Biosciences*, **7**, 329–339.
- BELLU, G., SACCOMANI, M.P., AUDOLY, S. & D’ANGIO, L. (2007). Daisy: a new software tool to test global identifiability of biological and physiological systems. *Comput Methods Programs Biomed*, **88**, 52–61.
- BENNER, S.A. & SISMOUR, A.M. (2005). Synthetic biology. *Nature Reviews Genetics*, **6**, 533–543.
- BERMAN, P., DASGUPTA, B. & KARPINSKI, M. (2009). Approximating transitive reductions for directed networks. In F. Dehne, M. Gavrilova, J.R. Sack & C. Toth, eds., *Algorithms and Data Structures*, vol. 5664 of *Lecture Notes in Computer Science*, 74–85, Springer Berlin Heidelberg.
- BIEGLER, L.T. (2007). An overview of simultaneous strategies for dynamic optimization. *Chemical Engineering and Processing: Process Intensification*, **46**, 1043–1053.
- BING, N. & HOESCHELE, I. (2005). Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics*, **170**, 533–542.
- BOLDT, J. & MÜLLER, O. (2008). Newtons of the leaves of grass. *Nature biotechnology*, **26**.
- BOX, G.E. & COX, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26**, 211–252.
- BOX, G.E. & DRAPER, N.R. (1987). *Empirical model-building and response surfaces*. John Wiley & Sons.
- BOX, G.E. & HILL, W. (1967). Discrimination among mechanistic models. *Technometrics*, **9**, 57–71.
- BOX, G.E. & LUCAS, H. (1959). Design of experiments in non-linear situations. *Biometrika*, 77–90.
- BOX, G.E., HUNTER, J.S. & HUNTER, W.G. (2005). *Statistics for experimenters: design, innovation, and discovery*, vol. 2. Wiley Online Library.
- BRAZHNİK, P., DE LA FUENTE, A. & MENDES, P. (2002). Gene networks: how to put the function in genomics. *Trends in Biotechnology*, **20**, 467–472.
- BREIMAN, L. (2001). Random forests. *Machine learning*, **45**, 5–32.
- BREM, R.B. & KRUGLYAK, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 1572–1577.



## REFERENCES

---

- BURMA, S., CHEN, B.P., MURPHY, M., KURIMASA, A. & CHEN, D.J. (2001). ATM phosphorylates histone H2AX in response to DNA double-strand breaks. *Journal of Biological Chemistry*, **276**, 42462–42467.
- BURNHAM, K.P. & ANDERSON, D.R. (2002). Information and likelihood theory: A basis for model selection and inference. In *Model Selection and Multimodel Inference*, 49–97, Springer New York.
- BUZZI-FERRARIS, G. & FORZATTI, P. (1983). A new sequential experimental design procedure for discriminating among rival models. *Chemical Engineering Science*, **38**, 225–232.
- BUZZI FERRARIS, G., FORZATTI, P., EMIG, G. & HOFMANN, H. (1984). Sequential experimental design for model discrimination in the case of multiple responses. *Chemical Engineering Science*, **39**, 81–85.
- CANMAN, C.E., LIM, D.S., CIMPRICH, K.A., TAYA, Y., TAMAI, K., SAKAGUCHI, K., APPELLA, E., KASTAN, M.B. & SILICIANO, J.D. (1998). Activation of the ATM kinase by ionizing radiation and phosphorylation of p53. *Science*, **281**, 1677–9.
- CARLBORG, O., DE KONING, D., MANLY, K.F., CHESLER, E., WILLIAMS, R.W. & HALEY, C.S. (2005). Methodological aspects of the genetic dissection of gene expression. *Bioinformatics*, **21**, 2383–2393.
- CHACHUAT, B. (2007). Nonlinear and dynamic optimization. *From Theory to Practice, Laboratoire dmAutomatique, Ecole Polytechnique Federale de Lausanne*.
- CHAN, D.W., CHEN, B.P.C., PRITHIVIRAJESINGH, S., KURIMASA, A., STORY, M.D., QIN, J. & CHEN, D.J. (2002). Autophosphorylation of the DNA-dependent protein kinase catalytic subunit is required for rejoining of DNA double-strand breaks. *Genes & Development*, **16**, 2333–2338.
- CHEN, B.H. & ASPREY, S.P. (2003). On the design of optimally informative dynamic experiments for model discrimination in multiresponse nonlinear situations. *Industrial & Engineering Chemistry Research*, **42**, 1379–1390.
- CHEN, T., HE, H.L. & CHURCH, G.M. (1999). Modeling gene expression with differential equations. In *Pacific symposium on biocomputing*, vol. 4, 4.
- CHURCHILL, G.A. & DOERGE, R.W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.
- CICCIA, A. & ELLEDGE, S.J. (2010). The DNA damage response: making it safe to play with knives. *Molecular Cell*, **40**, 179–204.
- COX, D.R. (1961). Tests of separate families of hypotheses.
- CUCINOTTA, F.A., PLUTH, J.M., ANDERSON, J.A., HARPER, J.V. & O’NEILL, P. (2008). Biochemical kinetics model of DSB repair and induction of  $\gamma$ -H2AX foci by non-homologous end joining. *Radiat. Res.*, **169**, 214–22.
- CUI, X., YU, Y., GUPTA, S., CHO, Y.M., LEES-MILLER, S.P. & MEEK, K. (2005). Autophosphorylation of DNA-dependent protein kinase regulates DNA end processing and may also alter double-strand break repair pathway choice. *Molecular and Cellular Biology*, **25**, 10842–10852.

## REFERENCES

---

- DAUB, C.O., STEUER, R., SELBIG, J. & KLOSKA, S. (2004). Estimating mutual information using B-spline functions-an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, **5**, 118.
- DAVIDSON, D., COULOMBE, Y., MARTINEZ-MARIGNAC, V., AMREIN, L., GRENIER, J., HODKINSON, K., MASSON, J.Y., ALOYZ, R. & PANASCI, L. (2012). Irinotecan and DNA-PKcs inhibitors synergize in killing of colon cancer cells. *Investigational New Drugs*, **30**, 1248–1256.
- DAVIDSON, D., AMREIN, L., PANASCI, L. & ALOYZ, R. (2013). Small molecules, inhibitors of DNA-PK, targeting DNA repair, and beyond. *Frontiers in Pharmacology*, **4**.
- DAVIS, J. & GOADRICH, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240, ACM.
- DAVISON, A.C. (1997). *Bootstrap methods and their application*, vol. 1. Cambridge university press.
- D’HAESELEER, P., WEN, X., FUHRMAN, S. & SOMOGYI, R. (1999). Linear modeling of mRNA expression levels during CNS development and injury. In *Pacific symposium on bio-computing*, vol. 4, 41–52.
- DICICCIO, T. & TIBSHIRANI, R. (1987). Bootstrap confidence intervals and bootstrap approximations. *Journal of the American Statistical Association*, **82**, 163–170.
- DIEDRICH, J. (2011). Optimale Versuchsplanung zur Modelldiskriminierung von dynamischen Modellen in der Systembiologie am Beispiel eines Signalmotivs.
- DISHISHA, T. (2013). *Microbial production of bio-based chemicals: a biorefinery perspective*. Ph.D. thesis, Lund University.
- DONALDSON, J.R. & SCHNABEL, R.B. (1987). Computational experience with confidence intervals for nonlinear least squares. *Technometrics*, **29**, 67–82.
- DONCKELS, B.M.R., DE PAUW, D.J.W., DE BAETS, B., MAERTENS, J. & VANROLLEGHEM, P.A. (2009). An anticipatory approach to optimal experimental design for model discrimination. *Chemometrics and Intelligent Laboratory Systems*, **95**, 53–63.
- DRUD, A. (1985). CONOPT: A GRG code for large sparse dynamic nonlinear optimization problems. *Mathematical Programming*, **31**, 153–191.
- DURZINSKY, M., WAGLER, A., WEISMANTEL, R. & MARWAN, W. (2008). Automatic reconstruction of molecular and genetic networks from discrete time series data. *BioSystems*, **93**, 181–190.
- EDERER, M. & GILLES, E.D. (2007). Thermodynamically feasible kinetic models of reaction networks. *Biophysical Journal*, **92**, 1846–1857.
- EFRON, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, **82**, 171–185.
- EFRON, B. & TIBSHIRANI, R.J. (1994). *An introduction to the bootstrap*, vol. 57. CRC press.
- EFRON, B., HASTIE, T., JOHNSTONE, I. & TIBSHIRANI, R. (2004). Least angle regression. *The Annals of statistics*, **32**, 407–499.

## REFERENCES

---

- EFROYMSON, M.A. (1960). Multiple regression analysis. In A. Ralston & H.S. Wilf, eds., *Mathematical Methods for Digital Computers*, 191–202, Wiley, New York.
- FERGUSON, D.O., SEKIGUCHI, J.M., CHANG, S., FRANK, K.M., GAO, Y., DEPINHO, R.A. & ALT, F.W. (2000). The nonhomologous end-joining pathway of dna repair is required for genomic stability and the suppression of translocations. *Proceedings of the National Academy of Sciences*, **97**, 6630–6633.
- FISHER, R.A. (1935). The design of experiments. *The design of experiments*.
- FLASSIG, R.J. & KLAMT, S. (2009). Dream 4 subchallenge 2; <http://wiki.c2b2.columbia.edu/dream/results/dream4/>.
- FLASSIG, R.J. & SUNDMACHER, K. (2012a). Nonlinear design of stimulus experiments for optimal discrimination of biochemical systems. In I.A. Karimi & R. Srinivasan, eds., *11th International Symposium on Process Systems Engineering-PSE2012*, vol. 31, 540–544, Elsevier.
- FLASSIG, R.J. & SUNDMACHER, K. (2012b). Optimal design of stimulus experiments for robust discrimination of biochemical reaction networks. *Bioinformatics*, **28**, 3089–3096.
- FLASSIG, R.J., HEISE, S., SUNDMACHER, K. & KLAMT, S. (2013). An effective framework for reconstructing gene regulatory networks from genetical genomics data. *Bioinformatics*, **29**, 246–254.
- FLASSIG, R.J., MAUBACH, G., TÄGER, C., SUNDMACHER, K. & NAUMANN, M. (2014). Experimental design, validation and computational modeling uncover DNA damage sensing by DNA-PK and ATM. *Molecular BioSystems*, **10**, 1978–1986.
- FLOUDAS, C.A. & GOUNARIS, C.E. (2009). A review of recent advances in global optimization. *Journal of Global Optimization*, **45**, 3–38.
- FRANCESCHINI, G. & MACCHIETTO, S. (2008). Model-based design of experiments for parameter precision: State of the art. *Chemical Engineering Science*, **63**, 4846–4872.
- FRIEDMAN, N., LINIAL, M., NACHMAN, I. & PEER, D. (2000). Using bayesian networks to analyze expression data. *J Comput Biol*, **7**, 601–20.
- GALVANIN, F., BAROLO, M. & BEZZO, F. (2009). Online model-based redesign of experiments for parameter estimation in dynamic systems. *Industrial & Engineering Chemistry Research*, **48**, 4415–4427.
- GARDNER, T.S. & FAITH, J.J. (2005). Reverse-engineering transcription control networks. *Physics of life reviews*, **2**, 65–88.
- GEURTS, P., IRRTHUM, A. & WEHENKEL, L. (2009). Supervised learning with decision tree-based methods in computational and systems biology. *Molecular BioSystems*, **5**, 1593–1605.
- GEYER, J., CHARLES (2010). Introduction to markov chain monte carlo. In B. Steve, G. Andrew, J. Galin & M. Xiao-Li, eds., *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/ CRC Press.
- GILKS, W.R., RICHARDSON, S. & SPIEGELHALTER, D.J. (1996). *Markov chain Monte Carlo in practice*, vol. 2. CRC Press.

## REFERENCES

---

- GILLES, E.D. (2002). Regelung-Schlüssel zum Verständnis biologischer Systeme. *Automatisierungstechnik*, **50**, 7–17.
- GRUCHATTKA, E., HÄDICKE, O., KLAMT, S., SCHÜTZ, V. & KAYSER, O. (2013). In silico profiling of escherichia coli and saccharomyces cerevisiae as terpenoid factories. *Microbial cell factories*, **12**, 84.
- GUSTAFSSON, F. & HENDEBY, G. (2008). On nonlinear transformations of stochastic variables and its application to nonlinear filtering. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 3617–3620, IEEE.
- GUTENKUNST, R.N., WATERFALL, J.J., CASEY, F.P., BROWN, K.S., MYERS, C.R. & SETHNA, J.P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, **3**, e189.
- HALL, P. & WILSON, S.R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 757–762.
- HANSEN, N. & OSTERMEIER, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, **9**, 159–195.
- HANSEN, P. (1984). Shortest paths in signed graphs. *North-Holland mathematics studies*, **95**, 201–214.
- HEARN, A.C. (1987). *REDUCE user's manual*. Rand Corporation.
- HECKER, M., LAMBECK, S., TOEPFFER, S., VAN SOMEREN, E. & GUTHKE, R. (2009). Gene regulatory network inference: data integration in dynamic models - a review. *Biosystems*, **96**, 86–103.
- HEINE, T., KAWOHL, A. & KING, R. (2008). Derivative-free optimal experimental design. *Chemical Engineering Science*, **63**, 4873–4880.
- HEINEMANN, M. & PANKE, S. (2006). Synthetic biology putting engineering into biology. *Bioinformatics*, **22**, 2790–2799.
- HEISE, S., FLASSIG, R.J. & KLAMT, S. (2013). Benchmarking a simple yet effective approach for inferring gene regulatory networks from systems genetics data. In A.d.l. Fuente, ed., *Gene Network Inference: Verification of Methods from Systems Genetics Data*, 33–47, Springer Berlin, Berlin.
- HICKSON, I., ZHAO, Y., RICHARDSON, C.J., GREEN, S.J., MARTIN, N.M.B., ORR, A.I., REAPER, P.M., JACKSON, S.P., CURTIN, N.J. & SMITH, G.C.M. (2004). Identification and characterization of a novel and specific inhibitor of the ataxia-telangiectasia mutated kinase ATM. *Cancer Research*, **64**, 9152–9159.
- HILL, P.D. (1978). A review of experimental design procedures for regression model discrimination. *Technometrics*, **20**, 15–21.
- HIMMELBLAU, D.M. (1970). *Process analysis by statistical methods*. Wiley New York.
- HINDMARSH, A., BROWN, P., GRANT, K., LEE, S., SERBAN, R., SHUMAKER, D. & WOODWARD, C. (2005). Sundials: Suite of nonlinear and differential/algebraic equation solvers. *ACM Trans. Math. Softw.*, **31**, 363–396.

- HOHEISEL, J.D. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet*, **7**, 200–210.
- HOLSTEIN, K., FLOCKERZI, D. & CONRADI, C. (2013). Multistationarity in sequential distributed multisite phosphorylation networks. *Bulletin of Mathematical Biology*, **75**, 2028–2058.
- HORST, R. (2000). *Introduction to global optimization*. Springer.
- HOVORKA, R., CANONICO, V., CHASSIN, L.J., HAUETER, U., MASSI-BENEDETTI, M., FEDERICI, M.O., PIEBER, T.R., SCHALLER, H.C., SCHAUPP, L. & VERING, T. (2004). Non-linear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological measurement*, **25**, 905.
- HUGHES, J.D., ESTEP, P.W., TAVAZOIE, S. & CHURCH, G.M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in *saccharomyces cerevisiae*. *Journal of molecular biology*, **296**, 1205–1214.
- HUNTER, W.G. & REINER, A.M. (1965). Designs for discriminating between two rival models. *Technometrics*, **7**, 307–323.
- HUYNH-THU, V., IRRTHUM, A., WEHENKEL, L. & GEURTS, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, **5**.
- ITO, K. & XIONG, K. (2000). Gaussian filters for nonlinear filtering problems. *Automatic Control, IEEE Transactions on*, **45**, 910–927.
- JACQUEZ, J.A. & GREIF, P. (1985). Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Mathematical Biosciences*, **77**, 201–227.
- JAENISCH, R. & BIRD, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, **33**, 245–254.
- JAMES, G.M. (1983). Model specification tests against non-nested alternatives. Tech. rep., Queen’s University, Department of Economics.
- JANSEN, R.C. (2003). Studying complex biological systems using multifactorial perturbation. *Nature Reviews Genetics*, **4**, 145–151.
- JANSEN, R.C. & NAP, J.P. (2001). Genetical genomics: the added value from segregation. *TRENDS in Genetics*, **17**, 388–391.
- JEBARA, T., KONDOR, R. & HOWARD, A. (2004). Probability product kernels. *The Journal of Machine Learning Research*, **5**, 819–844.
- JENKINSON, G. & GOUTSIAS, J. (2011). Thermodynamically consistent model calibration in chemical kinetics. *BMC Systems Biology*, **5**, 64.
- JIMENEZ, G.S., BRYNTESSON, F., TORRES-ARZAYUS, M.I., PRIESTLEY, A., BEECHE, M., SAITO, S., SAKAGUCHI, K., APPELLA, E., JEGGO, P.A., TACCIOLI, G.E., WAHL, G.M. & HUBANK, M. (1999). DNA-dependent protein kinase is not required for the p53-dependent response to DNA damage. *Nature*, **400**, 81–3.

## REFERENCES

---

- JOHNSON, N.L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, 149–176.
- JULIER, S.J. & UHLMANN, J.K. (1996). A general method for approximating nonlinear transformations of probability distributions. Tech. rep., Technical report, Robotics Research Group, Department of Engineering Science, University of Oxford.
- KAUFFMAN, S. (1993). *The origins of order: Self organization and selection in evolution*. Oxford University Press.
- KAY, S. (1993). *Fundamentals of statistical signal processing : estimation theory*, vol. 1. Prentice-Hall PTR.
- KEURENTJES, J.J., FU, J., TERPSTRA, I.R., GARCIA, J.M., VAN DEN ACKERVEKEN, G., SNOEK, L.B., PEETERS, A.J., VREUGDENHIL, D., KOORNNEEF, M. & JANSEN, R.C. (2007). Regulatory network construction in arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences*, **104**, 1708–1713.
- KHALIL, A.S. & COLLINS, J.J. (2010). Synthetic biology: applications come of age. *Nature Reviews Genetics*, **11**, 367–379.
- KINNER, A., WU, W., STAUDT, C. & ILIAKIS, G. (2008).  $\gamma$ -H2AX in recognition and signaling of DNA double-strand breaks in the context of chromatin. *Nucleic Acids Research*, **36**, 5678–5694.
- KLAMT, S. & VON KAMP, A. (2009). Computing paths and cycles in biological interaction graphs. *BMC Bioinformatics*, **10**, 181.
- KLAMT, S., FLASSIG, R.J. & SUNDMACHER, K. (2010). TRANSWESD: inferring cellular networks with transitive reduction. *Bioinformatics*, **26**, 2160–2168.
- KLAMT, S., FLASSIG, R.J., HEISE, S. & SAMAGA, R. (2013). Sbv improver, systems biology verification: Species translation subchallenge 4, rank 1, <https://www.sbvimprover.com/challenge-2/overview>.
- KREUTZ, C. (2011). *Statistical Approaches for Molecular and Systems Biology*. Ph.D. thesis, University of Freiburg.
- KRUHLAK, M.J., CELESTE, A., DELLAIRE, G., FERNANDEZ-CAPETILLO, O., MÜLLER, W.G., MCNALLY, J.G., BAZETT-JONES, D.P. & NUSSENZWEIG, A. (2006). Changes in chromatin structure and mobility in living cells at sites of DNA double-strand breaks. *The Journal of Cell Biology*, **172**, 823–834.
- KRUMSIEK, J., PÖLSTERL, S., WITTMANN, D.M. & THEIS, F.J. (2010). Odefy—from discrete to continuous models. *BMC Bioinformatics*, **11**, 233.
- KULLBACK, S. (1959). *Statistics and Information theory*. J. Wiley and Sons, New York.
- LÄHDESMÄKI, H., SHMULEVICH, I. & YLI-HARJA, O. (2003). On learning gene regulatory networks under the boolean network model. *Machine Learning*, **52**, 147–167.
- LAPAUGH, A.S. & PAPADIMITRIOU, C.H. (1984). The even path problem for graphs and digraphs. *Networks*, **14**, 507–513.

## REFERENCES

---

- LEE, T.I., RINALDI, N.J., ROBERT, F., ODOM, D.T., BAR-JOSEPH, Z., GERBER, G.K., HANNETT, N.M., HARBISON, C.T., THOMPSON, C.M. & SIMON, I. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
- LEES-MILLER, S.P., SAKAGUCHI, K., ULLRICH, S.J., APPELLA, E. & ANDERSON, C.W. (1992). Human DNA-activated protein kinase phosphorylates serines 15 and 37 in the amino-terminal transactivation domain of human p53. *Molecular and Cellular Biology*, **12**, 5041–9.
- LI, H., LU, L., MANLY, K.F., CHESLER, E.J., BAO, L., WANG, J., ZHOU, M., WILLIAMS, R.W. & CUI, Y. (2005). Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Human molecular genetics*, **14**, 1119–1125.
- LIANG, S., FUHRMAN, S. & SOMOGYI, R. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific symposium on biocomputing*, vol. 3, 2.
- LIEBERMEISTER, W. & KLIPP, E. (2006). Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data. *Theoretical Biology and Medical Modelling*, **3**, 42.
- LIEBERMEISTER, W., UHLENDORF, J. & KLIPP, E. (2010). Modular rate laws for enzymatic reactions: thermodynamics, elasticities and implementation. *Bioinformatics*, **26**, 1528–1534.
- LIMPERT, E., STAHEL, W.A. & ABBT, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, **51**, 341–352.
- LIU, B., DE LA FUENTE, A. & HOESCHELE, I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, **178**, 1763–1776.
- LIU, B., HOESCHELE, I. & DE LA FUENTE, A. (2010). Inferring gene regulatory networks from genetical genomics data. In *Handbook of Research on Computational Methodologies in Gene Regulatory Networks*, 79–107, IGI Global.
- LJUNG, L. (1999). *System identification (2nd ed.): theory for the user*. Prentice Hall PTR.
- LORENZ, S. (2006). *The model-data-overlap*. Ph.D. thesis, Freie Universität Berlin.
- LORENZ, S., DIEDERICH, E., TELGMANN, R. & SCHÜTTE, C. (2007). Discrimination of dynamical system models for biological and chemical processes. *Journal of Computational Chemistry*, **28**, 1384–1399.
- LOUIS, A., THOMAS & CARLIN, P., BRADLEY (2000). *Bayes and Empirical Bayes Methods for Data Analysis, Second Edition*. Chapman & Hall/CRC.
- LUDDEN, T.M., BEAL, S.L. & SHEINER, L.B. (1994). Comparison of the akaike information criterion, the schwarz criterion and the f test as guides to model selection. *Journal of Pharmacokinetics and Biopharmaceutics*, **22**, 431–445.
- MANGAN, S. & ALON, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, **100**, 11980–11985.
- MARBACH, D., SCHAFFTER, T., MATTIUSSI, C. & FLOREANO, D. (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, **16**, 229–239.

## REFERENCES

---

- MARCZYK, M., JAKSIK, R., POLANSKI, A. & POLANSKA, J. (2013). Adaptive filtering of microarray gene expression data based on gaussian mixture decomposition. *BMC Bioinformatics*, **14**, 101.
- MARGOLIN, A., NEMENMAN, I., BASSO, K., WIGGINS, C., STOLOVITZKY, G., FAVERA, R. & CALIFANO, A. (2006). Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.
- MARKOWETZ, F. & SPANG, R. (2007). Inferring cellular networks - a review. *BMC bioinformatics*, **8**, S5.
- MARTIN, M., GENESCA, A., LATRE, L., JACO, I., TACCIOLI, G.E., EGOZCUE, J., BLASCO, M.A., ILIAKIS, G. & TUSELL, L. (2005). Postreplicative joining of DNA double-strand breaks causes genomic instability in DNA-PKcs-deficient mouse embryonic fibroblasts. *Cancer Research*, **65**, 10223–10232.
- MATLAB (2010). version 7.10.0 (r2010a).
- MAURYA, M.R., RENGASWAMY, R. & VENKATASUBRAMANIAN, V. (2003). A systematic framework for the development and analysis of signed digraphs for chemical processes. 1. algorithms and analysis. *Industrial & Engineering Chemistry Research*, **42**, 4789–4810.
- MEEKER, W. & ESCOBAR, L. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *The American Statistician*, **49**, 48–53.
- MELYKUTI, B., AUGUST, E., PAPACHRISTODOULOU, A. & EL-SAMAD, H. (2010). Discriminating between rival biochemical network models: three approaches to optimal experiment design. *BMC Systems Biology*, **4**, 38.
- MEYER, P., ALEXOPOULOS, L.G., BONK, T., CALIFANO, A., CHO, C.R., DE LA FUENTE, A., DE GRAAF, D., HARTEMINK, A.J., HOENG, J., IVANOV, N.V., KOEPPL, H., LINDING, R., MARBACH, D., NOREL, R., PEITSCH, M.C., RICE, J.J., ROYYURU, A., SCHACHERER, F., SPRENGEL, J., STOLLE, K., VITKUP, D. & STOLOVITZKY, G. (2011). Verification of systems biology research in the age of collaborative competition. *Nat Biotech*, **29**, 811–815.
- MEYER, P., HOENG, J., RICE, J.J., NOREL, R., SPRENGEL, J., STOLLE, K., BONK, T., CORTHESEY, S., ROYYURU, A., PEITSCH, M.C. & STOLOVITZKY, G. (2012). Industrial methodology for process verification in research (improver): toward systems biology verification. *Bioinformatics*, **28**, 1193–1201.
- MICHAELSON, J.J., LOGUERCIO, S. & BEYER, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*, **48**, 265–276.
- MICHAELSON, J.J., ALBERTS, R., SCHUGHART, K. & BEYER, A. (2010). Data-driven assessment of eQTL mapping methods. *BMC Genomics*, **11**, 1–16.
- MICHALIK, C., STUCKERT, M. & MARQUARDT, W. (2009). Optimal experimental design for discriminating numerous model candidates: the AWDC criterion. *Industrial & Engineering Chemistry Research*, **49**, 913–919.
- MLADENOV, E. & ILIAKIS, G. (2011). Induction and repair of dna double strand breaks: The increasing spectrum of non-homologous end joining pathways. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **711**, 61–72.



## REFERENCES

---

- MOURI, K., NACHER, J.C. & AKUTSU, T. (2009). A mathematical model for the detection mechanism of DNA double-strand breaks depending on autophosphorylation of ATM. *PLoS ONE*, **4**, e5131–.
- MOYLES, D.M. & THOMPSON, G.L. (1969). An algorithm for finding a minimum equivalent graph of a digraph. *Journal of the ACM (JACM)*, **16**, 455–460.
- MÜLLER, G.T. (2002). *Modeling complex systems with differential equations*. Ph.D. thesis, Albert-Ludwigs-Universität.
- MUNK, A. & CZADO, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**, 223–241.
- MURPHY, K. & MIAN, S. (1999). Modelling gene expression data using dynamic bayesian networks. Tech. rep., Technical report, Computer Science Division, University of California, Berkeley, CA.
- NEAL, J.A. & MEEK, K. (2011). Choosing the right path: Does DNA-PK help make the decision? *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **711**, 73–86.
- NEAL, J.A., DANG, V., DOUGLAS, P., WOLD, M.S., LEES-MILLER, S.P. & MEEK, K. (2011). Inhibition of homologous recombination by DNA-dependent protein kinase requires kinase activity, is titratable, and is modulated by autophosphorylation. *Molecular and Cellular Biology*, **31**, 1719–1733.
- NELANDER, S., WANG, W., NILSSON, B., SHE, Q., PRATILAS, C., ROSEN, N., GENNEMARK, P. & SANDER, C. (2008). Models from experiments: combinatorial drug perturbations of cancer cells. *Molecular Systems Biology*, **4**.
- NEVISTIC, V. (1997). *Constrained control of nonlinear systems*. Ph.D. thesis, ETH Zurich.
- NORGAARD, M., POULSEN, N.K. & RAVN, O. (2000). New developments in state estimation for nonlinear systems. *Automatica*, **36**, 1627–1638.
- NUCCI, G. & COBELLI, C. (2000). Models of subcutaneous insulin kinetics. a critical review. *Computer methods and programs in biomedicine*, **62**, 249–257.
- OBERHARDT, M.A., PALSSON, B.O. & PAPIN, J.A. (2009). Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, **5**.
- PADULO, M., CAMPOBASSO, M.S. & GUENOV, M.D. (2007). Comparative analysis of uncertainty propagation methods for robust engineering design. In *International Conference on Engineering Design, Design for Society*, Paris, France.
- PARRY, M., LOWE, J. & HANSON, C. (2009). Overshoot, adapt and recover. *Nature*, **458**, 1102–1103.
- PEARSON, G., ROBINSON, F., BEERS GIBSON, T., XU, B.E., KARANDIKAR, M., BERMAN, K. & COBB, M.H. (2001). Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions 1. *Endocrine reviews*, **22**, 153–183.

## REFERENCES

---

- PERKINS, T.J., HALLETT, M. & GLASS, L. (2004). Inferring models of gene expression dynamics. *Journal of theoretical biology*, **230**, 289–299.
- PESARAN, M.H. & WEEKS, M. (2001). Non-nested hypothesis testing: an overview. *A Companion to Theoretical Econometrics*, 279–309.
- PINNA, A., SORANZO, N., HOESCHELE, I. & DE LA FUENTE, A. (2011). Simulating systems genetics data with sysgensim. *Bioinformatics*, **27**, 2459–2462.
- PINNA, A., HEISE, S., FLASSIG, R., DE LA FUENTE, A. & KLAMT, S. (2013). Reconstruction of large-scale regulatory networks based on perturbation graphs and transitive reduction: improved methods and their evaluation. *BMC Systems Biology*, **7**, 73.
- POHJANPALO, H. (1978). System identifiability based on the power series expansion of the solution. *Mathematical Biosciences*, **41**, 21–33.
- POLLARD JR, J., BUTTE, A.J., HOBERMAN, S., JOSHI, M., LEVY, J. & PAPPO, J. (2005). A computational model to define the molecular causes of type 2 diabetes mellitus. *Diabetes Technology & Therapeutics*, **7**, 323–336.
- POLTZ, R. & NAUMANN, M. (2012). Dynamics of p53 and nf-kappab regulation in response to dna damage and identification of target proteins suitable for therapeutic intervention. *BMC Systems Biology*, **6**, 125.
- PRESS, W.H., FLANNERY, B.P., TEUKOLSKY, S. & VETTERLING, W.T. (1989). *Numerical recipes in Pascal - The art of scientific computing*. Cambridge University Press.
- PRILL, R.J., MARBACH, D., SAEZ-RODRIGUEZ, J., SORGER, P.K., ALEXOPOULOS, L.G., XUE, X., CLARKE, N.D., ALTAN-BONNET, G. & STOLOVITZKY, G. (2010). Towards a rigorous assessment of systems biology models: The dream3 challenges. *PLoS ONE*, **5**, e9202.
- PRILL, R.J., SAEZ-RODRIGUEZ, J., ALEXOPOULOS, L.G., SORGER, P.K. & STOLOVITZKY, G. (2011). Crowdsourcing network inference: the dream predictive signaling network challenge. *Science Signaling*, **4**, mr7.
- PUERNICK, P.E. & WEISS, R. (2009). The second wave of synthetic biology: from modules to systems. *Nature Reviews Molecular Cell Biology*, **10**, 410–422.
- QUACKENBUSH, J. (2002). Microarray data normalization and transformation. *Nature Genetics*, **32**, 496–501.
- RAMKRISHNA, D. (1982). A cybernetic perspective of microbial growth. In *Foundations of Biochemical engineering: Kinetics and Thermodynamics in Biological Systems.*, American Chemical Society, Washington, DC.
- RAPPUOLI, R. & ADEREM, A. (2011). A 2020 vision for vaccines against HIV, tuberculosis and malaria. *Nature*, **473**, 463–469.
- RAU, A., JAFFREZIC, F. & NUEL, G. (2013). Joint estimation of causal effects from observational and intervention gene expression data. *BMC Systems Biology*, **7**, 111.
- RAUE, A., KREUTZ, C., MAIWALD, T., BACHMANN, J., SCHILLING, M., KLINGMÜLLER, U. & TIMMER, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, **25**, 1923–1929.

## REFERENCES

---

- REIMAND, J., VAQUERIZAS, J.M., TODD, A.E., VILO, J. & LUSCOMBE, N.M. (2010). Comprehensive reanalysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets. *Nucleic Acids Research*, **38**, 4768–4777.
- RICE, J.J., TU, Y. & STOLOVITZKY, G. (2005). Reconstructing biological networks using conditional correlation analysis. *Bioinformatics*, **21**, 765–773.
- ROCKMAN, M.V. (2008). Reverse engineering the genotype-phenotype map with natural genetic variation. *Nature*, **456**, 738–744.
- ROCKMAN, M.V. & KRUGLYAK, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics*, **7**, 862–872.
- ROLLIÉ, S., MANGOLD, M. & SUNDMACHER, K. (2012). Designing biological systems: systems engineering meets synthetic biology. *Chemical Engineering Science*, **69**, 1–29.
- ROSSNER, N., HEINE, T. & KING, R. (2010). Quality-by-design using a gaussian mixture density approximation of biological uncertainties. In *Computer Applications in Biotechnology*, vol. 11, 7–12.
- ROYSTON, P. (2007). Profile likelihood for estimation and confidence intervals. *Stata Journal*, **7**, 376–387.
- SAEZ-RODRIGUEZ, J., ALEXOPOULOS, L.G., EPPERLEIN, J., SAMAGA, R., LAUFFENBURGER, D.A., KLAMT, S. & SORGER, P.K. (2009). Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular Systems Biology*, **5**.
- SCHADT, E.E., LAMB, J., YANG, X., ZHU, J., EDWARDS, S., GUHATHAKURTA, D., SIEBERTS, S.K., MONKS, S., REITMAN, M. & ZHANG, C. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, **37**, 710–717.
- SCHENKENDORF, R. & MANGOLD, M. (2011). Qualitative and quantitative optimal experimental design for parameter identification of a map kinase model. In *Proceedings of the 18th World Congress The International Federation of Automatic Control*.
- SCHENKENDORF, R. & MANGOLD, M. (2013). Online model selection approach based on unscented kalman filtering. *Journal of Process Control*, **23**, 44 – 57.
- SCHENKENDORF, R., KREMLING, A. & MANGOLD, M. (2009). Optimal experimental design with the sigma point method. *IET systems biology*, **3**, 10–23.
- SCHLÖGL, F. (1972). Chemical reaction models for non-equilibrium phase transitions. *Zeitschrift für Physik*, **253**, 147–161.
- SEBER, G. & WILD, C.J. (2003). *Nonlinear Regression (Wiley Series in Probability and Statistics)*. Wiley-Interscience.
- SEDOGLAVIC, A. (2002). A probabilistic algorithm to test local algebraic observability in polynomial time. *Journal of Symbolic Computation*, **33**, 735–755.

## REFERENCES

---

- SHAHEEN, F.S., ZNOJEK, P., FISHER, A., WEBSTER, M., PLUMMER, R., GAUGHAN, L., SMITH, G.C.M., LEUNG, H.Y., CURTIN, N.J. & ROBSON, C.N. (2011). Targeting the DNA double strand break repair machinery in prostate cancer. *PLoS ONE*, **6**, e20311.
- SHANNON, W., CULVERHOUSE, R. & DUNCAN, J. (2003). Analyzing microarray data using cluster analysis. *Pharmacogenomics*, **4**, 41–52.
- SHEN-ORR, S.S., MILO, R., MANGAN, S. & ALON, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature genetics*, **31**, 64–68.
- SHIEH, S.Y., IKEDA, M., TAYA, Y. & PRIVES, C. (1997). DNA damage-induced phosphorylation of p53 alleviates inhibition by MDM2. *Cell*, **91**, 325–334.
- SHRIVASTAV, M., MILLER, C.A., DE HARO, L.P., DURANT, S.T., CHEN, B.P.C., CHEN, D.J. & NICKOLOFF, J.A. (2009). DNA-PKcs and ATM co-regulate DNA double-strand break repair. *DNA Repair*, **8**, 920–929.
- SILVESCU, A. & HONAVAR, V. (2001). Temporal boolean network models of genetic networks and their inference from gene expression time series. *Complex Systems*, **13**, 61–78.
- SINGH, S. (1998). On establishing the credibility of a model for a system. *Chemical Engineering Research and Design*, **76**, 657–668.
- SINGH, S. (1999). Model selection for a multiresponse system. *Chemical Engineering Research and Design*, **77**, 138–150.
- SKANDA, D. & LEBIEDZ, D. (2010). An optimal experimental design approach to model discrimination in dynamic biochemical systems. *Bioinformatics*, **26**, 939–945.
- SKANDA, D. & LEBIEDZ, D. (2013). A robust optimization approach to experimental design for model discrimination of dynamical systems. *Mathematical Programming*, **141**, 405–433.
- SOLDMANN, M. (2013). Diskriminierung von dynamischen, nichtlinearen Modellen mit verteilten Parametern in der Systembiologie.
- SPENCER, S.L., GAUDET, S., ALBECK, J.G., BURKE, J.M. & SORGER, P.K. (2009). Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature*, **459**, 428–432.
- STEWART, W.E., SHON, Y. & BOX, G.E.P. (1998). Discrimination and goodness of fit of multiresponse mechanistic models. *AIChE Journal*, **44**, 1404–1412.
- STOLOVITZKY, G., MONROE, D. & CALIFANO, A. (2007). Dialogue on reverse-engineering assessment and methods. *Annals of the New York Academy of Sciences*, **1115**, 1–22.
- STOLOVITZKY, G., PRILL, R.J. & CALIFANO, A. (2009). Lessons from the DREAM2 challenges. *Annals of the New York Academy of Sciences*, **1158**, 159–195.
- STUCKI, M. & JACKSON, S.P. (2006).  $\gamma$ H2AX and MDC1: Anchoring the DNA-damage-response machinery to broken chromosomes. *DNA Repair*, **5**, 534–543.
- SZEDERKENYI, G., BANGA, J. & ALONSO, A. (2011). Inference of complex biological networks: distinguishability issues and optimization-based solutions. *BMC Systems Biology*, **5**, 177.

## REFERENCES

---

- TAVAKKOLKHAH, P. & KÜFFNER, R. (2013). Extending partially known networks. In A.d.l. Fuente, ed., *Gene Network Inference: Verification of Methods from Systems Genetics Data*, 33–47, Springer Berlin, Berlin.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
- TOPSOE, F. (2000). Some inequalities for information divergence and related measures of discrimination. *Information Theory, IEEE Transactions on*, **46**, 1602–1609.
- TRESCH, A., BEISSBARTH, T., SÜLTMANN, H., KUNER, R., POUSTKA, A. & BUNESS, A. (2007). Discrimination of direct and indirect interactions in a network of regulatory effects. *Journal of Computational Biology*, **14**, 1217–1228.
- TUMMLER, K. (2010). Methods for the discrimination of competing dynamic models in systems biology using the example of an intracellular signaling cascade.
- VAJDA, S., RABITZ, H., WALTER, E. & LECOURTIER, Y. (1989). Qualitative and quantitative identifiability analysis of nonlinear chemical kinetic models. *Chemical Engineering Communications*, **83**, 191–219.
- VALLAT, L., KEMPER, C.A., JUNG, N., MAUMY-BERTRAND, M., BERTRAND, F., MEYER, N., POCHEVILLE, A., FISHER, J.W., GRIBBEN, J.G. & BAHRAM, S. (2013). Reverse-engineering the genetic circuitry of a cancer cell with predicted intervention in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences*, **110**, 459–464.
- VAN DER MERWE, R. (2004). *Sigma-point Kalman filters for probabilistic inference in dynamic state-space models*. Ph.D. thesis, University of Stellenbosch.
- VAN DOREN, J.F., DOUMA, S.G., VAN DEN HOF, P.M., JANSEN, J.D. & BOSGRA, O.H. (2009). Identifiability: from qualitative analysis to model structure approximation. In *Proceedings of the 15th IFAC Symposium on System Identification, Saint-Malo, France*.
- VANLIER, J., TIEMANN, C.A., HILBERS, P.A. & VAN RIEL, N.A. (2012). An integrated strategy for prediction uncertainty analysis. *Bioinformatics*, **28**, 1130–5.
- VANLIER, J., TIEMANN, C.A., HILBERS, P.A. & VAN RIEL, N.A. (2014). Optimal experiment design for model selection in biochemical networks. *BMC Systems Biology*, **8**, 20.
- VARIGONDA, S., GEORGIU, T.T. & DAOUTIDIS, P. (2001). A flatness based algorithm for optimal periodic control problems. In *Proceedings of the American Control Conference*, vol. 2, 831–836.
- VELLELA, M. & QIAN, H. (2009). Stochastic dynamics and non-equilibrium thermodynamics of a bistable chemical system: the Schlögl model revisited. *Journal of The Royal Society Interface*, **6**, 925–940.
- VERHEIJEN, P.J.T. (2003). Model selection: An overview of practices in chemical engineering. In S.P. Asprey & S. Macchietto, eds., *Computer Aided Chemical Engineering*, vol. 16, 85–104, Elsevier.

## REFERENCES

---

- VIGNES, M., VANDEL, J., ALLOUCHE, D., RAMADAN-ALBAN, N., CIERCO-AYROLLES, C., SCHIEX, T., MANGIN, B. & DE GIVRY, S. (2011). Gene regulatory network reconstruction using bayesian networks, the dantzig selector, the lasso and their meta-analysis. *PLoS ONE*, **6**, e29165.
- WAGNER, A. (2001). How to reconstruct a large genetic network from  $n$  gene perturbations in fewer than  $n^2$  easy steps. *Bioinformatics*, **17**, 1183–1197.
- WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **54**, 426–482.
- WALTER, E. & PRONZATO, L. (1997). *Identification of Parametric Models: from Experimental Data*. Communications and Control Engineering, Springer.
- WALTER, E., LECOURTIER, Y. & HAPPEL, J. (1984). On the structural output distinguishability of parametric models, and its relations with structural identifiability. *Automatic Control, IEEE Transactions on*, **29**, 56–57.
- WANG, H., WANG, M., WANG, H., BÖCKER, W. & ILIAKIS, G. (2005). Complex H2AX phosphorylation patterns by multiple kinases including ATM and DNA-PK in human cells exposed to ionizing radiation and treated with kinase inhibitors. *Journal of Cellular Physiology*, **202**, 492–502.
- WARD, I.M. & CHEN, J. (2001). Histone H2AX is phosphorylated in an ATR-dependent manner in response to replicational stress. *Journal of Biological Chemistry*, **276**, 47759–47762.
- WASSERMAN, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics, Springer.
- WEBER, P., KRAMER, A., DINGLER, C. & RADDE, N. (2012). Trajectory-oriented Bayesian experiment design versus Fisher A-optimal design: an in depth comparison study. *Bioinformatics*, **28**, 535–541.
- XIONG, M., LI, J. & FANG, X. (2004). Identification of genetic networks. *Genetics*, **166**, 1037–1052.
- YUAN, J., ADAMSKI, R. & CHEN, J. (2010). Focus on histone variant H2AX: To be or not to be. *FEBS Letters*, **584**, 3717–3724.
- ZABINSKY, Z.B. (2003). *Stochastic adaptive search for global optimization*, vol. 72. Springer.
- ZHU, J., WIENER, M.C., ZHANG, C., FRIDMAN, A., MINCH, E., LUM, P.Y., SACHS, J.R. & SCHADT, E.E. (2007). Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput Biol*, **3**, e69.
- ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.
- ZUCCHINI, W. (2000). An introduction to model selection. *J Math Psychol*, **44**, 41–61.

# List of Figures

1.1	The thesis in a picture . . . . .	4
2.1	Model identification scheme . . . . .	23
2.2	Robustification scheme . . . . .	27
3.1	Approximation of nonlinear PDF mapping. . . . .	31
3.2	Model overlap landscape . . . . .	42
3.3	Comparison of robust stimulus design in case of bistability . . . . .	44
3.4	Network structures and initial data (OED 0) . . . . .	49
3.5	OED I: Parameterization of the stimulus design, design criteria and data . . . . .	51
3.6	OED II: Design criteria and data . . . . .	52
3.7	Model structure for model A2 . . . . .	53
3.8	DNA DSB model prediction . . . . .	56
4.1	Gene regulatory network and what interactions in the gene space may represent . . . . .	61
5.1	Example for transitive reduction and its extension to TRANSWESD . . . . .	69
5.2	Illustration of precision-recall and receiver-operating characteristics curves . . . . .	79
5.3	Workflow of reconstruction methodology . . . . .	83
A.1	Profile likelihood number 1 of model A2 . . . . .	107
A.2	Profile likelihood number 2 of model A2 . . . . .	108
A.3	Profile likelihood number 3 of model A2 . . . . .	109
A.4	Profile likelihood number 4 of model A2 . . . . .	110
A.5	Profile likelihood number 5 of model A2 . . . . .	111
A.6	Profile likelihood number 6 of model A2 . . . . .	112
A.7	Profile likelihood number 7 of model A2 . . . . .	113
A.8	Profile likelihood number 8 of model A2 . . . . .	114
A.9	Profile likelihood number 9 of model A2 . . . . .	115
A.10	Profile likelihood number 10 of model A2 . . . . .	116
A.11	Profile likelihood number 11 of model A2 . . . . .	117
A.12	Profile likelihood number 12 of model A2 . . . . .	118

## LIST OF FIGURES

---

A.13 Profile likelihood number 13 of model A2 . . . . .	119
A.14 Profile likelihood number 14 of model A2 . . . . .	120
A.15 Profile likelihood number 15 of model A2 . . . . .	121
A.16 Profile likelihood number 16 of model A2 . . . . .	122
A.17 Profile likelihood number 17 of model A2 . . . . .	123
A.18 Profile likelihood number 18 of model A2 . . . . .	124
A.19 Profile likelihood number 19 of model A2 . . . . .	125



# List of Tables

3.1	Monte Carlo convergency . . . . .	37
3.2	Benchmark robust OED using a signaling cascade . . . . .	39
3.3	Fit statistics of OED runs . . . . .	48
3.4	Design criteria . . . . .	50
5.1	Reconstruction results for the DREAM5/3A . . . . .	88
5.2	Reconstruction results on yeast . . . . .	91
A.1	Final parameter set for model A2 . . . . .	106

## LIST OF TABLES

---

## Publications and Statements on Authorship

### Peer-reviewed contributions

1. **R. J. Flassig**, G. Maubach, C. Täger, K. Sundmacher, M. Naumann: Experimental design, validation and dynamic modeling uncover DNA damage sensing by DNA-PK and ATM, *Molecular BioSystems*, 2014, 10:1978-1986.  
**R. J. Flassig** developed the model, designed the experiment, analyzed the data and wrote the manuscript.
2. A. Pinna, S. Heise, **R. J. Flassig**, A. de la Fuente, S. Klamt: Reconstruction of large-scale regulatory networks based on perturbation graphs and transitive reduction: improved methods and their evaluation, *BMC Systems Biology*, 2013, 7(1):73.  
**R. J. Flassig** performed part of the analysis and wrote part of the manuscript.
3. **R. J. Flassig**, S. Heise, K. Sundmacher, S. Klamt: An effective framework for reconstructing gene regulatory networks from genetical genomics data, *Bioinformatics*, 2013, 29(2):246-54.  
**R. J. Flassig** developed the method, performed the analysis and wrote the manuscript.
4. **R. J. Flassig**, K. Sundmacher: Optimal design of stimulus experiments for robust discrimination of biochemical reaction networks, *Bioinformatics*, 2012, 28(23):3089-3096.  
**R. J. Flassig** developed the method, performed the analysis and wrote the manuscript.
5. S. Klamt, **R. J. Flassig**, K. Sundmacher: TRANSWESD: inferring cellular networks with transitive reduction, *Bioinformatics*, 2010, 26(17):2160-2168.  
**R. J. Flassig** co-developed the method, performed part of the analysis and wrote part of the manuscript.
6. **R. J. Flassig**, K. Sundmacher: Nonlinear design of stimulus experiments for optimal discrimination of biochemical systems, In Proceedings of the 11th International Symposium on Process Systems Engineering, 15-19 July 2012, Singapore.  
**R. J. Flassig** developed the method, performed the analysis and wrote the manuscript.

## **Book contribution**

1. S. Heise, **R. J. Flassig**, S. Klamt: Benchmarking a simple yet effective approach for inferring gene regulatory networks from systems genetics data. Chapter 3 in Verification of methods for gene network inference from Systems Genetics data. Ed. A. de la Fuente, Springer Berlin Heidelberg, 2013, p.33-47.

**R. J. Flassig** performed the analysis and wrote the manuscript.

## Oral Presentations

### Invited Talks

1. **R. J. Flassig**: An Effective Framework for Reconstructing Gene Regulatory Networks From Genetical Genomics Data, STATSEQ Meeting on Gene Network Inference with Systems Genetic Data and Beyond, Paris, 28-29 March, 2013.
2. **R. J. Flassig**, K. Sundmacher: Design of Optimal Stimulus Experiments for Robust Discrimination of Biochemical Reaction Networks, Spring School on Systems Biology, Kloster Seeon, Germany, 28-31 March, 2012.
3. **R. J. Flassig** and K. Sundmacher: Design of Robust Discrimination Experiments for Modeling Biochemical Reactions Networks, 7th International Workshop on Mathematics in Chemical Kinetics and Engineering, Heidelberg, 18-20 May, 2011.

### Talks

1. **R. J. Flassig** and K. Sundmacher: Nonlinear Propagation of Parameter Variabilities by Dynamic Systems for Robust Optimal Design Problems, 8th International Workshop on Mathematics in Chemical Kinetics and Engineering, Chennai, India, 4-6 February, 2013.
2. **R. J. Flassig**, K. Sundmacher: Design of Stimulus Experiments for Robust Model Discrimination, 12th international conference on Systems Biology, Heidelberg/Mannheim, Germany, 28 August - 1 September 2012.
3. **R. J. Flassig**, K. Sundmacher: Nonlinear Design of Stimulus Experiments for Optimal Discrimination of Biochemical Systems, Proceedings of the 11th International Symposium on Process Systems Engineering, Singapore, 15-19 July, 2012.

## Awards

1. Dream 4 In Silico Network Challenge (2009): **3rd** Place for network reconstruction using TRANSWESD ([www.the-dream-project.org](http://www.the-dream-project.org))
2. Sbv Improver Species Translation Sub-Challenge 2 (2013): **1st** Place for optimized network reconstruction including prior knowledge (Nature, 21 Nov, 2013, Naturejobs p. 12)

**Systems Biology Verification: Species Translation Challenge completed**  
**Congratulations to the best performing teams from the sbv IMPROVER Species Translation Challenge**

In October 2013, the results of the second sbv IMPROVER challenge were shared with the scientific community at the sbv IMPROVER Symposium 2013 in Athens, Greece. The best performing teams received, among other awards, research grants of USD 20,000 (Sub-Challenge 1-3) and travel bursaries. The results are planned to be published in early 2014.

Sub-Challenge 1: Rank 1 <b>Team PRB:</b> Adi L. Tarca, Roberto Romero	Sub-Challenge 2: Rank 1 <b>Team AMG:</b> Gyan Bhanot, Adel Dayarian, Michael Biehl, Sahand Hormoz	Sub-Challenge 3: Rank 1 <b>Team AMG:</b> Gyan Bhanot, Adel Dayarian, Michael Biehl, Sahand Hormoz	Sub-Challenge 4: Rank 1 <b>Team Reconstructors:</b> Steffen Klamt, Robert Johann Flassig, Sandra Heise, Regina Samaga	Sub-Challenge 4: Rank 2 <b>Team Vital-IT:</b> Anastasia Chasapi, Leonore Wigger, Julien Dorier, Ioannis Xenarios, Mark Ibberson, Nicolas Guex
<b>Team AMG:</b> Gyan Bhanot, Adel Dayarian, Michael Biehl, Sahand Hormoz	Sub-Challenge 2: Rank 2 <b>Team IGB:</b> Peter Sadowski, Michael Zeller	Sub-Challenge 3: Rank 2 <b>Team PRB:</b> Adi L. Tarca, Roberto Romero	<b>Team PITT.DBMLDREAM:</b> Luja Chen, Xinghua Lu	Sub-Challenge 4: Rank 3 <b>Team UPITT.Trans.Med:</b> Chunhui Cai
<b>Team Clemson:</b> Feng Luo, Zhiming Wang		<b>Team Edith:</b> Christoph Hafemeister		Sub-Challenge 4: Rank 4 <b>Team PNNL:</b> Hugh Mitchell, Susan Tilton, Jason McDermott, Joel G. Pounds

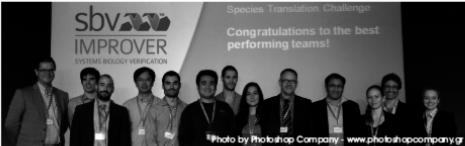



Photo by [Photoblog.Companys.com](http://Photoblog.Companys.com) / [www.photoblog.companys.com](http://www.photoblog.companys.com)

From left to right: Manuel Pelsch (Philip Morris International), Peter Sadowski, Michael Zeller, Xinghua Lu, Sahand Hormoz, Chunhui Cai, Christoph Hafemeister, Anastasia Chasapi, Michael Biehl, Gyan Bhanot, Sandra Heise, Gustavo Stolovitzky (Thomas J. Watson Research Center IBM), Julia Hoeng (Philip Morris International).

The third sbv IMPROVER challenge about Biological Network Verification started in October 2013 and runs until January 2014. For more details please visit [www.sbvimprover.com](http://www.sbvimprover.com)

sbv IMPROVER stands for systems biology verification, Industrial Methodology for Process Verification in Research, and it is a joint effort aimed at verification of systems biology in an industrial context by scientists from Philip Morris International's (PMI) Research and Development department and IBM's Thomas J. Watson Research Center. The project is funded by PMI.



## Student Assistance

1. K. Tummler: *Modelldiskriminierung konkurrierende dynamischer Modelle in der Systembiologie am Beispiel einer intrazellulären Signalkaskade* (Study project, 2010, Biosystemtechnik)
2. J. Diedrich: *Optimale Versuchsplanung zur Modelldiskriminierung von dynamischen Modellen in der Systembiologie am Beispiel eines Signalmotivs* (Bachelor thesis, 2011, Biosystemtechnik)
3. S. Heise: *Theoretische Methoden zur Generierung von Perturbationsgraphen aus experimentellen Daten für die Rekonstruktion genregulatorischer Netzwerke* (Diploma thesis, 2011, Biosystemtechnik)
4. I. Migal: *Modeling light harvesting of photosynthetic organisms* (Study project, 2012, Biosystemtechnik)
5. M. Soldman: *Diskriminierung von dynamischen, nichtlinearen Modellen mit verteilten Parametern in der Systembiologie* (Bachelor thesis, 2013, Biosystemtechnik)
6. I. Migal: *Dynamic-kinetic modeling of light energy conversion in microalgae* (Diploma thesis, 2013, Biosystemtechnik)

# Curriculum Vitae

## Personal Details

Name : Robert Johann Flassig  
Date of birth : 20 June 1981, Pritzwalk, Germany  
Marital status : Married  
Children : 3 (Alwin, Hedda, Almuth)  
Residential address : Gerwischerstr. 9, D-39114 Magdeburg

## Work Experience

Since 9/2012 : **Max Planck Institute for Dynamics of Complex Technical Systems**, Magdeburg, Germany, Research Assistant  
9/2009 – 8/2012 : **Otto-von-Guericke University**, Magdeburg, Germany, Research Assistant  
2/2009 – 8/2009 : **Max Planck Institute for Dynamics of Complex Technical Systems**, Magdeburg, Germany, Research Assistant  
7/2006 – 1/2007 : **Rolls-Royce Dtl.**, Dahlewitz, Germany, R&D in compressor design, Working Student  
10/2007 – 8/2008 : **LPMMC/CNRS**, Grenoble, France, Working Student  
4/2005 – 6/2005 : **LPMMC/CNRS**, Grenoble, France, Working Student  
8/2001 – 7/2002 : **Civilian service** in Hanover, Germany with professional training in first-aid and life-rescue  
8/2002 – 9/2002 : **Life-rescue** at Johanniter-Unfall-Hilfe e.V., Hanover, Germany

## Academic Training

9/2009 – 4/2014 : **Otto-von-Guericke University and Max Planck Institute for Dynamics of Complex Technical Systems**, Magdeburg, Ph.D. student (Ph.D. defense colloquium September 1, 2014)  
9/2005 – 9/2008 : **University Potsdam**, Germany, Dipl.-Phys. (with distinction)  
9/2004 – 8/2005 : **Université Joseph Fourier**, Grenoble, France, Studies in Physics  
10/2002 – 8/2004 : **University Potsdam**, Germany, Pre-Diploma thesis (very good)  
6/2001 : **Johann-Wolfgang-von-Goethe Gymnasium**, Pritzwalk, Germany, University-entrance Diploma  
8/1998 – 7/1999 : **Wautoma High School**, WI, USA

## Awards

2013 : **Sbv Improver** Species Translation Sub-Challenge 2: **1st Place** for optimized network reconstruction including prior knowledge (Nature, 21 Nov, 2013, Naturejobs p. 12)  
2009 : **DREAM 4** In Silico Network Challenge: **3rd Place** for network reconstruction using TRANSWESD  
2001 : **Deutsche Physikalische Gesellschaft**, Bad Honnef, Free membership